PRISM: A MULTI-DIMENSIONAL BENCHMARK FOR EVALUATING LLM PEER REVIEWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth in submissions to machine learning venues has strained the scientific peer-review system and intensified interest in LLM-based automated peer reviewers. However, how good these systems are actually, especially compared to human reviewers at catching scientific gaps, remains poorly understood. In this work, we introduce **PRISM** (**P**eer **R**eview **I**ntelligence via **S**tructured **M**ulti-dimensional assessment), a benchmarking framework that evaluates review quality across four dimensions: **Depth of Analysis**, **Novelty Assessment**, **Flaw Identification & Major Issues Prioritization**, and **Multi-dimensional Constructiveness**. Unlike most existing evaluations based on surface-level metrics like ROUGE and BLEU, or unconstrained LLM-as-a-judge prompting that conflates fluency with rigor, PRISM grounds each dimension in argument mining, retrieval-augmented verification, and consensus-based scoring. We apply PRISM to benchmark five leading automated reviewer systems and human reviewers on a stratified corpus of reviews from ICLR, ICML, and NeurIPS. The results reveal that LLMs can match or beat human reviewers on individual dimensions: comparable depth of analysis, stronger novelty verification, and higher flaw recall. However, no single system consistently matches the balanced performance of the human baseline across all dimensions at once. Each exhibits a distinct specialization profile with characteristic blind spots—failure modes that aggregate metrics miss entirely. The implication is that *LLM reviewers are best understood as targeted supplements to human review, effective within specific dimensions, but unreliable as standalone replacements.*

1 INTRODUCTION

Scientific peer review is under mounting strain: NeurIPS submissions grew from 15,671 (2024) to 21,575 (2025) Paper Copilot (2024b); Communications Chairs 2025 (2025) and ICML by 44.9% and 25.4% in consecutive years Paper Copilot (2023; 2024a; 2025), outpacing the reviewer pool and prompting load-management measures such as mandatory reviewing and author self-ranking Su & Su (2026). Meanwhile, Large Language Models (LLMs) have evolved from proofreading aids into autonomous reviewer agents that draft complete critiques Chang et al. (2025); Gao et al. (2024); Yu et al. (2024); Zhu et al. (2025); Weng et al. (2025), and an estimated 17–21% of reviews at recent top-tier venues already involve LLM assistance Liang et al. (2024); Wang et al. (2024); ICLR 2026 Program Chairs (2025). Venue policies now range from outright bans to mandatory disclosure ICML (2026).

This reality raises an important question:*Are LLMs sufficient reviewers to evaluate scientific work – and, critically, are they better at identifying gaps in a paper than human reviewers who increasingly work under time constraints and review overload?* Answering this question is important when growing evidence suggests that human review quality and reliability may be degrading under mounting pressures. For example, the NeurIPS consistency experiment Beygelzimer et al. (2023) suggested that as many as 23% of acceptance decisions may change depending purely on reviewer assignment.

We address this by introducing a benchmark to evaluate both LLM-generated and human reviews, grounded by official reviewer guidelines of established machine learning venues (e.g., ICLR, NeurIPS). A high-quality peer review must go beyond mere summarization to satisfy four core duties: evaluating technical soundness, contextualizing originality, diagnosing critical errors, and

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.

providing actionable feedback. Accordingly, our benchmark evaluates whether the reviewers can fulfill these mandates across four dimensions:

RQ1 Depth of Analysis: Do reviewers engage with a paper's methodological and empirical claims in depth, or default to surface-level assessment?

RQ2 Novelty Assessment: Are reviewers' novelty judgments grounded in prior literature, or do they rest on unverified or incorrect assertions?

RQ3 Flaw Identification & Major Issues Prioritization: How comprehensively do reviewers detect critical scientific flaws, and do they rank fatal methodological concerns above minor textual anomalies?

RQ4 Multi-dimensional Constructiveness: How actionable, solution-oriented, and professionally calibrated is the reviewers' feedback?

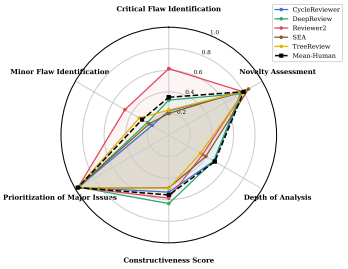

Figure 1: Macro-averaged PRISM profiles of five LLM reviewers and human reviewers.

We call this benchmark **PRISM** (**P**eer **R**eview **I**ntelligence via **S**tructured **M**ulti-dimensional assessment). Each dimension is operationalized through a dedicated evaluation pipeline, which is grounded in argument mining, retrieval-augmented verification, and consensus-based scoring. We then apply PRISM to compare five leading automated reviewer systems—TreeReview (Chang et al., 2025), Reviewer2 (Gao et al., 2024), SEA-E (Yu et al., 2024), DeepReview (Zhu et al., 2025), and CycleReviewer (Weng et al., 2025)—and human reviewers on a stratified corpus of papers drawn from ICLR, ICML, and NeurIPS. Figure 1 previews the outcome: LLM reviewers match or exceed humans on individual dimensions—CycleReviewer and DeepReview approaching human analytical depth, SEA-E in grounded novelty verification, Reviewer2 in flaw recall, and DeepReview in constructiveness—yet no single system dominates all four. Each occupies a distinct niche and leaves structured gaps that aggregate metrics hide, and only human review scores differ significantly between accepted and rejected papers (§4.2.5). *LLM reviewers are thus best understood as task-matched specialists—effective where deployed deliberately, but not yet general-purpose replacements for human reviewers.*

No single system dominates across all four dimensions: each excels in a distinct niche while leaving structured gaps invisible to aggregate metrics. This positions LLM reviewers as powerful, task-matched specialists—effective where deployed deliberately, but not yet near general-purpose replacements for human reviewers. In summary, the key contributions of this work are:

- **PRISM, a multi-dimensional benchmark for LLM peer reviewers.** Four verifiable evaluation pipelines operationalize RQ1–RQ4, applied to a stratified corpus of 600 manuscripts and expert human reviews from ICLR 2026, ICML 2025, and NeurIPS 2025.
- **A systematic human-vs-LLM analysis.** We benchmark five leading LLM reviewer systems against human reviewers on all four dimensions, validate the LLM judge against expert annotators, and reveal distinct specialization profiles and failure modes invisible to aggregate metrics.
- **Actionable deployment guidance.** We identify which systems best fit which roles within a human-assisted review pipeline.

2 Related Work

LLM-based Reviewer Systems. The rapid progress of large language models has spawned a growing family of specialized automated reviewing systems. One line of work improves review quality through structured reasoning: TreeReview Chang et al. (2025) decomposes evaluation into a hierarchical tree of questions that are recursively refined and aggregated, while DeepReview Zhu et al. (2025) emulates the slow, deliberate thinking process of expert reviewers. A complementary line focuses on optimizing the generation pipeline itself: Reviewer2 Gao et al. (2024) trains a two-stage model that first predicts review aspects and then conditions generation on them, and SEA Yu et al. (2024) standardizes heterogeneous review data before fine-tuning dedicated evaluation and analysis modules. Multi-agent collaboration offers yet another angle; CycleReviewer Weng et al. (2025) pairs a research agent with a reviewer agent in an iterative preference-training loop. While

these systems demonstrate impressive linguistic fluency, their corresponding evaluation protocols predominantly rely on generic n-gram metrics or monolithic LLM-as-a-judge scoring applied to the review as a whole. Although some works evaluate multiple criteria, these macro-level assessments are structurally blind to the granular logic of the critique: they cannot verify whether individual claims are substantiated by grounded premises, nor can they cross-check novelty assertions against retrieved prior literature.

Evaluation of AI-Generated Reviews. Evaluating AI-generated reviews is a distinct challenge from generating them. Early work relied on lexical overlap metrics—ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)—that reward surface similarity to reference reviews but are blind to reasoning quality and factual correctness (Novikova et al., 2017). Liang et al. (2024) measured point-level overlap between LLM and human feedback, finding comparable coverage but systematic gaps in methodological depth. LLM-as-a-judge Liu et al. (2023); Zheng et al. (2023) offers richer evaluation but introduces well-documented biases—position (Zheng et al., 2024), verbosity (Saito et al., 2023), and self-enhancement (Panickssery et al., 2024)—that are especially problematic when rigor, not fluency, is the target. ReviewEval (Garg et al., 2025) is the most structured prior framework, defining six dimensions including depth of analysis, constructiveness, and guideline adherence; however, it relies on end-to-end rubric prompting to assign scores and covers only 16 papers and three reviewer systems. DeepReview-Bench offers a large-scale evaluation set (1,000+ samples) but is restricted to a single venue (ICLR). RottenReviews (Ebrahimi et al., 2025) and the focus-level framework of Shin et al. (2025) study failure patterns and distributional biases in LLM reviews, but neither provides a reusable, per-review scoring protocol, while Dycke & Gurevych (2025) target reasoning faults only.

PRISM instead deploys a dedicated, verifiable pipeline per dimension—argument mining for depth, retrieval-augmented claim verification for novelty, consensus-weighted severity scoring for flaw identification and prioritization, and atomic-comment rating for constructiveness—in place of rubric-prompted LLM judging. It also benchmarks five reviewer systems and human reviewers on a stratified corpus of 600 papers from three recent venue-years (ICLR 2026, ICML 2025, and NeurIPS 2025).

3 The PRISM Framework

PRISM evaluates reviews across four independent pipelines designed to target the specific failure modes of LLMs in scientific discourse (Figure 2). Rather than asking an LLM judge for a holistic rating—which risks conflating stylistic fluency with scientific rigor—each of the pipelines in our framework decomposes the evaluation into structured evidence-extraction tasks: the LLM identifies and classifies discrete evidence units, while final scores are computed analytically. This approach ensures the evaluation is traceable and allows for precise control over metric formulation. Following sections (§3.1–3.4) detail the computational formulations and workflows for each dimension.

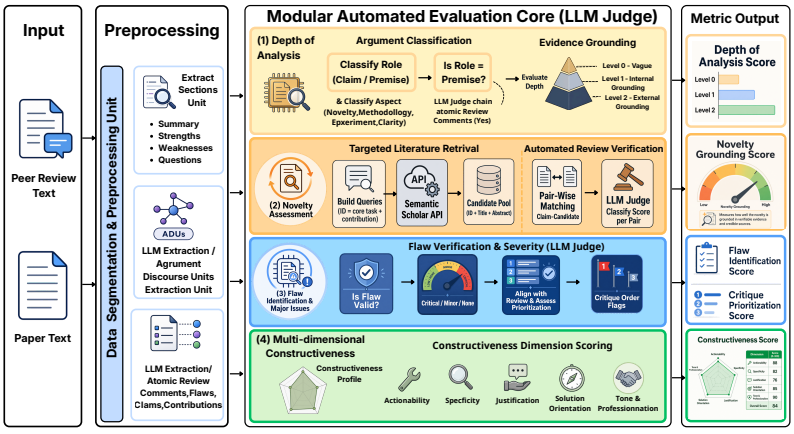

Figure 2: **Comprehensive overview of PRISM.** Reviews and manuscripts are segmented, then scored by four modular LLM-driven pipelines (§3.1–3.4), yielding a four-metric evaluation profile.

3.1 Depth of Analysis

A high-quality review is characterized not only by the presence of critical claims, but also by the substantive evidence supporting them Hua et al. (2019). We define *Depth of Analysis* (DoA) as the degree to which a reviewer substantiates judgments with objective, well-grounded premises: a shallow review relies on generic assertions, while a strong critique backs each argument with evidence.

Pipeline. We extract the core review sections (Summary, Strengths, Weaknesses) and break them into Argumentative Discourse Units (ADUs) (Peldszus & Stede, 2013). Each ADU is classified along two axes: (i) *argumentative role—Claim* (a point of contention or conclusion) or *Premise* (supporting evidence)—and (ii) *aspect topic* (Novelty, Methodology, Experiments, or Clarity). Identified premises are then assessed for *grounding level* $g(p) \in \{0, 1, 2\}$: Level 0 (Vague/Generic), Level 1 (Internal—references the manuscript directly), or Level 2 (External—references broader scientific literature).

Score Formulation. Let A be the set of all ADUs, $P \subseteq A$ the subset classified as premises, and $g_{\max} = 2$ as the maximum grounding level. We define the *Premise Ratio* $R_{\text{prem}} = |P|/|A|$ (evidence coverage) and the *normalized Average Grounding Score* $S_{\text{depth}} = \frac{1}{g_{\max}|P|} \sum_{p \in P} g(p) \in [0, 1]$ (evidence quality). DoA is defined as the harmonic mean: $\text{DoA} = \frac{2 \cdot R_{\text{prem}} \cdot S_{\text{depth}}}{R_{\text{prem}} + S_{\text{depth}}}$, which penalizes the imbalance: a review must excel in both the *proportion* and the *rigorousness* of its evidence to score highly. If $|P| = 0$, DoA $= 0$ by definition. Although aspect labels do not factor into the DoA score themselves, they reveal where reviewers direct their effort – toward substantive dimensions or surface-level concerns (Section 4.2.1).

3.2 Novelty Assessment

In scientific peer review, novelty is the degree to which a paper introduces non-trivial findings—such as new ideas, methods, data, or perspectives—relative to existing knowledge (Morgan, 1985; Mishra & Torvik, 2016; Zhao & Zhang, 2025). A genuine novelty judgment, therefore, requires situating the paper's claimed contributions within the prior literature. Our pipeline operationalizes this by verifying whether a reviewer's novelty comments are supported or refuted by retrievable prior work (Zhang et al., 2026).

Pipeline. The pipeline proceeds in three stages. ***Extraction***: a constrained LLM extracts the paper's core task, contribution anchors, and key terms, along with the set of verbatim novelty claims $\mathcal{C} = \{c_1, \ldots, c_n\}$ from the review. ***Retrieval***: we construct deterministic Semantic Scholar queries using the extracted anchors. Results are filtered for prior publications, duplication, and diversified via Maximal Marginal Relevance to form a candidate pool $\mathcal{B} = \{b_1, \ldots, b_k\}$. ***Verification***: for each claim-candidate pair (c_i, b_j), an LLM judge compares the review claim against both the paper context (abstract + introduction) and the candidate's prior work (title + abstract). It returns a discrete evidence-support score $s(c_i, b_j) \in \{-2, -1, 0, +1, +2\}$ ranging from *contradicted* to *fully supported*.

Score Formulation. Because each claim is evaluated against multiple candidates, we aggregate scores using a relevance-weighted top-3 policy ($\mathcal{T}_i$) rather than maximum pooling. This choice mitigates optimistic inflation from a single spuriously favorable match and better preserves the evidence ranking induced by retrieval. Let r_j denote the retrieval relevance of candidate b_j; the per-claim score is $S_{\text{claim}}(c_i) = \frac{\sum_{j \in \mathcal{T}_i} s(c_i, b_j)\, r_j}{\sum_{j \in \mathcal{T}_i} r_j}$. At the review level, we compute the mean claim score $\bar{S} = \frac{1}{n} \sum_{i=1}^{n} S_{\text{claim}}(c_i)$ and derive three normalized metrics— $NS(R) = \frac{\bar{S}+2}{4}$, $SR(R) = \frac{|\{c_i : S_{\text{claim}}(c_i) \geq 1\}|}{n}$, $SSR(R) = \frac{|\{c_i : S_{\text{claim}}(c_i) = 2\}|}{n}$, where $NS \in [0, 1]$ is the overall normalized score, SR and SSR measure the fraction of claims with partial and strict literature support, respectively. Together, these metrics distinguish well-grounded critiques from partial matches or unsupported hallucinations.

3.3 Flaw Identification & Major Issues Prioritization

Effective peer review requires both accurate diagnosis of scientific errors and clear structural organization. We define *Flaw Identification* as the ability to detect genuine methodological weaknesses in a manuscript while filtering minor surface-level issues. Because the absolute number of flaws in any

manuscript is unobservable, we establish a relative "ground truth" using a consensus mechanism that merges findings from both verified human and LLM reviewers. Furthermore, since authors prioritize issues encountered early in a reviewing text (Järvelin & Kekäläinen, 2002), we treat the burial of critical flaws beneath trivial formatting complaints as a significant failure in review quality.

Pipeline. The pipeline proceeds in two stages. ***Extraction***: we isolate the critical review sections (Summary, Weaknesses, Questions) from both the human and LLM reviews; an LLM parses them concurrently to extract distinct flaw arguments—specific criticisms regarding the manuscript. ***Consensus Verification***: grounded in the actual paper context, an LLM judge evaluates all extracted flaws, discarding invalid or hallucinated critiques; verified findings from both reviewer types are merged into a consensus ground truth and classified by severity into *Critical* (e.g., methodological errors, flawed proofs) or *Minor* (e.g., typos, formatting issues). ***Positional Recovery***: valid flaws are mapped back to their original sequential position within the review text, forming the ranked ordering used to compute the prioritization score.

Score Formulation. We represent the consensus sets of Critical and Minor flaws as F^{C}_{true} and F^{M}_{true}, respectively. The subsets of these valid flaws successfully identified by the reviewer under evaluation are denoted as F^{C}_{rev} and F^{M}_{rev}. **Diagnostic coverage** is measured by severity-stratified recall: Critical/Minor Recall $= \frac{|F^{C/M}_{\text{true}} \cap F^{C/M}_{\text{rev}}|}{|F^{C/M}_{\text{true}}|}$. **Structural ranking** quality is measured by the normalized Critique Prioritization Score ($nCPS$), inspired by NDCG (Järvelin & Kekäläinen, 2002). We assign severity weights $w_i \in \{2, 1\}$ for Critical/Minor flaws and let p_i be the position of the i-th valid flaw in the review: $nCPS = \frac{CPS}{iCPS}, CPS = \sum_{i=1}^{k} \frac{w_i}{\log_2(p_i+1)}$, where $iCPS$ is the ideal score (all Critical flaws preceding Minor), so an $nCPS$ approaches 1 indicates optimal prioritization.

3.4 Multi-Dimensional Constructiveness

While identifying flaws is essential, a review's real value lies in its ability to help authors improve. To measure this, we introduce the *Multi-Dimensional Constructiveness* metric, which quantifies the helpfulness of feedback. Grounded in discourse taxonomies like DISAPERE (Kennard et al., 2022), our framework systematically decomposes constructiveness into informational and social dimensions.

Pipeline. An LLM judge first breaks the review into Atomic Review Comments (ARCs), the smallest independent units of critique or suggestion. Each ARC (c_j) is then rated on a scale from 0 to 2 across five dimensions: **Actionability (D_1):** does the comment provide clear, implementable guidance rather than vague opinions?; **Specificity (D_2):** does it pinpoint concrete elements, such as specific sections or equations?; **Justification (D_3):** are assertions backed by logical reasoning or empirical evidence?; **Solution (D_4):** does the reviewer propose a path for improvement instead of just highlighting a problem?; **Tone (D_5):** is the language professional and encouraging? This dimension penalizes hostility, which can demoralize authors without improving scientific quality (Hyland & Jiang, 2020; Rao & Bareham, 2022).

Score Formulation. For a review R with n ARCs $\{c_1, \ldots, c_n\}$, the Comment-Level Constructiveness $CLC(c_j) = \frac{1}{10}\sum_{k=1}^{5} D_k(c_j) \in [0, 1]$ normalizes the five dimension scores, and the Mean Constructiveness Score $MCS(R) = \frac{1}{n}\sum_{j=1}^{n} CLC(c_j)$ averages over all comments. This formulation ensures that to achieve a perfect MCS of 1.0, a reviewer must consistently deliver specific, well-justified, actionable and professionally toned feedback across all constituent comments.

4 Experiment and analysis

4.1 Evaluation Setting

Dataset selection. PRISM is evaluated on 200 manuscripts per venue-year across three conference splits—**ICLR 2026**, **ICML 2025**, and **NeurIPS 2025** (600 papers in total; per-decision counts in Table 4, Appendix B.1). We deliberately anchor the benchmark to the most recently concluded venue-years available, such that the manuscripts and their peer reviews were released after the reported training cutoffs of both the evaluated reviewer systems and the LLM judge, where such cutoff information is available. This temporal separation is intended to reduce the likelihood that benchmark instances were encountered during model training, thereby mitigating memorization

and benchmark contamination in both review generation and evaluation. The dataset is stratified by decision category (*Reject*, *Poster*, *Spotlight*, *Oral*) and topic (keyword distribution in Figure 8, Appendix B.1). Sampling ensures every decision tier is well represented, over-sampling the rare Oral and Spotlight tiers, while remaining tractable for end-to-end multi-system evaluation.

Reviewer baselines and implementations. We evaluate five automated reviewer systems spanning two paradigms–*supervised fine-tuning* (SEA-E Yu et al. (2024), CycleReviewer Weng et al. (2025), DeepReview Zhu et al. (2025)) and *prompting-based* (Reviewer2 Gao et al. (2024), TreeReview Chang et al. (2025))—and human reviewers; see Appendix B for configuration details. **LLM-as-a-Judge implementation.** We adopt the LLM-as-a-Judge paradigm, using Gemini 2.5 Flash Lite Comanici et al. (2025) as our evaluation engine for all metric extraction and scoring tasks. Full configuration details and prompt templates are in Appendix C.

4.2 Result Analysis: LLMs vs Human-Reviewer Baselines

Table 1: Macro-Average Performance Across 3 Conferences (ICLR 2026, ICML 2025, NeurIPS 2025) compared to Human.

Baselines	Depth of Analysis	Novelty Assessment	Flaw Identification		Prioritization	Constructiveness
			Critical	Minor		
Human	0.493 ± 0.062	0.802 ± 0.195	0.311 ± 0.158	0.253 ± 0.080	0.975 ± 0.043	0.556 ± 0.072
CycleReviewer	$\mathbf{0.483} \pm 0.135$	0.789 ± 0.217	0.223 ± 0.294	0.179 ± 0.141	0.969 ± 0.114	0.530 ± 0.112
DeepReview	0.481 ± 0.132	0.769 ± 0.206	0.349 ± 0.325	0.231 ± 0.143	0.967 ± 0.076	$\mathbf{0.636} \pm 0.087$
Reviewer2	0.363 ± 0.137	0.800 ± 0.214	$\mathbf{0.644} \pm 0.304$	$\mathbf{0.470} \pm 0.181$	0.972 ± 0.052	0.586 ± 0.096
SEA	0.397 ± 0.151	$\mathbf{0.854} \pm 0.188$	0.197 ± 0.252	0.208 ± 0.118	$\mathbf{0.984} \pm 0.055$	0.491 ± 0.085
TreeReview	0.339 ± 0.116	0.813 ± 0.209	0.245 ± 0.302	0.316 ± 0.142	0.977 ± 0.044	0.495 ± 0.130

Table 1 reports macro-averaged PRISM scores for five LLM reviewer systems and the human baseline across all four dimensions; the following subsections unpack each in turn. Extended quantitative breakdowns appear in Appendices D–E and qualitative examples in Appendix F.

4.2.1 Depth of Analysis

Table 2: Core Depth of Analysis Metrics.

System	Prem. Ratio	Grounding
Human	0.557 ± 0.091	0.481 ± 0.062
CycleReviewer	$\mathbf{0.614} \pm 0.185$	0.437 ± 0.137
DeepReview	0.588 ± 0.177	0.446 ± 0.134
Reviewer2	0.376 ± 0.185	0.432 ± 0.151
SEA	0.393 ± 0.210	$\mathbf{0.464} \pm 0.119$
TreeReview	0.302 ± 0.136	0.445 ± 0.115

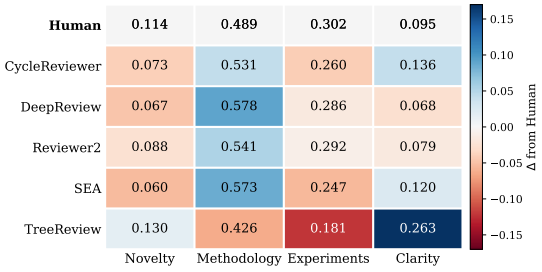

Figure 3: Premise-level aspect distributions across four topics (cf. Table 6), colored by deviation from the Human baseline (Δ).

Human reviewers set the benchmark with the highest overall DoA score (0.493). Among automated systems, **DeepReview** (0.481) and **CycleReviewer** (0.483) closely match this standard, driven by a robust *Premise Ratio* (≈ 0.60) that compensates for their slight gap in Grounding.

Table 2 reveals that Grounding scores remain consistent across humans and LLMs (0.432–0.481), so the DoA disparity is primarily driven by the Premise Ratio: while TreeReview falls short, CycleReviewer (0.614) and DeepReview (0.588) match or exceed the human baseline (0.557) in consistently substantiating their claims. Aspect distributions (Figure 3) further show that cognitive alignment is heavily architecture-dependent. Advanced pipelines (DeepReview, CycleReviewer, Reviewer2, SEA) mirror human focus by dedicating most grounded premises to Methodology and Experimental Design, while keeping *Clarity* close to human levels ($\sim$ 7–14%; Appendix E.2). By contrast, TreeReview devotes $\sim 26\%$ of its grounded premises to Clarity and presentation at the expense of methodological rigor—a degradation in evaluative depth recently observed in in-the-wild LLM peer reviews (Wu

et al., 2026). The "surface-level trap" is thus not an inherent LLM flaw, but an artifact of reasoning frameworks that lack explicit, domain-specific constraints.

Key Insight: *Human depth rests on a high Premise Ratio and a focus on methodology over formatting; the best LLMs reach comparable depth by generating abundant premises, which compensates for slightly weaker grounding.*

4.2.2 Novelty Assessment

In contrast to the human-dominated Depth of Analysis, Novelty Assessment yields uniformly high evidence-grounding scores across automated baselines: all systems fall within 0.769–0.854 (Table 1), meaning that many of their extracted novelty claims can be matched to supportive prior-work evidence under the PRISM retrieval-and-verification pipeline. Importantly, this metric does not certify the manuscript's objective novelty or human-level agreement; it measures how well the claims a reviewer chose to make are grounded in retrieved literature, so a review can score highly while still differing from humans in claim selection or calibration. Within this perspective, **SEA** achieves the highest macro-average score (**0.854**), slightly above the human baseline (0.802), suggesting that structured prompting helps models articulate retrievably justifiable novelty claims.

Figure 4 reveals that review systems diverge considerably in novelty stance: SEA endorses novelty in 83% of claims—far above the human rate of 62%—reflecting a tendency to agree with authors rather than scrutinize their contributions, whereas DeepReview adopts the most skeptical lens (41% *Novel*, 21% *Not novel*, 35% *Unclear*), suggesting its multi-step reasoning actively searches for counter-evidence. In parallel, Figure 5 exposes a recurring cross-reviewer pattern: claims labeled *Novel* generally attract stronger literature grounding than those labeled *Not novel* or *Unclear*, although the magnitude of this effect varies across reviewers. This reflects the asymmetric evidentiary bar of retrieval-based verification—*affirming novelty is corroborated whenever no closely matching prior work surfaces, whereas a "not novel" critique is fully grounded only when retrieval locates a specific prior paper anticipating the claimed contribution, a strictly harder target*. The pattern holds across reviewer pipelines and humans alike, confirming it is an intrinsic property of the retrieval-verification task rather than an LLM artifact.

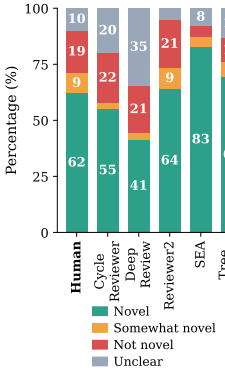

Figure 4: Novelty verdict distribution per system.

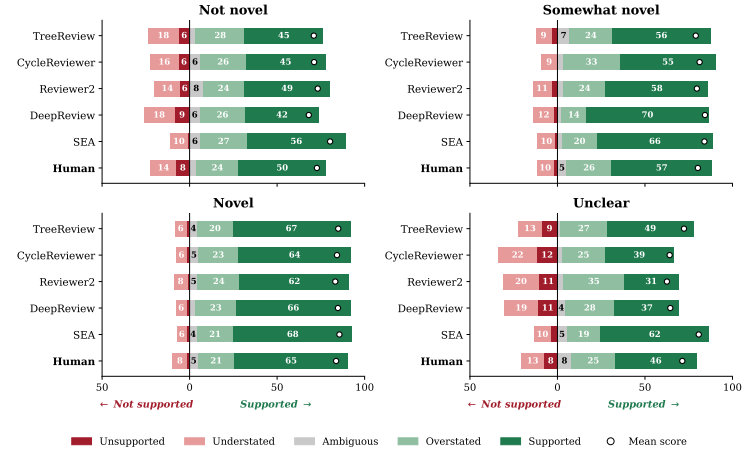

Figure 5: PRISM evidence-support breakdown for novelty comments, stratified by the reviewer's verdict.

Key Insight: *LLMs' well-grounded novelty claims reflect a preference for easily verifiable claims, not human-level judgment. For models and humans alike, affirming novelty is better grounded than denying it, which requires pinpointing a specific anticipating prior work.*

4.2.3 Flaw Identification & Major Issues Prioritization

Table 1 reveals distinct specialization profiles in diagnostic precision. **Reviewer2** stands out as an exhaustive flaw scanner, achieving the highest recall for both Critical (**0.644**) and Minor (**0.470**) issues—substantially exceeding the human baseline (0.311 and 0.253). This suggests that structured LLM pipelines can systematically surface vulnerabilities that time-constrained human reviewers

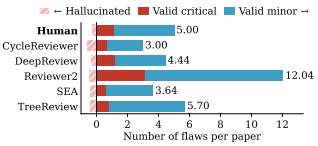

Figure 6: Valid vs. hallucinated flaws per system; all systems hallucinate minor flaws only—zero hallucinated critical flaws.

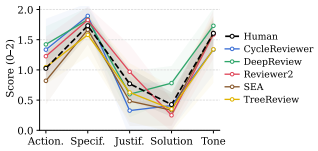

Figure 7: Constructiveness Sub-dimensions (D1-D5) Comparison.

overlook. By contrast, **DeepReview** stays close to the human baseline on Critical recall (0.349 vs. 0.311) while flagging far fewer issues overall.

Figure 6 contextualizes raw recall by decomposing extracted flaws into valid and hallucinated counts. Reviewer2 recovers an exceptionally high volume of valid flaws at a low hallucination rate ($\sim$3.6%), whereas CycleReviewer's high hallucination rate ($\sim$16.2%) signals a precision deficit. Critically, hallucinations are confined to minor issues across every system: no reviewer—human or LLM—fabricates a fatal methodological breakdown, so Critical flaw flags remain factually grounded. Aspect-level analysis (Appendix E.4) further shows that both LLMs and humans adapt their diagnostic focus by severity, concentrating on core methodology for Critical flaws and shifting toward presentation and clarity for Minor anomalies. Notably, all systems—including humans—achieve near-identical nCPS scores (≈ 0.97), suggesting that prioritizing critical over minor flaws may reflect a near-universal baseline behavior rather than a discriminating capability at current performance levels.

Key Insight: *Some LLMs act as high-sensitivity scanners, catching more critical flaws than humans; critical-first ordering, however, is universal across systems, not a distinguishing advantage.*

4.2.4 Multi-dimensional Constructiveness

The Multi-Dimensional Constructiveness Score evaluation reveals that LLMs can emulate, and in some cases exceed, the professional and supportive tone expected in academic peer review. While human reviewers establish a solid constructiveness baseline of 0.556, **DeepReview** significantly outperforms both human reviewers and other LLMs, achieving the highest score of $\mathbf{0.636}$. This suggests that DeepReview's multi-stage reasoning pipeline is exceptionally effective at not only identifying weaknesses but also formulating specific, actionable and professionally communicated suggestions for author improvement.

Figure 7 and Appendix E.5 decompose constructiveness into five dimensions (D1–D5), where each score reflects the *per-ARC average* across all atomic comments—not a binary presence indicator; a lower score means lower *density* of that attribute, not its absence. The results reveal an intriguing divergence. Both humans (1.732) and **CycleReviewer** (1.894) excel at *Specificity* (D2), yet both lag on *Solution* provision (D4 = 0.427 and 0.410)—they identify problems more readily than they propose fixes. **DeepReview** fills this gap most convincingly, leading on both *Actionability* (D1 = 1.422) and *Solution* (D4 = 0.781): it does not merely flag issues but formulates explicit, implementable improvements. **Reviewer2**'s elevated *Justification* score (D3 = 0.969) may partly reflect its verbose style rather than genuine reasoning depth, as its low *Solution* rate (D4 = 0.249) leaves critiques largely unactionable. On *Tone* (D5), DeepReview (1.730) is the most professional, whereas CycleReviewer, TreeReview, and Reviewer2 fall below the human baseline (1.605).

Key Insight: *Constructiveness is a design property, not an intrinsic LLM trait: purpose-built pipelines such as DeepReview propose actionable, professionally toned fixes that other systems rarely provide, and more often than human reviewers do.*

4.2.5 Review Sensitivity to Paper Quality: Accept vs. Reject

The RQ1–RQ4 results above show that LLM reviewers can match or exceed human performance on individual dimensions in isolation. This raises a sharper question: do the resulting scores vary with a paper's eventual outcome, or are they roughly the same regardless of the decision? To test this directly, we compare each reviewer's per-paper scores for accepted versus rejected submissions using

a two-sided Mann-Whitney U test (Holm-corrected across metrics within each reviewer; full results in Appendix E, Table 11 and Figure 29).

Human review scores differ clearly with the eventual decision, which these reviews themselves informed: reviews of accepted papers make better-grounded novelty claims ($\Delta = +0.051^{*}$) and have tighter Prioritization ($\Delta = +0.006^{***}$), while reviews of rejected papers recover significantly more of the consensus Critical ($\Delta = -0.049^{***}$) and Minor ($\Delta = -0.018^{**}$) flaws—with the Critical-flaw gap nearly $3\times$ larger than the Minor-flaw gap, showing that human critique intensifies specifically where it matters most. In sharp contrast, the same comparison is overwhelmingly non-significant across all five automated systems: only DeepReview's Prioritization Score reaches significance ($\Delta = +0.011^{***}$), while every other metric for every other LLM is statistically indistinguishable between accepted and rejected papers.

Key Insight: *Although LLM reviewers match or exceed humans on isolated dimensions (RQ1–RQ4), only human review scores differ significantly between accepted and rejected papers. LLM reviewers behave as* evaluatively invariant *scanners, applying a uniform standard regardless of merit—the strongest evidence that per-dimension strength does not yet translate into the merit-sensitive judgment that defines expert human review.*

4.3 Human Validation of LLM-as-a-Judge

To validate our evaluation engine, we conducted a targeted human evaluation with 20 experienced researchers, including doctoral researchers, senior researchers, and professors, with prior reviewing experience at major machine learning venues (ICLR, ICML, and NeurIPS). They re-scored the same pre-extracted components (ADUs, novelty claims, flaw arguments, ARCs) on a subset of 30 papers, using the exact criteria given to our LLM evaluator; the full annotation protocol is in Appendix C.4. As summarized in Table 3, the results yield **substantial agreement on objective dimensions** (Aspect Identification $\kappa = 0.699$; Grounding $\kappa_w = 0.482$). For more **subjective dimensions** involving nuanced judgments (Novelty, Constructiveness), the model achieves fair-to-moderate agreement. Crucially, the exceedingly high within-1 agreement rates (74.8%–100%) across ordinal tasks indicate that human–LLM disagreements are primarily minor boundary cases rather than fundamental misalignments. Furthermore, high true-class F1 scores (0.875 for Flaw Validity) confirm strong practical consistency, even when raw κ values are suppressed by class imbalance.

Task (dimension)	κ / κ_w	**Agreement**
Aspect ID (DoA)	0.699	77.9%
Grounding (DoA)	0.482	98.4%†
Novelty score	0.345	74.8%†
Aspect class. (Flaw)	0.514	63.5%
Severity (Flaw)	0.347	100%†
Validity (Flaw)	0.875 (F1)	76.8%
D1–D5 (Constr.)	0.14–0.52	89–99%†

Table 3: Human-expert vs. Gemini 2.5 Flash Lite agreement (†within-one on ordinal scales).

5 Conclusion

Actionable deployment recommendations. We recommend a targeted ensemble deployment rather than a standalone approach: use **Reviewer2** for flaw scanning (highest diagnostic recall); use **DeepReview** for constructive feedback drafting (highest actionability and solution density); use **SEA** for literature-grounding checks of novelty claims (highest literature support rate). Ultimately, these systems are most effective as specialist co-pilots within a human-assisted pipeline rather than autonomous reviewers.

Limitations and future works. Our primary evaluation pipeline relies on `Gemini 2.5 Flash Lite` as the core judge model. While we conducted preliminary robustness checks using an alternative model (Xiaomi `MiMo V2.5 Pro` Xiaomi MiMo Team (2026)) on a subset of the data to verify metric stability (See Appendix E), a comprehensive multi-judge study across diverse LLM families remains necessary to fully eliminate judge-specific biases. Furthermore, the benchmark corpus covers ML/AI venues only, and PRISM may require recalibration for other scientific domains. Full limitation details are in Appendix G. Future work will extend PRISM to non-ML domains and multi-judge settings (App. G).

AI USE STATEMENT

Large language models are the object of study in this work: the five automated reviewer systems we benchmark are LLM-based, and `Gemini 2.5 Flash Lite` (with `MiMo V2.5 Pro` for the robustness check) serves as the evaluation engine inside PRISM, as described in Section 3 and Appendix C; a subset of LLM-produced judgments used in our analysis were validated against expert annotators (Section 4.3). Generative AI tools were also used to assist with language editing, code debugging, and formatting of selected figures and tables. All AI-assisted outputs were manually reviewed and verified by the authors before inclusion. Generative AI was not used to autonomously determine the scientific conclusions of this work. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

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 arXiv:2306.03262*, 2023.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15651–15682, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.790. URL `https://aclanthology.org/2025.emnlp-main.790/`.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL `https://arxiv.org/abs/2507.06261`.

Communications Chairs 2025. Reflections on the 2025 review process from the program committee chairs. `https://blog.neurips.cc/2025/09/30/reflections-on-the-2025-review-process-from-the-program-committee-chairs/`, 2025. Accessed: 2026-05-07.

Nils Dycke and Iryna Gurevych. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *arXiv preprint arXiv:2508.21422*, 2025. doi: 10.48550/arXiv.2508.21422. URL `https://arxiv.org/abs/2508.21422`.

Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. Rottenreviews: Benchmarking review quality with human and llm-based judgments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 5642–5649, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720406. doi: 10.1145/3746252.3761506. URL `https://doi.org/10.1145/3746252.3761506`.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing review generation through prompt generation, 2024. URL `https://arxiv.org/abs/2402.10886`.

Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. ReviewEval: An evaluation framework for AI-generated reviews. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of*

the Association for Computational Linguistics: EMNLP 2025, pp. 20542–20564, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1120. URL https://aclanthology.org/2025.findings-emnlp.1120/.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2131–2137, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1219. URL https://aclanthology.org/N19-1219/.

Ken Hyland and Feng Kevin Jiang. "This work is antithetical to the spirit of research": An anatomy of harsh peer reviews. *Journal of English for Academic Purposes*, 46:100867, 2020. doi: 10.1016/j.jeap.2020.100867.

ICLR 2026 Program Chairs. Policies on large language model usage at ICLR 2026. https://iclr.cc/FAQ/LLM, August 2025. Accessed: 2026-05-07.

ICML. Icml 2026 policy for llm use in reviewing. https://icml.cc/Conferences/2026/LLM-Policy, 2026. Accessed: 2026-04-28.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. DISAPERE: A dataset for discourse structure in peer review discussions. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1234–1249, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.89. URL https://aclanthology.org/2022.naacl-main.89/.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024. doi: 10.1056/AIoa2400196. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400196.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Shubhanshu Mishra and Vetle I. Torvik. Quantifying conceptual novelty in the biomedical literature. *D-Lib magazine : the magazine of the Digital Library Forum*, 22 9-10, 2016. doi: 10.1045/september2016-mishra. URL https://api.semanticscholar.org/CorpusID:1732032.

Peter P. Morgan. Originality, novelty and priority: Three words to reckon with in scientific publishing. *Canadian Medical Association Journal*, 132(1):8–9, 1985. URL https://pubmed.ncbi.nlm.nih.gov/3965061/.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pp. 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238. URL https://aclanthology.org/D17-1238.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Paper Copilot. ICML 2023 statistics. https://papercopilot.com/statistics/icml-statistics/icml-2023-statistics/, 2023. Accessed: 2026-04-06.

Paper Copilot. ICML 2024 statistics. https://papercopilot.com/statistics/icml-statistics/icml-2024-statistics/, 2024a. Accessed: 2026-04-06.

Paper Copilot. NeurIPS 2024 Statistics. https://papercopilot.com/statistics/neurips-statistics/neurips-2024-statistics/, 2024b. Accessed: 2026-05-07.

Paper Copilot. ICML 2025 statistics. https://papercopilot.com/statistics/icml-statistics/icml-2025-statistics/, 2025. Accessed: 2026-05-06.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, January 2013. ISSN 1557-3958. doi: 10.4018/jcini.2013010101. URL https://doi.org/10.4018/jcini.2013010101.

Rahul Tony Rao and Beth Bareham. Regression towards the mean—a plea for civility in peer review. *BMJ*, 379:o2886, 2022. doi: 10.1136/bmj.o2886.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=magEgFpK1y.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Mind the blind spots: A focus-level evaluation framework for LLM reviews. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35630–35656, Suzhou, China, nov 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1805. URL https://aclanthology.org/2025.emnlp-main.1805/.

Weijie Su and Buxin Su. Introducing ICML 2026 policy for self-ranking in reviews. https://blog.icml.cc/2026/01/12/introducing-icml-2026-policy-for-self-ranking-in-reviews/, January 2026. Accessed: 2026-05-06.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL http://dx.doi.org/10.1007/s11704-024-40231-1.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Wenqing Wu, Chengzhi Zhang, Yi Zhao, and Tong Bao. Impact of large language models on peer review opinions from a fine-grained perspective: Evidence from top conference proceedings in AI. *arXiv preprint arXiv:2604.19578*, 2026. URL https://arxiv.org/abs/2604.19578.

Xiaomi MiMo Team. MiMo-V2.5-Pro. https://huggingface.co/XiaomiMiMo/MiMo-V2.5-Pro, 2026. Hugging Face model card. Accessed: 2026-05-07.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.595. URL https://aclanthology.org/2024.findings-emnlp.595/.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, et al. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *arXiv preprint arXiv:2601.01576*, 2026. doi: 10.48550/arXiv.2601.01576. URL https://arxiv.org/abs/2601.01576.

Yi Zhao and Chengzhi Zhang. A review on the novelty measurements of academic papers. *Scientometrics*, 130:727 – 753, 2025. doi: 10.1007/s11192-025-05234-0. URL https://api.semanticscholar.org/CorpusID:275954035.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/54dd9e0cff6d9214e20d97eb2a3bae49-Paper-Conference.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

APPENDIX TABLE OF CONTENTS

A Formal Problem Definition

B Experimental Details

B.1 Dataset Selection

B.2 Reviewer Baselines and Implementations

B.3 Review Generation Process

C PRISM Evaluation Framework: Pipeline Details

C.1 The PRISM Evaluation Pipeline

C.2 PRISM Judge Setup

C.3 Prompt Templates by Dimension

C.4 Human Validation Protocol

D Metric Independence Analysis via Pearson Correlation

E Full Cross-Dataset Quantitative Results

E.1 Statistical Significance Testing Protocol

E.2 Depth of Analysis
E.3 Novelty Assessment
E.4 Flaw Identification & Prioritization
E.5 Multi-Dimensional Constructiveness
E.6 Review Sensitivity to Paper Quality
E.7 Evaluator Robustness Across LLM Backends

F Qualitative Analysis & Case Studies

F.1 Depth of Analysis
F.2 Novelty Assessment
F.3 Flaw Identification & Major Issues Prioritization
F.4 Multi-dimensional Constructiveness

G Limitations

A FORMAL PROBLEM DEFINITION

The fundamental challenge in benchmarking automated peer reviewers lies in the highly subjective, domain-specific, and unstructured nature of scientific critiques. While existing literature often treats LLMs as either pure text generators or generic evaluators, assessing a scientific peer review requires measuring cognitive depth rather than mere linguistic fluency.

To systematically evaluate this, we formalize the peer review benchmarking process. Let P denote a submitted scientific manuscript. In our setting, an LLM-based reviewer baseline $\mathcal{M}$ processes P to generate an automated review, denoted as $R_{LLM} = \mathcal{M}(P)$. Simultaneously, we possess a high-quality human expert review R_{human} corresponding to the same manuscript P, which serves as our reference .

The central problem addressed in this work is to construct a multi-dimensional evaluation function, denoted as $\mathcal{E}$. Rather than relying on superficial n-gram matching metrics (like ROUGE) or unconstrained prompting, our framework requires $\mathcal{E}$ to process the generated review, the human reference, and the original paper to output a comprehensive capability profile:

$$\mathcal{S} = \mathcal{E}(R_{LLM}, R_{human}, P)$$

where $\mathcal{S}$ represents a set of quantitative scores spanning diverse cognitive aspects. The goal of our benchmarking protocol is to design $\mathcal{E}$ such that it accurately measures the analytical gap between R_{LLM} and R_{human}, specifically penalizing superficial summarization, hallucinated flaws, ungrounded novelty claims and un-actionable feedback.

B EXPERIMENTAL DETAILS

B.1 DATASET SELECTION.

Table 4: Distribution of selected papers across conferences and decision categories.

Venue (Year)	Oral	Spotlight	Poster	Reject	Total
ICLR 2026	26	-	80	94	200
ICML 2025	50	39	59	52	200
NeurIPS 2025	50	50	50	50	200
Total	**126**	**89**	**189**	**196**	**600**

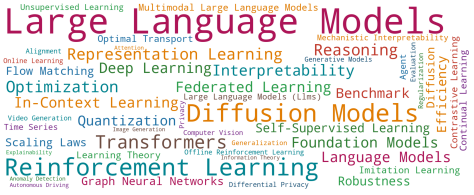

Figure 8: Top 50 popular keywords within our benchmark.

We evaluate PRISM on a stratified benchmark drawn from three conference splits: **ICLR 2026**, **ICML 2025**, and **NeurIPS 2025** (Table 4). These venue-years were selected as the most recently concluded at the time of benchmark construction, so that both the manuscripts and their peer reviews postdate the training cutoff of every evaluated baseline (see Appendix G for further discussion). For each of the three venue-years, we construct a representative subset of exactly 200 manuscripts, stratified across their final decision categories (*Reject*, *Poster*, *Spotlight*, *Oral*) and cover various topics (Figure 8). During the sampling process, we ensure that every decision tier is sufficiently represented, over-sampling the rare Oral and Spotlight tiers relative to their natural rates. This approach ensures comprehensive outcome coverage and high-fidelity review-quality diversity, while keeping the benchmark tractable for end-to-end multi-system evaluation.

B.2 REVIEWER BASELINES AND IMPLEMENTATIONS.

B.2.1 TAXONOMY OF BASELINE LLM REVIEWERS

We compare human reviews against five automated reviewer systems: **SEA-E**, **DeepReview**, **Reviewer2**, **CycleReviewer**, and **TreeReview**. These systems span two broad paradigms.

Supervised fine-tuning methods. SEA-E Yu et al. (2024) is a structured evaluation model trained to output review components such as summaries, strengths, weaknesses, and questions. CycleReviewer Weng et al. (2025) is optimized through an iterative preference-based training framework in

which a reviewer model is progressively refined from win/lose comparisons. DeepReview Zhu et al. (2025) uses a multi-stage reasoning pipeline that explicitly models multi-perspective analysis, and reliability checking before producing a final review.

Prompting-based methods. Reviewer2 Gao et al. (2024) is based on a two-stage rubric-driven process that first generates aspect-specific questions and then answers them to synthesize the final review. TreeReview Chang et al. (2025) follows a hierarchical reasoning strategy, decomposing the review into a tree of sub-questions and aggregating leaf-level evidence into a complete critique. In our experiments, prompting-based baselines are executed under a standardized backbone configuration to isolate the effect of the prompting framework rather than confounding it with backbone choice.

B.2.2 Baseline Implementation and Configuration

SEA-E SEA-E operates as a structured evaluation model, utilizing the model *ECNU-SEA/SEA-E* to generate comprehensive review components such as summaries, strengths, weaknesses, and numerical ratings. To accommodate full-length manuscripts, the engine is configured with a 70,000-token context window. The pipeline processes a batch size of 4 papers simultaneously, generating the final critique with the maximum output length capped at 8,000 tokens. To ensure a balance between analytical diversity and factual coherence, the generation hyperparameters are strictly configured with a temperature of 0.7 and a top-p of 0.9 to maintain high-quality and academically rigorous outputs.

CycleReviewer CycleReviewer utilizes a model *WestlakeNLP/CycleReviewer-ML-Llama-3.1-8B* optimized through an iterative preference-based reasoning framework. All inference workloads are executed on NVIDIA RTX A5000 GPUs. The model employs a 24,000-token context window to accommodate and process complete manuscript texts. For each manuscript, the system executes 2 to 3 iterative refinement passes to progressively enhance the review quality. The 8B engine operates on a single GPU configuration. It generates critiques with a maximum generation length of 3,000 tokens. Generation hyperparameters are configured with a temperature of 0.7, top-p of 0.9, top-k of 50, and a repetition penalty of 1.2.

DeepReview DeepReview utilizes the *WestlakeNLP/DeepReviewer-14B* core reasoning engine alongside a retrieval-augmented subsystem powered by OpenScholar. All inference workloads are executed on NVIDIA RTX A5000 GPUs. Aligning with the original architecture, the retrieval module employs *Llama-3.1-OpenScholar-8B* (configured with a 70,000-token context limit) for evidence synthesis and *Qwen-2.5-3B-Instruct* (configured with a 10,000-token context limit) for query processing. For each manuscript, the system transforms generated questions into search keywords to retrieve approximately 30 candidate papers, utilizing a dedicated reranking model to select the top 10 most relevant sources for grounding. The core 14B engine operates across two GPUs via tensor parallelism. It processes a batch size of 8 papers with a maximum generation length of 7,000 tokens. Generation hyperparameters are configured with a temperature of 0.8, top-p of 0.9, top-k of 50, and a repetition penalty of 1.2.

Reviewer2 The Reviewer2 framework originally operates on a two-stage prompting methodology utilizing custom checkpoints (*GitBag/Reviewer2_Mp* and *GitBag/Reviewer2_Mr*). However, due to the suboptimal generation quality observed from these native models during our preliminary evaluations, we replace the underlying generation engine with the open-weights *Qwen-3.5-14B* model. Crucially, we strictly retain Reviewer2's official prompt templates to preserve the methodological integrity of their two-stage pipeline. All inference workloads are executed on NVIDIA RTX A5000 GPUs. To process extensive manuscripts, the engine is configured with an 80,000-token context window. The pipeline processes a batch size of 4 papers simultaneously. During Phase 1, the model generates specific review questions, and in Phase 2, it synthesizes the final critique based on these generated prompts, with the maximum generation length capped at 7,000 tokens. Across both stages, the generation hyperparameters are strictly configured with a temperature of 0.8, top-p of 0.9, top-k of 50, and a repetition penalty of 1.2 to ensure diverse yet academically rigorous outputs.

TreeReview TreeReview models the peer review process as a hierarchical and bidirectional question-answering framework. While the original implementation utilizes GPT-4o, we standardized the backbone to *Qwen3-14B* to ensure a fair comparison across all prompting-based baselines. All

inference workloads are executed on NVIDIA RTX A5000 GPUs. To accommodate the full paper text alongside the dynamically expanding tree of sub-questions, the engine is configured with an 80,000-token context window. The pipeline processes a batch size of 4 papers simultaneously. Following its core logic, the system recursively decomposes high-level review objectives into fine-grained sub-questions and aggregates answers from leaf to root to synthesize the final critique, with the maximum output length capped at 7,000 tokens. Across all reasoning stages, the generation hyperparameters are strictly set to a temperature of 0.8, top-p of 0.9, top-k of 50, and a repetition penalty of 1.2.

B.2.3 Review generation process.

Before applying our evaluation framework, we must first generate the corresponding AI reviews. For each of the 600 papers in our dataset, we provide the complete textual content—including all sections and tables represented in text form—to all five LLM reviewer baselines. Figures and other visual elements are excluded, as the LLM reviewers considered in this study do not yet reliably support multimodal (vision–language) understanding. Each model then independently generates a complete peer review based on its respective methodology. This process yields a comprehensive corpus of 3,000 automated reviews, which serves as the primary testbed for all subsequent PRISM evaluations.

C PRISM Evaluation Framework: Pipeline Details and Experimental Setup

To ensure full reproducibility and provide transparency into our methodology, this section details the hyperparameters used for all evaluated baselines and the exact prompt templates deployed for multi-dimensional assessment.

C.1 The PRISM Evaluation Pipeline

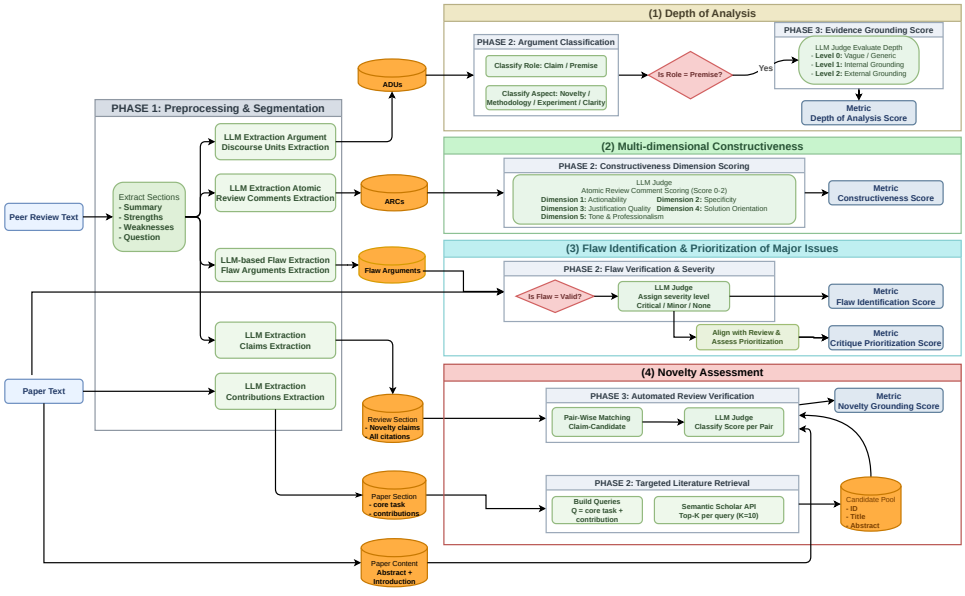

Figure 9: Detailed Flow of PRISM

C.2 PRISM Judge Setup.

To compute the diverse evaluation metrics defined in our framework PRISM, we adopt the LLM-as-a-Judge paradigm. We deploy **Gemini 2.5 Flash Lite** (`gemini-2.5-flash-lite`) (Comanici

et al., 2025) as the core evaluation engine for all metric extraction and scoring tasks. To ensure strict reproducibility and minimize generation variance, we explicitly configure the model parameters by setting the temperature to 0.0 and top-p to 0.95, without utilizing top-k sampling. During the evaluation phase, the model is strictly prompted with our standardized rubrics to systematically extract arguments, verify flaw validity against the ground truth, and compute component scores for both human and automated reviewers equally.

C.3 Prompt Templates by Dimension

Dimension 1: Depth of Analysis. The evaluation of Depth of Analysis is executed through a sequential, three-phase prompting pipeline designed to isolate and score scientific arguments. The first phase segments the raw review text into discrete Argumentative Discourse Units (ADUs), as detailed in Figure 10. Subsequently, the second phase classifies each extracted ADU into its corresponding argument role (Claim or Premise) and maps it to one of four predefined aspect topics, illustrated in Figure 11. Finally, the third phase evaluates the empirical depth of the review by computing a categorical Grounding Score specifically for the identified Premises, which is presented in Figure 12.

Phase 1: Argumentative Discourse Unit (ADU) Segmentation

ROLE AND OBJECTIVE
You are an expert NLP researcher specializing in Scholarly Argumentation Mining. Your task is to segment the following scientific peer review text into distinct Argumentative Discourse Units (ADUs).

GUIDELINES

1. After each independent logical unit, insert the exact marker `<sep>`.
2. Keep the original text in the exact same order WITHOUT adding, removing, or altering any words.
3. CRITICAL: DO split complex/compound sentences when a conclusion/claim is joined with its supporting reason/evidence. Split at:
   - Logical and causal conjunctions ("because", "as", "since", "due to", "but", "however", etc).
   - Relative pronouns (e.g., "which", "that", "who", etc).
   - Participial phrases indicating result/proof ("demonstrating", "showing", "proving", "making", "resulting in", etc).
4. IGNORE structural headings entirely (e.g., "**Summary:**", "**Strengths:**"). DO NOT append `<sep>` to them.

INPUT REVIEW TEXT:
{raw_review_text}

Figure 10: PRISM Depth of Analysis Prompt (1/3): Argumentative Discourse Unit (ADU) Segmentation.

Phase 2: Argument Role and Aspect Topic Classification

ROLE AND OBJECTIVE
Classify the Argument Role and Aspect Topic for the provided list of segmented ADUs. Use the full review as macro-context. For EACH argument in the list, classify its Argument Role and Aspect Topic.

1. ARGUMENT ROLE CLASSIFICATION (Choose exactly ONE)

- **Claim (Conclusion / Point):** The statement that is being argued *for*. It is the controversial statement or the central point that needs support.
- **Premise (Reason / Support):** The statement that is used to *support* the Claim. It provides the reasons, evidence, justifications, or grounds to accept the Claim.

Example:

- *Macro-Context:* "The proposed method is not novel. Similar architectures were already introduced by Smith et al. (2023)."
- *ADU:* "The proposed method is not novel." → **Claim** (Needs evidence to be proven).
- *ADU:* "Similar architectures were already introduced by Smith et al. (2023)." → **Premise** (Direct evidence to support the Claim).

2. ASPECT TOPIC CLASSIFICATION (Choose exactly ONE)

- **Novelty & Related Work:** Discusses originality, plagiarism, or literature review.
- **Methodology & Theoretical Soundness:** Discusses math, algorithms, architecture, or dataset guidelines.
- **Experimental Design & Evaluation:** Discusses empirical setups, baselines, ablation studies, and metrics.
- **Clarity, Presentation & Reproducibility:** Discusses writing quality, typos, or formatting.

OUTPUT FORMAT
Respond ONLY with a valid JSON object.
MACRO-CONTEXT: {macro_context}
LIST OF ARGUMENTS: {argument_list}

Figure 11: PRISM Depth of Analysis Prompt (2/3): Argument Role and Aspect Topic Classification.

Phase 3: Premise Grounding Score Evaluation

ROLE AND OBJECTIVE
You are an expert NLP researcher. I will provide a full peer review (for context) and a list of PREMISES (evidence). Evaluate the "Grounding Score" (Depth of Evidence) for EACH premise.

GROUNDING SCORE DEFINITIONS

- **Score 0 (Generic / Vague):** The premise is vague and lacks specific anchors (e.g., "The datasets", "Past research", "The equations").
- **Score 1 (Internal Grounding):** The premise explicitly anchors to elements INSIDE the manuscript (e.g., "Equation 4", "Table 2", "The proposed module").
- **Score 2 (External / Comparative):** The premise anchors to EXTERNAL knowledge outside the manuscript (e.g., "(Smith et al., 2023)", "RoBERTa", "The GLUE benchmark").

OUTPUT FORMAT
Respond ONLY with a valid JSON object.
MACRO-CONTEXT: {macro_context}
LIST OF PREMISES: {premise_list}

Figure 12: PRISM Depth of Analysis Prompt (3/3): Premise Grounding Score Evaluation.

Running Example: Depth of Analysis. To illustrate the full Depth of Analysis pipeline, consider the following excerpt from a raw review: *"3 seeds is too few to get any statistical confidence, especially without doing independent hyperparameter sweeps for each baseline. While in the past this has been standard, as a field we continually have shown that the statistical power of our experiments are laughably poor. The performance of the proposed goal-conditioned RL algorithm on the most challenging tasks was less than 50%. QRL assumes deterministic dynamics of the environment, while TD InfoNCE learns without such assumption."*

Processing this text through our three-phase framework yields the following structured output:

- **Claim:** *"3 seeds is too few to get any statistical confidence..."*
  → **Aspect:** Experimental Design & Evaluation.
- **Premise 1:** *"While in the past this has been standard, as a field we continually have shown that the statistical power of our experiments are laughably poor."*
  → **Aspect:** Experimental Design & Evaluation.
  → **Grounding Score: 0** (Generic/Vague — purely a community-level assertion with no specific anchor).
- **Premise 2:** *"The performance of the proposed goal-conditioned RL algorithm on the most challenging tasks was less than 50%."*
  → **Aspect:** Experimental Design & Evaluation.
  → **Grounding Score: 1** (Internal — directly references a quantitative result reported inside the manuscript).
- **Premise 3:** *"QRL assumes deterministic dynamics of the environment, while TD InfoNCE learns without such assumption."*
  → **Aspect:** Methodology & Theoretical Soundness.
  → **Grounding Score: 2** (External/Comparative — explicitly references QRL, a published external baseline with a known stated assumption).

Finally, the overall DoA score is defined as the harmonic mean of the Premise Ratio and the Normalized Grounding Score. This formulation ensures that a high DoA score requires both a sufficient volume of supporting arguments and a high degree of external/internal grounding:

$$DoA = 2 \times \frac{R_{premise} \times S_{grounding}}{R_{premise} + S_{grounding}}$$

Calculation for the Running Example: In the excerpt above, the pipeline extracted a total of 4 ADUs (1 Claim and 3 Premises). The premises received grounding scores of 0, 1, and 2. Applying our aggregation formulas yields:

$$R_{premise} = \frac{3}{4} = 0.75$$

$$S_{grounding} = \frac{0 + 1 + 2}{3 \times 2} = 0.5$$

$$DoA = 2 \times \frac{0.75 \times 0.5}{0.75 + 0.5} = 0.6$$

Dimension 2: Novelty Assessment. The evaluation of Novelty Assessment is executed through a three-phase pipeline that grounds reviewer novelty claims against an external body of retrieved prior work, enabling verifiable, evidence-based scoring rather than purely introspective LLM judgment. The pipeline is summarized in Figure 9.

Phase 1: Structured Target Extraction. The first phase, illustrated in Figure 13, processes both the submitted paper and the peer review in a single LLM call. From the paper, the model extracts a `core_task` (the concrete problem addressed, $\leq$20 words), up to three structured `contributions` (each with a verbatim author claim, a normalized paraphrase, and a source location hint), a set of `key_terms`, and a list of `must_have_entities` (named models, datasets, metrics). From the review, the model identifies `novelty_claims` — verbatim review sentences asserting that the paper is or is not novel — and annotates each with its argumentative stance (`not_novel` / `somewhat_novel` / `novel` / `unclear`), confidence language, and any cited prior-work strings. A deterministic regex fallback independently augments the citation list from the raw review text to prevent hallucinated references.

Phase 2: Related Work Retrieval (non-LLM). The second phase is entirely deterministic and requires no LLM call. The `core_task` and contribution names extracted in Phase 1 are composed into structured search queries, which are issued to the Semantic Scholar Academic Graph API. Raw candidates are then deduplicated via approximate title-abstract similarity ($\geq$0.96 threshold), filtered for non-technical documents (editorials, errata, etc.), and temporally constrained to papers published no later than the submission year. Finally, a Maximal Marginal Relevance (MMR) algorithm Carbonell & Goldstein (1998) selects a diverse top-K candidate pool ($K = 30$ by default) that balances retrieval relevance against redundancy.

Phase 3: LLM Judge Verification. The third phase, depicted in Figure 16, instantiates an LLM Judge for each (novelty claim, related-work candidate) pair. Given the review sentence, the abstract and introduction of the paper under review, and the title and abstract of the related work, the judge assigns a five-level verdict on a $[-2, +2]$ integer scale.

A verdict of **SUPPORTED** $(+2)$ indicates that the retrieved evidence confirms the reviewer's novelty assessment. **OVERSTATED** $(+1)$ flags cases where partial similarity exists but the reviewer's claim of "same as / not novel" is too strong. **AMBIGUOUS** (0) denotes unverifiable claims. **UNDERSTATED** (-1) penalizes reviewers who miss closely related prior work that is present in the candidate pool. **UNSUPPORTED** (-2) indicates that retrieved evidence contradicts the claim or that no supporting evidence was found.

Per-sentence scores are aggregated across the retrieved candidate pool using a configurable policy. In the experiments reported in this paper, the default is a relevance-weighted top-3 rule, not `max`: each claim is scored against its three highest-relevance candidates so that one spuriously favorable match does not dominate the result. We retain `max` only as an ablation alternative, where it yields a more optimistic estimate of support. The final **Novelty Verification Score (NS)** for a reviewer is the mean aggregated score over all their novelty claims.

Phase 1: Structured Target Extraction from Paper and Review

ROLE AND OBJECTIVE
TASK: Extract structured targets for verifiable novelty checking.
You will receive TWO sources below: PAPER TEXT and REVIEW TEXT.
Return STRICT JSON only (no markdown, no code fences, no extra keys).
The output MUST contain BOTH top-level keys: `"paper"` and `"review"`.
For novelty claims, the `"text"` field MUST be verbatim from the REVIEW TEXT (1–2 sentences max).
If the review contains no novelty claims, return an empty `novelty_claims` list but still include `review`.

OUTPUT JSON SCHEMA (must match exactly):

```
{
  "paper": {
    "core_task": "string (<=20 words)",
    "contributions": [
      {
        "name": "short name for contribution (<=15 words)",
        "author_claim_text": "verbatim quote from paper (<=40
words)",
        "description": "normalized paraphrase (<=60 words)",
        "source_hint": "location tag e.g. Abstract, Introduction
§1, Conclusion"
      }
    ],
    "key_terms": ["5-12 short phrases"],
    "must_have_entities": ["model/dataset/metric names if any"]
  },
  "review": {
    "novelty_claims": [
      {
        "claim_id": "C1",
        "text": "verbatim review claim (1-2 sentences max)",
        "stance": "not_novel | somewhat_novel | novel | unclear",
        "confidence_lang": "high | medium | low",
        "mentions_prior_work": true,
        "prior_work_strings": ["author-year strings or titles as
written"],
        "evidence_expected": "method_similarity | task_similarity |
results_similarity | theory_overlap | dataset_overlap"
      }
    ],
    "all_citations_raw": ["everything that looks like a
citation/title/arxiv id/url"]
  }
}
```

INPUT: {paper_text} {review_text}

Figure 13: PRISM Novelty Assessment Prompt (1/2): Extracting structured paper contributions and review novelty claims as verifiable targets.

Auxiliary Prompt: Core Task Extraction

ROLE AND OBJECTIVE
TASK: Extract the core task from a research paper.
You will receive the full text of a paper below.
Return STRICT JSON only (no markdown, no code fences, no extra keys).

OUTPUT JSON SCHEMA (must match exactly):

```
{
  "core_task":  "string (<=20 words)"
}
```

INPUT: {paper_text}

Figure 14: Auxiliary novelty prompt for extracting the paper's core task.

Auxiliary Prompt: Contribution Extraction

ROLE AND OBJECTIVE
TASK: Extract the main contributions claimed by the authors.
You will receive the full text of a paper below.
Return STRICT JSON only (no markdown, no code fences, no extra keys).

OUTPUT JSON SCHEMA (must match exactly):

```
{
  "contributions":  [
    {
      "name":  "complete method type phrase (<=10 words, e.g.  'A
gradient-based adversarial attack for ViTs')",
      "author_claim_text":  "verbatim quote from paper (<=40
words)",
      "description":  "normalized paraphrase (<=60 words)",
      "source_hint":  "location tag e.g.  Abstract, Introduction §1,
Conclusion"
    }
  ]
}
```

INPUT: {paper_text}

Figure 15: Auxiliary novelty prompt for extracting author-claimed contributions.

Phase 3: LLM Judge — Novelty Claim Verification

ROLE AND OBJECTIVE
INSTRUCTION:
You are an impartial Judge that verifies whether the review sentence is a claim about the paper, and how it relates to the related work evidence.
Use ONLY the provided text. If the claim is too vague or evidence is missing, return `"insufficient"`.

CLASSIFICATION

- **`claim`**: 1 if the sentence is a reviewer claim about the paper being reviewed; else 0.
- **`proof`**: 1 if the sentence provides evidence or support for a claim about the paper; else 0.

AXIS 1 — EVIDENCE SUPPORT (`stance_alignment`)

- `"aligned"`: reviewer claim aligns with and is supported by the related work evidence.
- `"partial"`: some relation exists but evidence is not conclusive.
- `"insufficient"`: claim is too vague, evidence is missing, or unverifiable.
- `"contradicted"`: evidence contradicts the reviewer claim or no supporting evidence was found.

AXIS 2 — CALIBRATION

- `"accurate"`: reviewer's strength of claim matches the actual evidence.
- `"overstated"`: reviewer claims too strongly given the evidence.
- `"understated"`: reviewer should have been stronger given the evidence.
- `"N/A"`: not applicable when evidence is insufficient to judge calibration.

VERDICT SCALE ($[-2, +2]$)

- $+2$ **SUPPORTED**: Reviewer's novelty assessment aligns with the retrieved evidence.
- $+1$ **OVERSTATED**: Some relation exists, but the reviewer's "same as / not novel" claim is too strong.
- 0 **AMBIGUOUS**: Claim is too vague or unverifiable with the provided evidence.
- -1 **UNDERSTATED**: Reviewer misses very close prior work present in the candidate pool.
- -2 **UNSUPPORTED**: Evidence contradicts the claim, or no supporting evidence was found.

OUTPUT FORMAT
Return STRICT JSON only (no markdown, no code fences, no extra keys).

```
{
  "review_sentence_id": "S_001",
  "related_paper_id": "P123",
  "classification": {"claim": 1, "proof": 0},
  "stance_alignment": "aligned",
  "calibration": "accurate",
  "score": 2,
  "label": "SUPPORTED",
  "explanation": "Short explanation"
}
```

INPUTS: {review_sentence} {paper_abstract_intro} {related_work_title_abstract}

Figure 16: PRISM Novelty Assessment Prompt (2/2): LLM Judge verifying each novelty claim against a retrieved related-work candidate on a five-level evidence scale.

Running Example: Novelty Assessment. To illustrate the full Novelty Assessment pipeline, consider a human review of a NeurIPS 2025 oral paper (`qYkhCah8OZ`), *Boosting Knowledge Utilization in Multimodal Large Language Models via Adaptive Logits Fusion and Attention Reallocation*. The benchmark concatenates all human reviews for the same submission; Phase 1 extracts three novelty claims from distinct reviewers:

- **C1** (*not_novel*, from Reviewer 1's Weaknesses): *"The proposed method appears incremental, as the techniques involving attention weighting and logits fusion are already well-known and mainly borrowed from previous works."*
- **C2** (*novel*, from Reviewer 2's Strengths): *"The proposed two modules, attention reallocation and adaptive logits fusion, offer a novel and effective perspective to enhance MLLM performance in knowledge-intensive tasks."*
- **C3** (*somewhat_novel*, from Reviewer 4's Originality assessment): *"The paper proposes a novel approach to reallocate attentions and fuse knowledge, although similar ideas for attention reallocations have been introduced by earlier work."*

Phase 2 issues structured queries to the Semantic Scholar API, retrieving 20 candidate related works. We report the three most relevant per the top-3 relevance-weighted aggregation policy: (RW1) MambaTrans: Multimodal Fusion Image Translation via LLM Priors; (RW2) Can Multimodal LLMs be Guided to Improve Industrial Anomaly Detection?; and (RW3) CAT+: Investigating and Enhancing Audio-Visual Understanding in LLMs.

Phase 3 evaluates each (claim, related-work) pair via the LLM Judge:

- **C1** (*not_novel*): RW1 is off-topic (image fusion, not attention reallocation) and "does not contain information about attention weighting or logits fusion techniques" - Unsupported (-2). RW2 corroborates that attention weighting is a known technique, as its related work section discusses MLLM enhancement via attention mechanisms - Supported ($+2$). RW3 similarly confirms these are established techniques in multimodal retrieval augmented generation - Supported ($+2$).
- **C2** (*novel*): RW1 again "does not contain any information about the paper being reviewed" - Unsupported (-2). RW2 "discusses a novel multi-expert framework" for MLLM tasks, validating the novelty claim - Supported ($+2$). RW3 "explicitly states [a] proposed [module] addressing audio-visual understanding enhancement in LLMs, aligning with the novelty claim" - Supported ($+2$).
- **C3** (*somewhat_novel*): All three related works corroborate the nuanced stance. RW1 discusses "multimodal fusion with attention, supporting the claim that similar ideas exist" - Supported ($+2$). RW2 "demonstrates that attention-based approaches are common, supporting the 'somewhat novel' stance" - Supported ($+2$). RW3 "proposes attention-based enhancement for LLMs, confirming that similar attention reallocation ideas have been introduced by earlier work" - Supported ($+2$).

The Novelty Verification Score is the mean aggregated score over all K novelty claims, where each claim's aggregated score s_k is the relevance-weighted average of its top-3 pair verdicts $v_{k,j}$. Throughout the main results, this top-3 relevance-weighted policy is the adopted default; a max-pooled variant is treated only as a sensitivity check because it typically produces more optimistic scores. Because the raw verdicts lie on a $[-2, +2]$ scale, the raw review-level mean $\bar{s}(R)$ is linearly normalized into the final $NS(R) \in [0, 1]$ via:

$$\bar{s}(R) = \frac{1}{K}\sum_{k=1}^{K} s_k, \quad s_k = \sum_{j=1}^{3} w_j \cdot v_{k,j}, \quad NS(R) = \frac{\bar{s}(R) + 2}{4}$$

Calculation for the Running Example: With $K = 3$ claims and equal relevance weights ($w_j = 1/3$):

$$s_{C_1} = \frac{(-2) + 2 + 2}{3} = 0.667, \quad s_{C_2} = \frac{(-2) + 2 + 2}{3} = 0.667, \quad s_{C_3} = \frac{2 + 2 + 2}{3} = 2.0$$

$$\bar{s}(R) = \frac{0.667 + 0.667 + 2.0}{3} = \frac{3.333}{3} = 1.111, \quad NS(R) = \frac{1.111 + 2}{4} = \mathbf{0.778}$$

This example illustrates three key behaviors of the pipeline. First, the *somewhat_novel* claim (C3) achieves the highest per-claim score because its nuanced stance ("novel… although similar ideas exist") is confirmed by all retrieved evidence, demonstrating that calibrated claims are rewarded. Second, the *not_novel* (C1) and *novel* (C2) claims receive identical aggregated scores (0.667) despite opposing stances, because the same off-topic related work (MambaTrans, which is about image fusion rather than MLLM knowledge utilization) penalizes both equally, highlighting that retrieval quality, not just claim content, drives the verdict. Third, the overall normalized score $NS(R) = 0.778$ reflects a review whose novelty assessments are partially well-grounded but sensitive to the composition of the retrieval pool.

Dimension 3: Flaw Identification & Prioritization. The evaluation of Flaw Identification and Prioritization is implemented as a two-phase pipeline that cross-validates reviewer arguments against the manuscript and quantifies both detection coverage and critical issue ordering. The first phase, illustrated in Figure 17, consolidates raw review texts from multiple reviewers (Human and LLM) by atomizing and grouping arguments into a structured inventory of *Micro-flaws*, each categorized within a hierarchical taxonomy of seven macro-topics (e.g., *Novelty & Contribution*, *Methodology & Theoretical Soundness*, *Experimental Design & Evaluation*). The second phase, presented in Figure 18, acts as an independent LLM meta-reviewer that verifies each Micro-flaw against the paper text, assigning a binary validity label (`is_valid`) and a severity rating (`Critical` or `Minor`) according to a predefined ontology grounded in whether fixing the issue requires new experiments or is purely editorial.

From these two phases, three complementary metrics are derived. **Critical Recall** and **Minor Recall** measure the proportion of ground-truth Critical and Minor flaws, respectively, that a given reviewer successfully identified, enabling fine-grained diagnosis of detection coverage across severity strata. The primary ranking metric, the **normalized Critique Prioritization Score (nCPS)**, adapts the standard NDCG formulation Järvelin & Kekäläinen (2002) to capture whether a reviewer front-loads their most severe critiques within each structural section of their review.

Phase 1: Micro-flaw Atomization and Grouping

ROLE AND OBJECTIVE
You are an expert meta-reviewer for top-tier computer science conferences. Analyze raw review texts from multiple reviewers and consolidate their arguments into a structured list of unique *Micro-flaws*.

GROUPING RULES (STRICTLY ENFORCED)

1. **Conceptual Consistency (Must Split):** Arguments grouped into the same Micro-flaw MUST address the same fundamental problem. Do not merge distinct scientific issues merely because they share a broad topic.
2. **Allowed Aggregation (Can Group):** Arguments MAY be grouped if they point to the exact same specific error in the paper (e.g., multiple reviewers citing the same missing baseline).
3. **No Forced Fit:** If an argument does not fit any existing Micro-flaw precisely, create a new one.
4. **No Upper Bound:** There is no limit on the number of Micro-flaws. Multiple Micro-flaws may share the same Macro-topic.

TAXONOMY (7 Macro-topics): Novelty & Contribution; Clarity & Presentation; Applicability, Scalability & Limitations; Experimental Design & Evaluation; Related Work & Citations; Methodology & Theoretical Soundness; Reproducibility & Open Science.

OUTPUT FORMAT
Respond ONLY with a valid JSON object.

INPUT REVIEWS: {input_text}

Figure 17: PRISM Flaw Identification Prompt (1/2): Atomizing and grouping reviewer arguments into a canonical Micro-flaw inventory.

Phase 2: Meta-reviewer Validity and Severity Judgement

ROLE AND OBJECTIVE
You are a strict and objective Meta-Reviewer. Given the full paper text and a JSON list of Micro-flaws raised by reviewers, independently verify each flaw against the manuscript.

FOR EACH MICRO-FLAW, ANSWER:

1. **`is_valid` (True/False):** Does this flaw genuinely exist in the paper? Return `False` if it is a hallucination, misunderstanding, or unreasonable request.
2. **`severity` ("Critical" / "Minor"):** If valid, assign severity per the ontology below.

SEVERITY ONTOLOGY

- **Critical** ($w = 2$): Flaws requiring new experiments, proof revisions, or core claim changes — covering Methodology, Experimental Design, Novelty, and severe Reproducibility/Applicability issues.
- **Minor** ($w = 1$): Flaws fixable via textual/editorial revision — covering Clarity & Presentation, missing citations, or documentation gaps.
- **Borderline rule:** Prefer Critical if fixing the issue can plausibly alter main conclusions; prefer Minor if the fix is purely editorial.

OUTPUT FORMAT
Respond ONLY with a valid JSON object.

INPUT: {paper_text} {micro_flaws_json}

Figure 18: PRISM Flaw Identification Prompt (2/2): LLM meta-reviewer verifying flaw validity and assigning severity labels as ground truth.

Running Example: Flaw Identification & Major Issues Prioritization. To illustrate both the Flaw Identification and Prioritization pipelines, consider a reviewer evaluating a graph neural network paper. The **Ground Truth** flaw set, defined as the union of all valid flaws identified across all reviewers for this paper, consists of:

- **Critical flaws (GT):**
  - FC1: *"Missing comparison to GraphSAGE baseline in Table 2."*
  - FC2: *"Convergence proof contains a gap in Lemma 3: the Lipschitz assumption is invoked but never verified."*
  - FC3: *"No ablation study on the effect of message-passing depth."*
- **Minor flaws (GT):**
  - FM1: *"Inconsistent notation:* $\mathbf{A}$ *and* $\tilde{\mathbf{A}}$ *used interchangeably in Eq. 4 and Eq. 5."*
  - FM2: *"Figure 2 axes are unlabeled."*
  - FM3: *"Related work omits Liu et al. (2022)."*

Reviewer X produces the following flaw list, in the order they appear in the review:

1. **[Minor]** FM1 — inconsistent notation D
2. **[Critical]** FC1 — missing GraphSAGE comparison D
3. **[Minor]** FM2 — Figure 2 axes unlabeled D
4. **[Critical]** FC2 — convergence proof gap D

Calculation for Flaw Identification:

Reviewer X matched 2 out of 3 critical flaws (missed FC3) and 2 out of 3 minor flaws (missed FM3):

$$\text{Critical Recall} = \frac{|\{\text{FC1, FC2}\}|}{|\{\text{FC1, FC2, FC3}\}|} = \frac{2}{3} \approx 0.667$$

$$\text{Minor Recall} = \frac{|\{\text{FM1, FM2}\}|}{|\{\text{FM1, FM2, FM3}\}|} = \frac{2}{3} \approx 0.667$$

Calculation for Major Issue Prioritization (nCPS):

The $nCPS$ is computed over all k GT-matched valid flaws identified by Reviewer X, ordered by their position of appearance in the review. Each flaw receives a position discount $\log_2(p_i + 1)$ and a severity weight $w_i \in \{2, 1\}$ for Critical/Minor respectively. The GT-matched valid flaws are: FM1 (Minor, position 1), FC1 (Critical, position 2), FM2 (Minor, position 3), FC2 (Critical, position 4).

$$\begin{aligned} CPS &= \frac{w_{\text{FM1}}}{\log_2(1+1)} + \frac{w_{\text{FC1}}}{\log_2(2+1)} + \frac{w_{\text{FM2}}}{\log_2(3+1)} + \frac{w_{\text{FC2}}}{\log_2(4+1)} \\ &= \frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{1}{\log_2 4} + \frac{2}{\log_2 5} \\ &= 1.000 + 1.262 + 0.500 + 0.861 \;=\; 3.623 \end{aligned}$$

The ideal score $iCPS$ places all Critical flaws first, then Minor:

$$\begin{aligned} iCPS &= \frac{2}{\log_2 2} + \frac{2}{\log_2 3} + \frac{1}{\log_2 4} + \frac{1}{\log_2 5} \\ &= 2.000 + 1.262 + 0.500 + 0.431 \;=\; 4.193 \end{aligned}$$

$$nCPS = \frac{CPS}{iCPS} = \frac{3.623}{4.193} \approx \mathbf{0.864}$$

Dimension 4: Multi-dimensional Constructiveness. To minimize cognitive overload and ensure high-fidelity scoring, the evaluation of constructiveness is decoupled into a two-phase pipeline. Phase 1, detailed in Figure 19, decomposes the holistic peer review into discrete, non-overlapping Atomic Review Comments (ARCs) while retaining their original context via anchor quotes. Phase 2, illustrated in Figure 20, then evaluates each isolated ARC against the five granular dimensions of constructiveness (D1-D5) using a stringent $[0, 2]$ rubric.

Phase 1: Atomic Review Comment (ARC) Extraction

ROLE AND OBJECTIVE
You are an expert peer-review analyst. Your task is to decompose a comprehensive peer review into distinct Atomic Review Comments (ARCs).

EXTRACTION RULES

1. Extract ALL distinct points from Summary, Strengths, Weaknesses, Questions, and Suggestions.
2. **One point per ARC:** If a single sentence contains two critiques, split it into two separate ARCs.
3. **Anchor Quote:** Provide a verbatim 5-25 word substring copied EXACTLY from the review to anchor the comment.
4. **Comment Type:** Classify each ARC as exactly one of: `weakness`, `strength`, `question`, `suggestion`, or `observation`.

OUTPUT FORMAT
Respond ONLY with a valid JSON object.

INPUT REVIEW TEXT:
{raw_review_text}

Figure 19: PRISM Constructiveness Prompt (1/2): Decomposing the review into distinct Atomic Review Comments (ARCs).

Phase 2: Multi-dimensional Constructiveness Scoring

ROLE AND OBJECTIVE
You are an expert peer-review analyst. Given the full peer review as macro-context and a list of extracted Atomic Review Comments (ARCs), score EACH comment on 5 dimensions.

SCORING RUBRIC (Score 0, 1, or 2 for each)

- **D1_actionability** — Can the author act on this?
  - **0**: Opinion with no guidance (e.g., "poorly written")
  - **1**: General direction (e.g., "needs more baselines")
  - **2**: Specific, implementable (e.g., "add [MethodX] on CIFAR-10")
- **D2_specificity** — References concrete paper elements?
  - **0**: Vague (e.g., "has issues")
  - **1**: Semi-specific (e.g., "methodology section unclear")
  - **2**: Pinpoints exact element (e.g., "Eq 7 in Sec 4.2 missing term")
- **D3_justification** — Evidence-backed?
  - **0**: Bare assertion (e.g., "not novel")
  - **1**: Partial reasoning (e.g., "similar to prior work on X")
  - **2**: Full evidence (e.g., "same loss as [Author2020] Eq 3")
- **D4_solution** — Suggests improvements?
  - **0**: Problem-only (e.g., "baselines weak")
  - **1**: Implicit fix (e.g., "lacks recent SOTA")
  - **2**: Explicit fix (e.g., "add [M2023] achieving X%")
- **D5_tone** — Respectful?
  - **0**: Hostile / dismissive
  - **1**: Neutral, factual
  - **2**: Professional-constructive, encouraging

OUTPUT FORMAT
Respond ONLY with a valid JSON object containing the `arc_id` and the numerical scores for `D1` through `D5`.

MACRO-CONTEXT: {raw_review_text}
LIST OF ARCS: {arc_json_list}

Figure 20: PRISM Constructiveness Prompt (2/2): Evaluating isolated ARCs against the five-dimensional constructiveness rubric.

Running Example: Multi-dimensional Constructiveness. To illustrate the MCS pipeline, consider the following excerpt from a human review of a theoretical machine learning paper: *"The paper lacks a clear comparison of its theoretical results (Table 1, Section 5) with prior related work. No experimental results. The toy example should correspond to the motivation example. Provide concrete toy examples illustrating setup and theorems, including specific distributions and query complexity bounds. A detailed comparison to existing results in the bandits literature is needed."*

Processing this text through Gemini yields the following Atomic Review Comments (ARCs):

- **ARC 1 (Weakness):** *"Lacks clear comparison of theoretical results (Table 1, Sec. 5) with prior related work."*
  → **Scores:** D1=1 (identifies gap but no specific fix), D2=2 (names Table 1, Sec. 5), D3=1 (partial reasoning), D4=0 (no fix), D5=1 (neutral tone).

- **ARC 2 (Weakness):** *"No experimental results to validate theoretical findings."*
  → **Scores:** D1=2 (clear: add experiments), D2=2 (specific), D3=0 (no justification why), D4=1 (implicit: conduct experiments), D5=1 (factual).
- **ARC 3 (Question):** *"Provide concrete toy examples with distributions and query complexity bounds."*
  → **Scores:** D1=2 (actionable), D2=2 (specific), D3=0 (no WHY explained), D4=1 (suggested fix), D5=2 (professional-constructive framing).
- **ARC 4 (Weakness):** *"Detailed comparison to existing bandit results needed."*
  → **Scores:** D1=2 (actionable), D2=2 (references bandit literature), D3=0 (no justification), D4=0 (problem-only), D5=1 (neutral).

Aggregating dimension-wise over these 4 ARCs:

$$\overline{D_1} = \frac{1+2+2+2}{4} = 1.75, \quad \overline{D_2} = \frac{2+2+2+2}{4} = 2.00$$
$$\overline{D_3} = \frac{1+0+0+0}{4} = 0.25, \quad \overline{D_4} = \frac{0+1+1+0}{4} = 0.50$$
$$\overline{D_5} = \frac{1+1+2+1}{4} = 1.25$$

The Multi-dimensional Constructiveness Score is the normalized average:

$$\text{MCS} = \frac{1}{5 \times 2} \left(\overline{D_1} + \overline{D_2} + \overline{D_3} + \overline{D_4} + \overline{D_5} \right)$$

Calculation for the Running Example:

$$\text{MCS} = \frac{1.75 + 2.00 + 0.25 + 0.50 + 1.25}{10} = \frac{5.75}{10} = \mathbf{0.575}$$

C.4 Human Validation Protocol

To rigorously validate our evaluation engine, we conducted a targeted human evaluation on a subset of 30 papers. We recruited 20 senior researchers, including doctoral researchers, senior researchers, and professors, with prior reviewing experience at venues such as ICLR, ICML, NeurIPS. We bypassed the initial extraction phases; granular components (Argumentative Discourse Units, Novelty Claims, Flaw Arguments, Atomic Review Comments) were pre-extracted.

Our validation scale is comparable to, and in terms of annotator coverage extends, those adopted in prior work on automated peer review. For example, TreeReview Chang et al. (2025) conducted human evaluation on 20 papers with five experts, while ReviewEval Garg et al. (2025) evaluated 30 papers with five experts. Our study therefore operates within a similar paper-level evaluation scale while involving a broader pool of human annotators.

To focus the study specifically on the reliability of the evaluation stage, we bypassed the initial extraction phases and provided annotators with pre-extracted granular components, including Argumentative Discourse Units (ADUs), novelty claims, flaw arguments, and Atomic Review Comments (ARCs). Requiring annotators to independently repeat the extraction process would substantially increase the annotation burden, as each paper yields numerous fine-grained units across multiple evaluation dimensions, each of which must subsequently be assessed under dimension-specific criteria. More importantly, using pre-extracted components controls for extraction variability: human and LLM evaluators assess identical inputs, so their agreement reflects differences in evaluation rather than differences in what each evaluator extracts from the review. Annotators were then asked to score these pre-segmented inputs using the same criteria provided to our LLM evaluator:

- **Depth of Analysis:** Annotators evaluate pre-segmented ADUs through three tasks: (1) Role Classification (Claim vs. Premise), (2) Aspect Classification (assigning one aspect topic), and (3) Grounding Score (0/1/2, for premises only).
- **Novelty Assessment:** Given pre-extracted novelty claims and pre-retrieved related papers, annotators assign a literature-grounded score (-2 to +2) for each claim.

- **Flaw Identification:** Annotators work pairwise (Human Reviewer vs. LLM Reviewer), creating a local flaw pool, assigning validity labels, and classifying severity (Minor/Critical) alongside aspect topics.
- **Multi-dimensional Constructiveness:** Using pre-extracted ARCs, annotators score each unit across five dimensions (Actionability, Specificity, Justification, Solution, Tone) on a 0/1/2 scale.

Table 3 in Section 4.3 reports the resulting Cohen's κ/κ_w and exact/± 1 agreement rates between annotators and `Gemini 2.5 Flash Lite` for each task above.

D Metric Independence Analysis via Pearson Correlation

D.1 Motivation and Objective

A key requirement for a multi-dimensional evaluation benchmark is that its constituent metrics should capture *distinct, non-overlapping* aspects of review quality. If two metrics were highly correlated, they would convey redundant information and effectively reduce the dimensionality of the evaluation, undermining the claim that different facets of peer review are independently assessed. To verify this property, we conduct a pairwise Pearson correlation analysis across the five evaluation dimensions of our benchmark: Depth of Analysis ($\mathbf{DoA}_{\mathrm{HM}}$), Novelty Assessment (**NS**), Flaw Identification (**Critical Recall**, **Minor Recall**), Issue Prioritization (**nCPS**) and Multi-dimensional Constructiveness (**MCS**).

D.2 Statistical Method

Pearson Correlation Coefficient. For two metric vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ measured over n paper-review samples, the Pearson correlation coefficient is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where $r_{xy} \in [-1, 1]$. A value of $r = 0$ indicates no linear association; $|r| = 1$ indicates perfect linear dependence.

Significance Test. To assess whether an observed r_{xy} differs significantly from zero, we apply the two-tailed t-test under the null hypothesis $H_0 \colon \rho = 0$ (no linear correlation in the population). The test statistic is:

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}} \tag{2}$$

which follows a Student's t-distribution with $n - 2$ degrees of freedom under H_0. We report two-tailed p-values with significance thresholds $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and label non-significant results as *ns*.

Effect Size Interpretation. Statistical significance alone is insufficient because large samples can render even trivially small correlations significant. We therefore assess the *practical magnitude* of each $|r_{xy}|$ using the conventional thresholds of (Cohen, 1988): $|r| < 0.10$ (negligible), $0.10 \leq |r| < 0.30$ (small), $0.30 \leq |r| < 0.50$ (moderate) and $|r| \geq 0.50$ (large). Only correlations that are both statistically significant *and* of moderate-to-large magnitude are considered substantively meaningful.

D.3 Results and Discussion

Figure 21 presents the full pairwise Pearson correlation matrix across all six metrics. The results consistently show very weak inter-metric associations, with a maximum absolute coefficient of $|r|_{\max} = 0.193$.

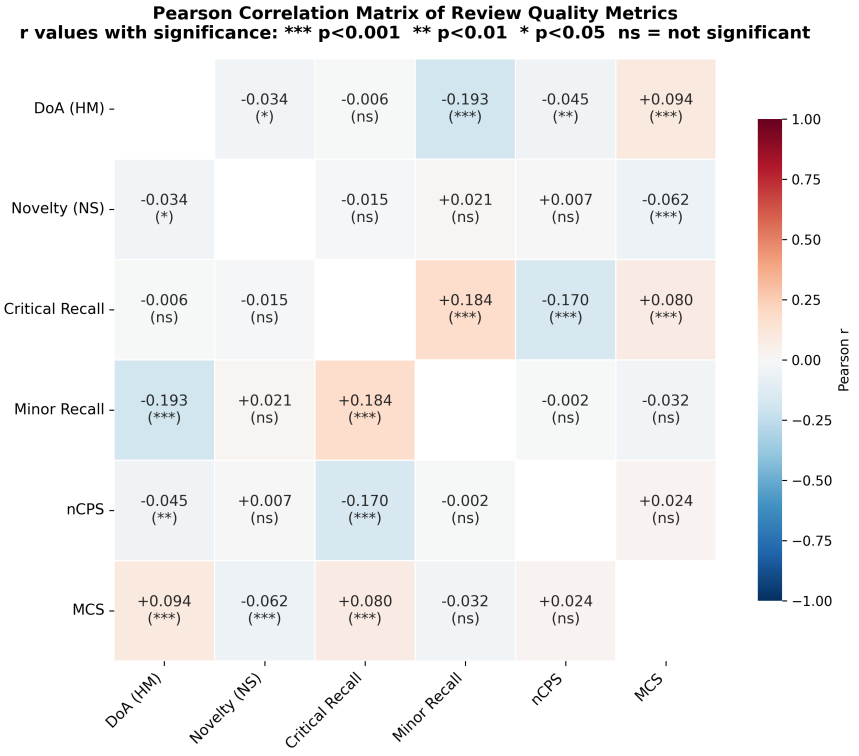

Figure 21: Pearson correlation matrix of the six review quality metrics. Each cell reports the Pearson r coefficient with significance annotation: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; unmarked cells are not significant ($p \geq 0.05$). Blue shading denotes negative correlation; red shading denotes positive correlation.

Cross-dimension independence. The most critical finding concerns correlations *across* evaluation dimensions. **Novelty (NS)** shows no significant association with any flaw-related metric (Critical Recall: $r = -0.015$, $p = 0.40$; Minor Recall: $r = +0.021$, $p = 0.23$; nCPS: $r = +0.007$, $p = 0.70$), confirming that evaluating the novelty of reviewer claims is entirely decoupled from flaw detection ability. **DoA$_{\text{HM}}$** likewise exhibits no meaningful linear relationship with Critical Recall ($r = -0.006$, $p = 0.72$), demonstrating that structural argumentation depth is independent of a reviewer's capacity to identify methodological flaws. The correlation between DoA$_{\text{HM}}$ and MCS is statistically significant ($r = +0.094$, $p < 0.001$), yet the effect size remains negligible ($r^2 < 0.01$), confirming that argumentative depth and constructiveness constitute distinct dimensions.

Overall assessment. Seven of fifteen metric pairs show no statistically significant correlation ($p \geq 0.05$). All significant pairs have $|r| < 0.20$, placing them in the *negligible-to-small* range with shared variance below 4% ($r^2 < 0.04$). These results collectively confirm that the six metrics are **empirically near-orthogonal**: each captures a distinct dimension of peer review quality, thereby justifying their joint use as a comprehensive multi-dimensional evaluation benchmark.

E Full Cross-Dataset Quantitative Results

E.1 Statistical Significance Testing Protocol

To rigorously assess the performance differences between the LLM baselines and the human ground-truth, we conduct non-parametric statistical testing across all metrics. Given the non-normal distribution of the evaluation scores, we employ the **Wilcoxon signed-rank test** to compute the uncorrected p-values for paired comparisons.

Furthermore, to stringently control the Family-Wise Error Rate (FWER) while evaluating the cross-venue consistency of the models, we construct our hypothesis families based on model-metric pairs. Specifically, the **Holm-Bonferroni step-down correction** is applied independently for each LLM baseline within each specific evaluation dimension across the 3 conferences ($N = 3$ comparisons per family, corresponding to the 3 venues used throughout this paper: ICLR 2026, ICML 2025, and NeurIPS 2025). This approach ensures that any statistically significant result reflects a model's robust and consistent capability across different peer-review distributions, rather than an isolated success at a single venue.

Throughout the subsequent tables, we report the effect size (rank-biserial correlation r) and denote the Holm-corrected statistical significance using the following standard notation:

- **ns** (Not Significant): $p_{holm} \geq 0.05$
- $*$: $p_{holm} < 0.05$
- $**$: $p_{holm} < 0.01$
- $***$: $p_{holm} < 0.001$

Results marked as **ns** suggest that the model's performance is statistically indistinguishable from the human baseline, whereas the starred results indicate a robust difference that survives rigorous multiple-comparison correction.

E.2 Depth of Analysis

Table 5: Detailed Cross-Venue Performance for **Depth of Analysis**. Statistical significance is computed independently for each baseline across the 3 conferences to evaluate consistency.

Baselines	Depth of Analysis Scores (Mean ± Std)		
	ICLR 2026	**ICML 2025**	**NeurIPS 2025**
Human	0.497 ± 0.061	0.483 ± 0.063	0.500 ± 0.059
CycleReviewer	$\mathbf{0.479 \pm 0.121}^{\text{ns}}$	$0.476 \pm 0.145^{\text{ns}}$	$\mathbf{0.494 \pm 0.138}^{\text{ns}}$
DeepReview	$0.471 \pm 0.142^{\text{ns}}$	$\mathbf{0.482 \pm 0.126}^{\text{ns}}$	$0.490 \pm 0.128^{\text{ns}}$
Reviewer2	$0.346 \pm 0.146^{***}$	$0.363 \pm 0.135^{***}$	$0.379 \pm 0.128^{***}$
SEA	$0.382 \pm 0.161^{***}$	$0.403 \pm 0.139^{***}$	$0.407 \pm 0.152^{***}$
TreeReview	$0.335 \pm 0.115^{***}$	$0.335 \pm 0.123^{***}$	$0.346 \pm 0.111^{***}$

The granular DoA results presented in Table 5 highlight a stark contrast in the evidentiary capabilities of the evaluated baselines. TreeReview, SEA, and Reviewer2 yield consistently lower DoA scores across all three venues, with statistical significance ($p < 0.005$) underscoring their systematic deficiency in substantiating critiques compared to human reviewers. Conversely, CycleReviewer and DeepReview successfully bridge this gap. The consistent lack of statistical significance (^{ns}) when compared to the human baseline proves that these models achieve a comparable level of analytical depth. As analyzed previously, this statistical parity is primarily driven by their robust internal

grounding mechanisms and high premise ratios, which effectively compensate for the inherent limitations of standard LLM generation.

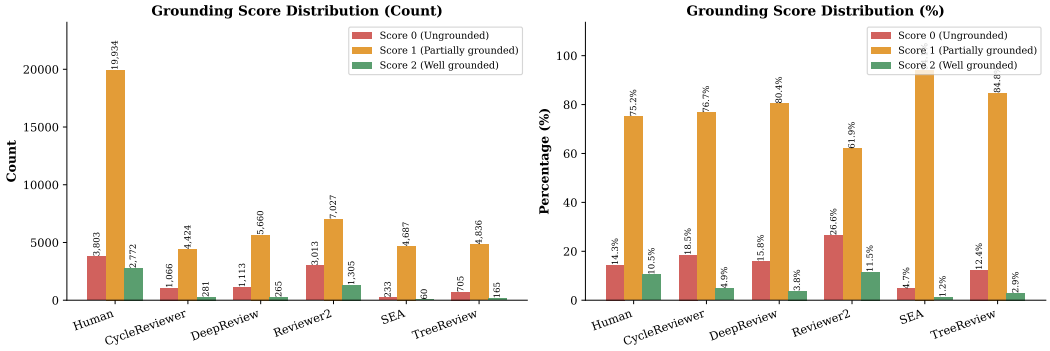

Figure 22: Absolute count (left) and percentage distribution (right) of extracted premises categorized by grounding scores for the human baseline and five evaluated LLMs. Score 0 represents ungrounded or vague premises, Score 1 indicates premises anchored to internal manuscript elements, and Score 2 denotes premises anchored to external literature or benchmarks.

As illustrated in Table 5 and Figure 22, the disparity in DoA scores stems from a complex interplay between the Premise Ratio (the consistency of providing justification) and the Grounding Score (the evidentiary quality of that justification).

The Illusion of Depth in Reviewer2. Reviewer2 presents an intriguing paradox at the grounding level. Despite producing an overwhelming absolute volume of vague, unanchored statements (Score 0), it surprisingly generates more externally grounded premises (Score 2) than other LLM baselines. However, as the per-aspect analysis in Figure 24 confirms, this marginal grounding advantage is entirely negated by its severely low Premise Ratio, which dilutes the overall analytical depth across all aspects.

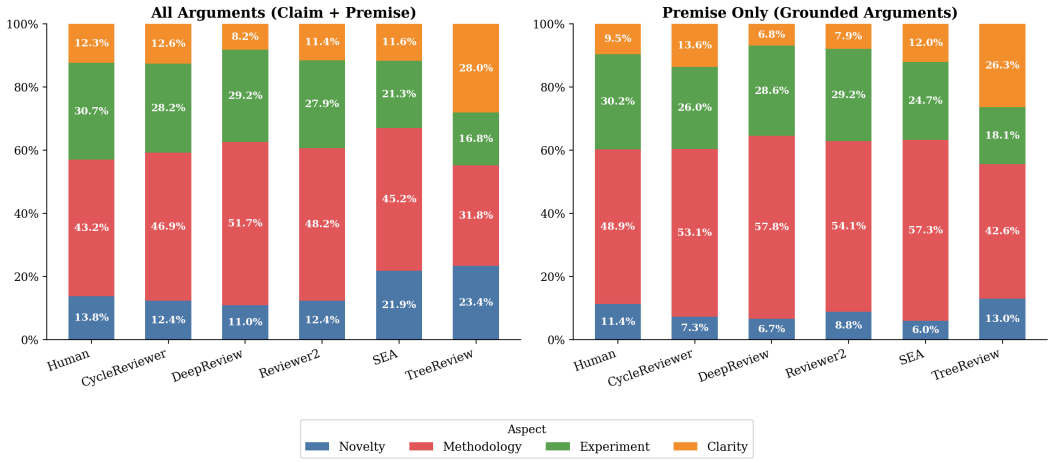

Figure 23: Distribution of review focus across 4 key aspects. The left panel shows the distribution across all extracted arguments, while the right panel highlights the distribution strictly for premises (grounded arguments).

Table 6: Macro-average aspect distribution (premise-level) and alignment with Human reviewers measured by Jensen-Shannon Distance (JSD). JSD $\in [0, 1]$; lower values indicate closer alignment to the Human aspect distribution. H (bits) denotes Shannon entropy of the macro-averaged aspect distribution (max $= \log_2 4 \approx 2$ bits).

Reviewer	Novelty	Methodology	Experiment	Clarity	JSD ↓	H (bits) ↑
Human	0.114	0.489	0.302	0.095	—	1.706
CycleReviewer	0.073	0.531	0.260	0.136	0.088	1.657
DeepReview	0.067	**0.578**	0.286	**0.068**	0.094	1.499
Reviewer2	0.088	0.541	0.292	0.079	**0.052**	1.596
SEA	0.060	0.573	0.247	0.120	0.108	1.569
TreeReview	**0.130**	0.426	0.181	**0.263**	0.206	1.860

As illustrated in Figure 23 and Table 6, analyzing the substantive distribution of review focus reveals critical divergences in how analytical effort is allocated across aspects. Fundamentally, both human experts and LLMs dedicate the largest proportion of their critiques to a paper's core technical components: Methodology (∼43%) and Experimental Design (∼31%), confirming a broad consensus on the most critical review dimensions. To quantify the degree of alignment between each LLM and Human reviewers, we compute the Jensen-Shannon Distance (JSD, the square root of the Jensen-Shannon Divergence) between each system's macro-averaged premise-level aspect distribution and the Human distribution. JSD $\in [0, 1]$, where 0 indicates identical distributions.

Alignment with Human Priorities. Most advanced baselines successfully mirror human intuitive focus, dedicating the largest proportion of their grounded premises to core technical components: Methodology ($\sim 43 - 58\%$) and Experimental Design ($\sim 18 - 29\%$). Notably, **Reviewer2** achieves the closest alignment to Human reviewers with the lowest JSD of 0.052. Its premise distribution (Methodology 54.1%, Experiment 29.2%, Clarity 7.9%) closely traces the human pattern (48.9%, 30.2%, 9.5%), demonstrating a well-calibrated allocation of critical effort. **DeepReview** achieves the lowest Clarity proportion (6.8%) and the highest Methodology concentration (57.8%). While this results in the lowest Shannon entropy ($H = 1.499$ bits)—indicating a narrower, highly specialized focus—it confirms the model's capacity to firmly anchor its feedback in the most critical dimensions of the submission.

The Surface-Level Trap. Rather than being an inherent LLM limitation, the "surface-level trap" manifests when automated frameworks lack explicit, domain-specific evaluation constraints. Remarkably, even highly structured pipelines can fall into this trap. This is starkly pronounced in **TreeReview**, which, despite its complex reasoning topology, allocates an excessive 26.3% of its premise-level effort to Clarity, nearly $2.8\times$ the proportion of Human reviewers (9.5%). Consequently, TreeReview records the highest JSD against Humans (0.206) and the lowest Methodology coverage (42.6%), confirming that without strict dimensional guidance, its analytical distribution naturally diverges toward superficial nitpicking.

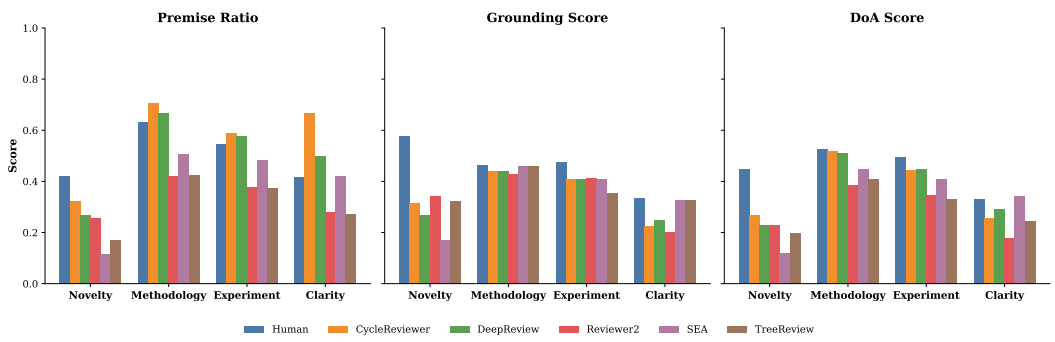

Figure 24: Per-aspect Depth of Analysis scores for Human and five LLM baselines, averaged across the three conferences. Each group corresponds to one of the four review aspects (Novelty & Related Work, Methodology, Experimental Design, Clarity)

Figure 24 decomposes the overall DoA score into its four constituent aspects—Novelty & Related Work, Methodology, Experimental Design, and Clarity—revealing the underlying drivers of the scores reported in Figure 22 and Table 6. For each aspect, the harmonic mean of Premise Ratio and Grounding Score is reported per reviewer, allowing a fine-grained comparison of *where* and *how deeply* each system substantiates its arguments.

Shared Prioritization of Core Technical Aspects. A consistent pattern emerges across both human and LLM reviewers: DoA scores are substantially higher for Methodology and Experimental Design than for Novelty and Clarity. For Human reviewers, the Methodology aspect achieves the highest DoA Score (0.527 ± 0.067), followed closely by Experiment (0.494 ± 0.090), while Novelty (0.447 ± 0.207) and Clarity (0.332 ± 0.189) trail significantly behind. This pattern holds uniformly across all five baselines, confirming that both humans and LLMs inherently recognize the need to anchor their most substantive arguments in the methodological and experimental core of a paper. The effect is directly traceable to elevated Premise Ratios and Grounding Scores in these two aspects: for instance, Human reviewers achieve a Premise Ratio of 0.630 on Methodology versus only 0.417 on Clarity, indicating that technical claims are far more consistently backed by evidence than presentation-level observations.

Evidence Density as the Key Differentiator Across Baselines. The second insight concerns the systematic gap between baselines. Across all four aspects, CycleReviewer and DeepReview consistently achieve DoA scores closest to—and in several cases statistically indistinguishable from—Human reviewers. This cross-aspect parity is not coincidental: both systems maintain the highest Premise Ratios among LLM baselines across every aspect (e.g., CycleReviewer reaches 0.707 and DeepReview 0.668 on Methodology, both exceeding the Human value of 0.630), demonstrating that their superior overall DoA is driven by a systematic tendency to substantiate claims with grounded evidence regardless of the aspect under discussion. In contrast, Reviewer2, SEA, and TreeReview show markedly lower Premise Ratios particularly on Novelty (0.255, 0.114, 0.171 respectively), producing the steepest per-aspect DoA drops and confirming that their aggregate weakness is not confined to any single dimension but reflects a globally deficient evidentiary discipline.

Summary of Analytical Depth. In conclusion, achieving a human-level Depth of Analysis requires more than merely generating a high volume of text. Models that fall into the trap of unsupported verbosity (Reviewer2) or surface-level nitpicking (TreeReview) are severely penalized. To bridge the analytical gap, automated reviewers must systematically substantiate their claims and strictly prioritize core technical dimensions over formatting issues.

E.3 NOVELTY ASSESSMENT

Table 7: Detailed Cross-Venue Performance for **Novelty Assessment**. Statistical significance is computed independently for each baseline across the 3 conferences to evaluate consistency.

Baselines	Novelty Assessment Scores (Mean ± Std)		
	ICLR 2026	**ICML 2025**	**NeurIPS 2025**
Human	0.764 ± 0.195	0.833 ± 0.189	0.814 ± 0.183
CycleReviewer	$0.793 \pm 0.219^{\text{ns}}$	$0.781 \pm 0.210^{*}$	$0.793 \pm 0.228^{\text{ns}}$
DeepReview	$0.737 \pm 0.220^{\text{ns}}$	$0.785 \pm 0.188^{**}$	$0.784 \pm 0.203^{*}$
Reviewer2	$0.792 \pm 0.215^{\text{ns}}$	$0.792 \pm 0.216^{\text{ns}}$	$0.815 \pm 0.195^{\text{ns}}$
SEA	$\mathbf{0.831 \pm 0.184}^{***}$	$\mathbf{0.870 \pm 0.187}^{*}$	$\mathbf{0.860 \pm 0.202}^{**}$
TreeReview	$0.813 \pm 0.201^{**}$	$0.808 \pm 0.216^{\text{ns}}$	$0.818 \pm 0.197^{\text{ns}}$

Table 7 reports the cross-venue scalar *Novelty Assessment* scores. These values should be interpreted carefully: they do not certify a manuscript's objective novelty, but rather measure whether the novelty claims expressed in a review can be grounded in retrievable prior work under the PRISM pipeline. Across almost all venues, mean scores fall between 0.737 and 0.870, indicating that both human reviewers and automated baselines frequently produce novelty claims that the retrieval-and-verification procedure can resolve with substantial evidence. The non-significant differences (^{ns}) for Reviewer2 in all three venues and TreeReview in two of three therefore suggest similar *evidence-grounding performance on this scalar metric*, not full claim-by-claim agreement with human reviewers.

The Outperformance of SEA. The most notable result on this scalar metric is the performance of the **SEA** baseline. While SEA is not uniformly strongest on the other review dimensions, it achieves the highest novelty-assessment score in all three venues and significantly exceeds the human baseline in ICLR 2026 ($p_{Holm} < 0.001$), ICML 2025 ($p_{Holm} < 0.05$), and NeurIPS 2025 ($p_{Holm} < 0.01$). Within the interpretation above, this suggests that SEA's structured generation style tends to produce novelty claims that are especially easy for the PRISM retrieval-and-verification pipeline to ground in prior work. As the agreement analysis below will show, this should not be conflated with universally stronger alignment to human novelty judgments.

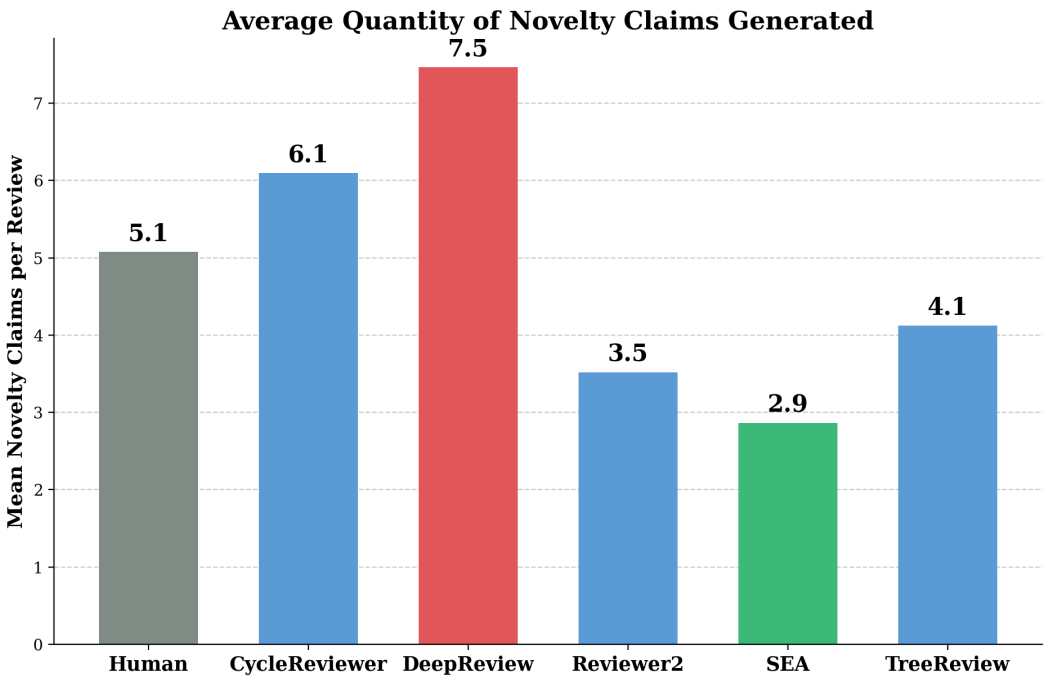

Figure 25: Average number of novelty claims generated per review across the human and LLM Reviewers.

Claim Volume and Generation Verbosity. Figure 25 reveals another important difference that scalar novelty scores alone do not capture: review sources vary substantially in how many novelty claims they choose to make. Because the human input concatenates multiple independent reviews for the same paper, raw human claim counts must be normalized by the number of individual reviews before they are interpreted as per-review verbosity. In the unnormalized benchmark output, the concatenated human-review bundle averages 5.1 extracted novelty-related claims per paper; the benchmark code therefore reports the human claim-volume statistic as claims per individual human review whenever review-count metadata is available. **DeepReview** is markedly more verbose at 7.5 claims per generated review, reflecting a finer-grained decomposition of contributions and comparisons. By contrast, **SEA** (2.9 claims) and **Reviewer2** (3.5 claims) are much more conservative, concentrating their novelty discussion into fewer statements. This matters because downstream agreement depends not only on how claims are scored once extracted, but also on claim granularity and boundary choices at the extraction stage.

Summary. Taken together, these results show that LLM reviewers can produce novelty claims that are often well grounded under the PRISM retrieval-and-verification pipeline. However, this scalar score should not be read as evidence that LLMs can certify the objective novelty of a manuscript, nor that they fully replicate human novelty judgments. The next subsection addresses the harder question: once all review sources are normalized through the same pipeline, do they actually arrive at the same claim-level novelty conclusions?

E.4 Flaw Identification & Prioritization of Major Issue

Table 8: Detailed Cross-Venue Performance for **Flaw Identification & Prioritization**. Statistical significance is computed independently for each baseline across the 3 conferences to evaluate consistency.

Metric / Baselines	Scores (Mean ± Std)		
	ICLR 2026	**ICML 2025**	**NeurIPS 2025**
Critical Flaw Identification			
Human	0.327 ± 0.149	0.299 ± 0.167	0.306 ± 0.159
CycleReviewer	$0.186 \pm 0.263^{***}$	$0.254 \pm 0.319^{***}$	$0.232 \pm 0.302^{***}$
DeepReview	0.309 ± 0.291^{ns}	0.360 ± 0.351^{ns}	0.383 ± 0.335^{ns}
Reviewer2	$\mathbf{0.593 \pm 0.317}^{***}$	$\mathbf{0.683 \pm 0.304}^{***}$	$\mathbf{0.657 \pm 0.290}^{***}$
SEA	$0.174 \pm 0.237^{***}$	$0.219 \pm 0.259^{***}$	$0.201 \pm 0.262^{***}$
TreeReview	$0.216 \pm 0.281^{***}$	$0.254 \pm 0.333^{**}$	$0.265 \pm 0.298^{**}$
Minor Flaw Identification			
Human	0.250 ± 0.080	0.262 ± 0.085	0.248 ± 0.073
CycleReviewer	$0.186 \pm 0.151^{***}$	$0.195 \pm 0.151^{***}$	$0.154 \pm 0.120^{***}$
DeepReview	0.256 ± 0.147^{ns}	$0.233 \pm 0.143^{***}$	$0.203 \pm 0.135^{***}$
Reviewer2	$\mathbf{0.467 \pm 0.185}^{***}$	$\mathbf{0.480 \pm 0.182}^{***}$	$\mathbf{0.463 \pm 0.179}^{***}$
SEA	$0.224 \pm 0.124^{***}$	$0.207 \pm 0.114^{***}$	$0.193 \pm 0.116^{***}$
TreeReview	$0.339 \pm 0.147^{***}$	$0.303 \pm 0.135^{***}$	$0.306 \pm 0.143^{***}$
Prioritization of Major Issue (nCPS)			
Human	0.969 ± 0.047	0.984 ± 0.047	0.973 ± 0.034
CycleReviewer	$0.969 \pm 0.110^{*}$	0.970 ± 0.108^{ns}	$0.968 \pm 0.127^{*}$
DeepReview	0.968 ± 0.053^{ns}	0.978 ± 0.048^{ns}	0.956 ± 0.110^{ns}
Reviewer2	$0.970 \pm 0.078^{*}$	$0.975 \pm 0.033^{***}$	0.970 ± 0.033^{ns}
SEA	$\mathbf{0.986 \pm 0.037}^{***}$	$\mathbf{0.984 \pm 0.038}^{ns}$	$\mathbf{0.981 \pm 0.080}^{**}$
TreeReview	0.977 ± 0.044^{ns}	0.980 ± 0.042^{ns}	0.975 ± 0.046^{ns}

The detailed results in Table 8 unequivocally confirm the exhaustive diagnostic capability of Reviewer2. Across every single evaluated venue, Reviewer2 consistently achieves the highest Recall for both *Critical* (ranging from 0.593 to 0.683) and *Minor* flaws. More importantly, the statistical tests ($p_{Holm} < 0.005$, denoted as ***) validate that this over-performance relative to the human baseline is structurally ingrained in the model's generation style, not merely a statistical artifact of a specific dataset. This elevated Recall is directly correlated with its overall generation volume: as previously illustrated in the valid flaw counts (Figure 6 - main text), Reviewer2 acts as a high-volume "flaw scanner," extracting an unprecedented absolute number of valid issues per review. By casting a wider diagnostic net, it naturally captures a higher proportion of both fatal methodologies and minor anomalies compared to the more conservative human baseline. Conversely, other LLM baselines generally exhibit a diagnostic deficit compared to human experts. Models like CycleReviewer and SEA consistently underperform the human ground-truth in extracting both major and minor flaws across all three venues. An interesting anomaly, however, is observed in **TreeReview**. While it struggles to detect fatal methodological errors (consistently scoring lower than humans in Critical Recall), it consistently outperforms humans in *Minor Flaw Identification* (ranging from 0.303 at

ICML 2025 to 0.339 at ICLR 2026, all with strong statistical significance). This further corroborates our earlier finding that TreeReview suffers from a "surface-level trap," heavily over-indexing its analytical effort on presentation and formatting anomalies rather than scientific rigor. DeepReview, maintaining a highly conservative profile, yields scores that are statistically indistinguishable (^{ns}) from human experts in every venue on Critical Recall, demonstrating a strong alignment with human reviewing patterns.

Concluding the table analysis, the granular breakdown of the Critique Prioritization Score ($nCPS$) solidifies a key macro-level observation: the ability to strategically rank flaws is near ceiling for modern LLMs. Across all three conferences, every baseline achieves near-perfect scores ($nCPS > 0.95$). Regardless of their diagnostic capability, the automated systems' structural organization adheres to the established academic norm of surfacing major flaws first, with every remaining difference from the human ground-truth below 0.02.

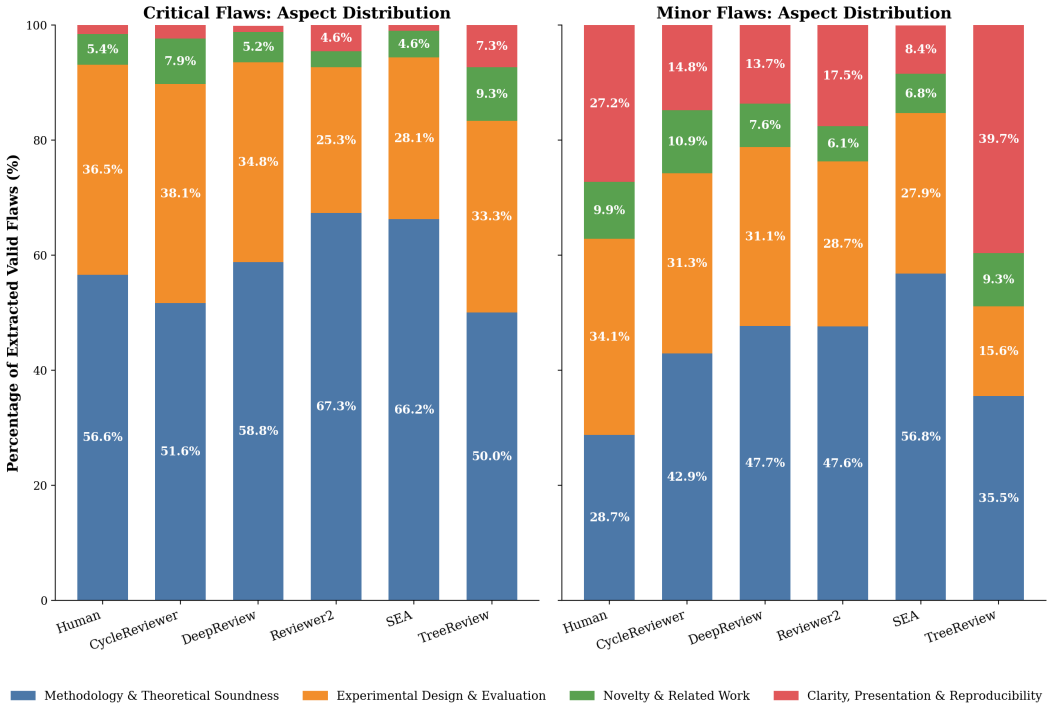

Figure 26: Comparative Aspect Topic Distribution of extracted valid flaws, stratified by severity (*Critical* vs. *Minor*).

To gain a deeper understanding of reviewer behavior, we stratify the aspect topic distribution by flaw severity (Critical vs. Minor), as visualized in Figure 26. This decomposition reveals profound insights into the contextual adaptability—or lack thereof—within LLM baselines compared to human experts.

Alignment on Critical Vulnerabilities. When evaluating *Critical* flaws, human experts demonstrate a laser-focused approach, allocating an overwhelming 93.1% of their critiques to core technical components (56.6% on Methodology and 36.5% on Experimental Design). Formatting issues (Clarity) account for merely 1.5% of critical flaws. **DeepReview** exhibits a remarkably identical signature, dedicating 93.5% of its critical flaw detection to Methodology and Experiments, with only 1.1% wasted on Clarity. This confirms that at the highest severity level, well-calibrated models like DeepReview successfully emulate the human capacity to isolate fatal scientific errors. Conversely, **TreeReview** assigns 7.3% of its critical flaws to Clarity, consistent with its broader Clarity skew.

The Adaptive Focus on Minor Flaws. The distribution of *Minor* flaws reveals a compelling alignment between human experts and LLM baselines. Unlike critical flaws, which are overwhelmingly technical, the cognitive focus for minor flaws naturally shifts toward surface-level anomalies. Both human reviewers and LLMs significantly reduce their extreme scrutiny on Methodology, while substantially increasing their attention to *Clarity, Presentation & Reproducibility*. For instance, the human focus on Clarity surges to 27.2% for minor flaws (up from a mere 1.5% in critical flaws). Remarkably, the automated baselines successfully mirror this contextual shift. **TreeReview** captures a massive 39.7% of clarity-related minor issues, while even verbose models like **Reviewer2** nearly quadruple their attention to presentation (17.5%, up from 4.6%) compared to their critical flaw distribution.

Emulating Human Flaw Categorization. This dynamic shift demonstrates that LLMs are not rigidly applying a single evaluation template. Instead, they contextually adapt the nature of their critiques based on severity. While they still capture minor methodological nitpicks (as seen in DeepReview and SEA), they correctly associate a large portion of minor flaws with presentation errors, directly aligning with how human experts intuitively classify minor review points.

Summary. These findings highlight a sophisticated capability in modern automated peer review: LLMs can successfully mimic human intuition in identifying and categorizing different types of flaws. By instinctively mapping core technical breakdowns to *Critical* severity and correctly shifting their lens toward presentation issues for *Minor* anomalies, LLMs prove they possess a nuanced, human-like understanding of manuscript evaluation.

E.5 Multi-Dimensional Constructiveness

Table 9: Detailed Cross-Venue Performance for **Constructiveness Score (MCS)**. Statistical significance is computed independently for each baseline across the 3 conferences to evaluate consistency.

Baselines	Constructiveness Scores (Mean ± Std)		
	ICLR 2026	**ICML 2025**	**NeurIPS 2025**
Human	0.601 ± 0.063	0.521 ± 0.070	0.546 ± 0.057
CycleReviewer	$0.540 \pm 0.120^{***}$	0.523 ± 0.095^{ns}	$0.527 \pm 0.119^{*}$
DeepReview	$\mathbf{0.636 \pm 0.091}^{***}$	$\mathbf{0.638 \pm 0.085}^{***}$	$\mathbf{0.635 \pm 0.087}^{***}$
Reviewer2	0.587 ± 0.100^{ns}	$0.584 \pm 0.101^{***}$	$0.586 \pm 0.087^{***}$
SEA	$0.502 \pm 0.098^{***}$	$0.487 \pm 0.084^{***}$	$0.485 \pm 0.070^{***}$
TreeReview	$0.490 \pm 0.144^{***}$	$0.498 \pm 0.125^{**}$	$0.496 \pm 0.120^{***}$

Table 9 presents the granular breakdown of the MCS and its sub-dimensions across the three evaluated conference venues. A longitudinal analysis reveals that while the constructiveness of both human and automated reviewers fluctuates slightly across different venues, DeepReview consistently maintains a substantial lead. Specifically, DeepReview achieves the highest MCS across all three conferences (ranging from 0.635 at NeurIPS 2025 to 0.638 at ICML 2025). Importantly, Holm's post-hoc tests confirm that DeepReview's superiority over the human baseline (0.521 at ICML 2025 to 0.601 at ICLR 2026) is statistically significant ($p_{Holm} < 0.05$) in every venue evaluated. Furthermore, Reviewer2 demonstrates competitive performance, particularly at ICML 2025 and NeurIPS 2025, where its MCS scores (0.584 and 0.586) significantly exceed the human ground-truth. In stark contrast, models such as SEA and TreeReview consistently underperform compared to the human baseline across all three venues, highlighting a persistent limitation in their capacity to formulate genuinely constructive critiques.

Table 10: Detailed Constructiveness Sub-dimensions (D1-D5) across the 3 Conferences evaluated on a raw scale of $[0, 2]$.

System	D1: Actionability	D2: Specificity	D3: Justification	D4: Solution	D5: Tone
Human	1.026 ± 0.270	1.732 ± 0.163	0.770 ± 0.259	0.427 ± 0.202	1.605 ± 0.230
CycleReviewer	1.335 ± 0.511	$\mathbf{1.894} \pm 0.211$	0.326 ± 0.452	0.410 ± 0.395	1.335 ± 0.413
DeepReview	$\mathbf{1.422} \pm 0.299$	1.832 ± 0.202	0.598 ± 0.381	$\mathbf{0.781} \pm 0.284$	$\mathbf{1.730} \pm 0.290$
Reviewer2	1.226 ± 0.324	1.826 ± 0.222	$\mathbf{0.969} \pm 0.447$	0.249 ± 0.259	1.586 ± 0.401
SEA	0.819 ± 0.395	1.666 ± 0.248	0.484 ± 0.413	0.333 ± 0.259	1.613 ± 0.266
TreeReview	1.045 ± 0.283	1.581 ± 0.354	0.627 ± 0.487	0.351 ± 0.382	1.340 ± 0.556

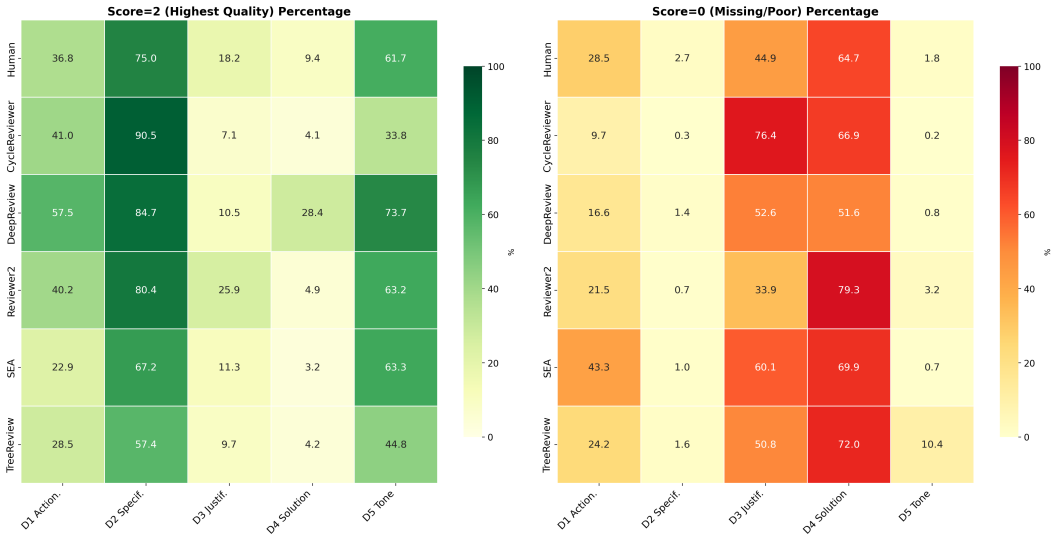

Figure 27: Heatmap detailing the performance across the five core sub-metrics of constructiveness (Actionability, Specificity, Justification, Solution, and Tone).

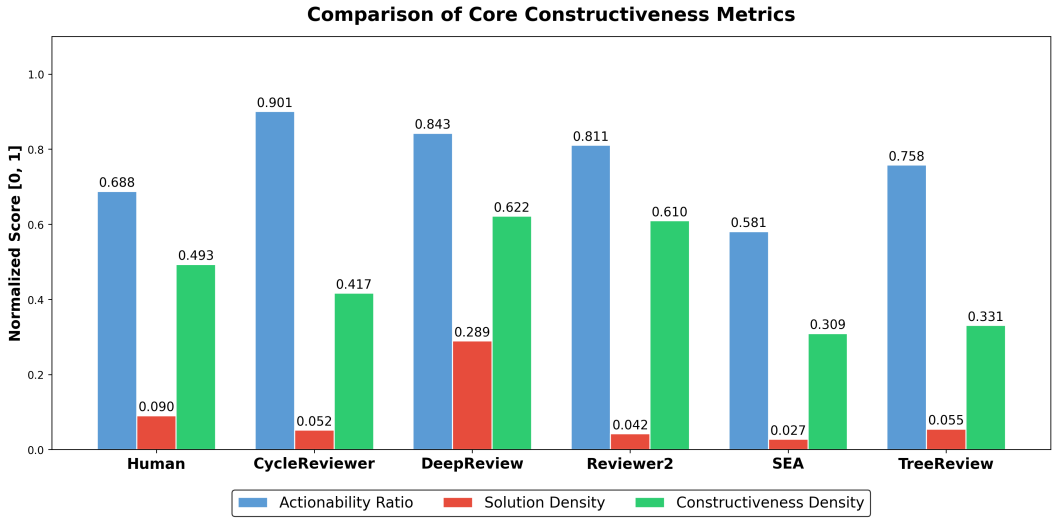

Figure 28: Comparison of core constructiveness metrics (Actionability Ratio, Solution Density, and Constructiveness Density).

To provide a more nuanced evaluation, we derive three auxiliary density metrics from the five core dimensions ($D_1 - D_5$). Let $\mathbb{I}(\cdot)$ be an indicator function. We define: (i) the **Actionability Ratio** (AR) as the proportion of comments providing at least a general direction ($D_1 \geq 1$); (ii) the **Solution Density** (SD) as the percentage of comments offering explicit, implementable fixes ($D_4 = 2$); and (iii) the **Constructiveness Density** (CD) as the proportion of comments achieving high-quality structural alignment ($CLC \geq 0.5$). Formally:

$$AR(R) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(D_1(c_j) \geq 1), \tag{3}$$

$$SD(R) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(D_4(c_j) = 2), \tag{4}$$

$$CD(R) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(CLC(c_j) \geq 0.5) \tag{5}$$

As presented in the macro-averaged results (Figure 27) and the density metrics (Figure 28), a granular examination of these metrics reveals distinct behavioral signatures distinguishing human reviewers from LLM baselines.

The Solution Bottleneck. The most profound divergence between human experts and top LLMs lies in the capacity to propose explicit improvements. The data reveals a systemic limitation in traditional peer review: while humans are highly proficient at pinpointing concrete flaws (D_2 Specificity ≈ 1.72), they frequently fail to provide actionable remedies. This is evidenced by the human baseline's remarkably low Solution Density ($SD \approx 0.090$), indicating that only 9% of human comments contain an explicit fix ($D_4 = 2$), though this is still second only to DeepReview. In stark contrast, DeepReview fundamentally alters this paradigm. It consistently achieves an SD approaching 0.29 and D_4 scores near 0.80 across all venues. DeepReview excels because it does not merely diagnose problems; it proactively prescribes explicit, implementable solutions, thereby maximizing its utility to the authors.

The Verbosity Paradox of Reviewer2. The sub-dimensional breakdown elegantly explains Reviewer2's deceptively high overall scores. Reviewer2 achieves exceptional Justification scores

($D_3 > 0.90$) and a strong Constructiveness Density ($CD \approx 0.61$), largely as an artifact of its highly verbose, two-stage rubric-driven generation style. It excels at providing extensive reasoning for its claims. However, its abysmal Solution Density ($SD \approx 0.04$) and low Solution score ($D_4 \approx 0.25$) reveal a critical flaw: it exhaustively explains *why* something is wrong but almost never tells the author *how* to fix it, rendering its voluminous critiques practically inert.

Actionability without Depth in CycleReviewer. Conversely, CycleReviewer exhibits a contrasting failure mode. It achieves the highest Specificity ($D_2 \approx 1.89$) and Actionability Ratio ($AR > 0.90$), meaning almost all of its comments reference concrete paper elements and offer at least a general direction ($D_1 \geq 1$). Yet, it suffers a catastrophic drop in Justification ($D_3 \approx 0.32$) and Solution Density ($SD \approx 0.05$). This signature points to a shallow, "checklist-style" reviewing behavior: it successfully targets specific sections with general demands (e.g., "needs more baselines"), but fails to provide the evidentiary backing or the explicit fixes required of a profound scientific critique.

Tone and Professionalism. Finally, the Tone dimension (D_5) confirms that well-calibrated LLMs can systematically elevate the discourse of peer review. **DeepReview** ($D_5 \approx 1.73$) consistently outputs more professional, neutral, and encouraging feedback than the human baseline ($D_5 \approx 1.61$), effectively mitigating the dismissive or hostile language occasionally encountered in human peer reviews.

Summary of Constructiveness. This multi-dimensional analysis highlights a fundamental difference in reviewing paradigms. Human experts primarily act as *diagnosticians*, highly effective at pinpointing errors but lacking in actionable guidance. LLM baselines often mask this deficiency with extreme verbosity (Reviewer2) or superficial demands (CycleReviewer). In contrast, **DeepReview** transcends these limitations, acting more as a *collaborator* by bridging the critical gap between identifying flaws (D_2) and formulating explicit, professionally toned solutions (D_4, D_5), it represents a significant step toward genuinely constructive automated peer review.

E.6 Review Sensitivity to Paper Quality: Accept vs. Reject Analysis

Statistical Methodology. To evaluate whether the generated reviews effectively distinguish between high-quality and low-quality submissions, we compare the metric scores assigned to accepted versus rejected papers (Figure 29). Specifically, we apply the Mann-Whitney U test (two-sided, unpaired) to compare per-paper metric scores between accepted and rejected papers for each reviewer baseline. To control the family-wise error rate within each reviewer, p-values are adjusted using the Holm-Bonferroni procedure across all metrics tested for that reviewer. The mean score differences ($\Delta = \text{Accept} - \text{Reject}$) and their significance levels are detailed in Table 11.

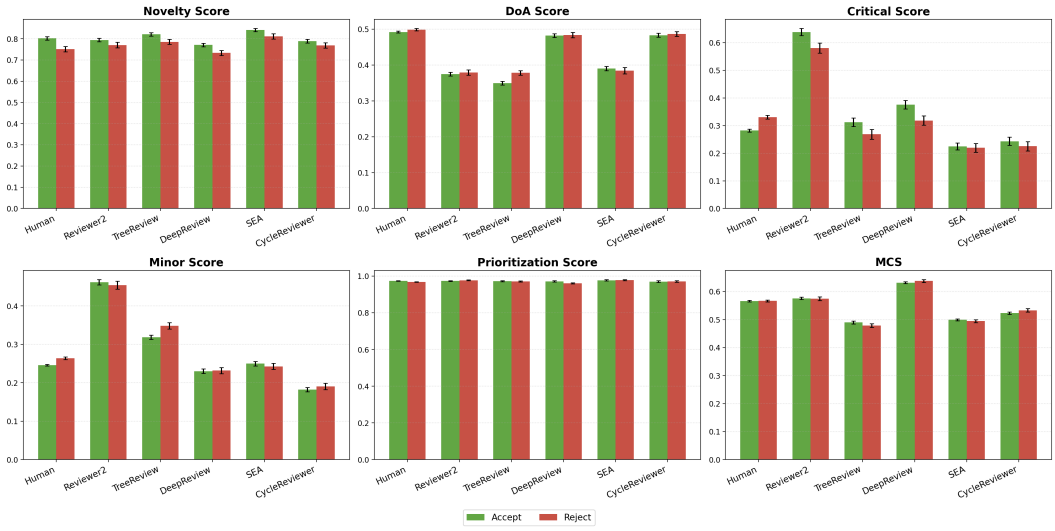

Figure 29: Mean scores of Human and LLM reviewers on six evaluation dimensions, stratified by paper decision (Accept vs. Reject)

Table 11: Metric differences across reviewers. Bold values indicate statistically significant results; (ns) denotes non-significant.

Metric	Human	Reviewer2	TreeReview	DeepReview	SEA	CycleReviewer
Novelty Score	$+\mathbf{0.051}^{*}$	$+0.024^{\text{ns}}$	$+0.035^{\text{ns}}$	$+0.037^{\text{ns}}$	$+0.029^{\text{ns}}$	$+0.021^{\text{ns}}$
DoA Score	-0.006^{ns}	-0.004^{ns}	-0.028^{ns}	-0.001^{ns}	-0.006^{ns}	-0.003^{ns}
Critical Score	$\mathbf{-0.049}^{***}$	$+0.058^{\text{ns}}$	$+0.043^{\text{ns}}$	$+0.057^{\text{ns}}$	$+0.005^{\text{ns}}$	$+0.018^{\text{ns}}$
Minor Score	$\mathbf{-0.018}^{**}$	-0.008^{ns}	-0.030^{ns}	-0.002^{ns}	-0.007^{ns}	-0.004^{ns}
Prioritization Score	$+\mathbf{0.006}^{***}$	-0.004^{ns}	$+0.001^{\text{ns}}$	$+\mathbf{0.011}^{***}$	-0.001^{ns}	-0.001^{ns}
MCS	-0.001^{ns}	$+0.001^{\text{ns}}$	$+0.011^{\text{ns}}$	-0.006^{ns}	$+0.004^{\text{ns}}$	-0.010^{ns}

Human Reviews Exhibit Strong Predictive Validity. Although human reviewers evaluate manuscripts blindly without any knowledge of the final editorial outcome, their assessments exhibit a robust, statistically significant correlation with the eventual decisions. Specifically, manuscripts that are ultimately accepted garner blind reviews with significantly higher Novelty Scores ($\Delta = +0.051^{*}$) and tighter structural prioritization ($\Delta_{\text{Prior.}} = +0.006^{***}$). Conversely, reviews of papers that are ultimately rejected recover significantly more of the consensus Critical ($\Delta = -0.049^{***}$) and Minor ($\Delta = -0.018^{**}$) flaws. The magnitude of the Critical score difference is notably $2.7\times$ larger than that of the Minor score. This confirms that human reviewers naturally calibrate the severity and focus of their critiques according to the inherent scientific merit of the manuscript, providing a valid signal that organically drives the final accept/reject decision.

LLMs Exhibit Evaluative Invariance Across Quality Tiers. In contrast to human reviewers, whose critiques strongly correlate with the final editorial outcome, LLM reviewers exhibit a highly invariant and consistent evaluative pattern regardless of whether a paper is ultimately accepted or rejected. Across all five automated systems, the metric differences are predominantly uniform, with only DeepReview's Prioritization Score reaching statistical significance ($\Delta = +0.011^{***}$). While the LLMs demonstrate a slight directional alignment with humans on interpretive dimensions—yielding marginally positive Δ for Novelty and negative Δ for Minor flaws—these shifts lack statistical power. Notably, this stability is most pronounced in the *Critical Score* dimension. Rather than being swayed by the overall quality of the submission, LLMs apply their internal diagnostic heuristics

independently: Reviewer2 ($+0.058^{\text{ns}}$), DeepReview ($+0.057^{\text{ns}}$), and TreeReview ($+0.043^{\text{ns}}$) swing inversely to humans, while SEA and CycleReviewer remain stable near zero.

Summary of Evaluative Invariance. The results in Table 11 highlight two fundamentally different review paradigms. Human reviewers display high sensitivity, naturally adjusting the severity of their critiques based on the overall scientific merit of the submission. In contrast, LLMs function as invariant diagnostic scanners. They generate reviews with remarkably stable metric distributions, applying a uniform evaluative standard across all papers. While this invariant behavior suggests that LLMs are not susceptible to the "halo effect" of high-quality submissions, it also implies a trade-off: their strict consistency limits their capacity to organically discriminate and highlight fatal methodological flaws with the same adaptive precision as human experts.

E.7 Evaluator Robustness Across LLM Backends

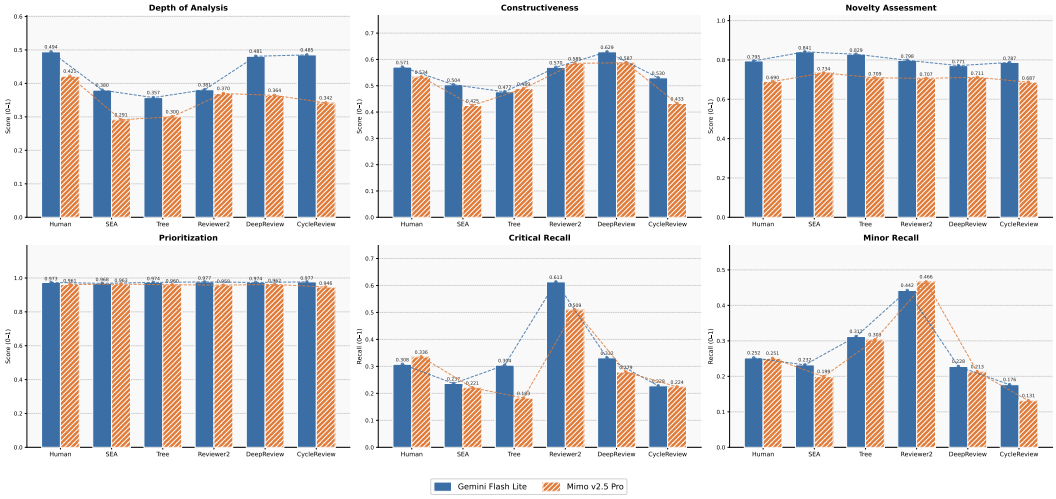

Figure 30: Comparison of all evaluation metrics between Gemini Flash Lite and Mimo v2.5 Pro as evaluator backends, aggregated over 150 papers (50 per conference: ICLR 2026, ICML 2025, NeurIPS 2025). Dashed lines indicate cross-group trends for each model.

To assess whether our metric framework is sensitive to the choice of evaluator LLM, we re-ran the full evaluation pipeline using Mimo v2.5 Pro Xiaomi MiMo Team (2026) as an alternative backend and compared results against Gemini 2.5 Flash Lite across all six metrics for our four aspects (Figure 30).

Moreover, Gemini generally assigns higher values, with Mimo scoring higher only for Reviewer2's Minor Recall (0.466 vs. 0.442), TreeReview's and Reviewer2's Constructiveness (0.489 vs. 0.477; 0.585 vs. 0.570), and the human Critical Recall (0.336 vs. 0.308). Importantly, Prioritization score shows the smallest divergence between evaluators, confirming that critical-flaw prioritization is effectively evaluator-agnostic.

Despite absolute score offsets, the relative ordering of reviewer types is largely consistent between the two evaluators. Reviewer2 and DeepReview consistently achieve the highest Constructiveness (MCS: Gemini 0.570/0.629 vs. Mimo 0.585/0.587) and Critical Recall (Gemini 0.613/0.332 vs. Mimo 0.509/0.279), while SEA and TreeReview rank lowest on Depth of Analysis and Constructiveness. For Novelty Assessment, both evaluators agree that SEA produces the highest scores (Gemini 0.841, Mimo 0.734). The main exception is Depth of Analysis, where under Mimo Reviewer2 (0.370) edges out DeepReview (0.364) and CycleReviewer (0.342), and all LLMs fall further below the human baseline (0.421). Apart from this ordering, the conclusions drawn from Gemini hold under Mimo, confirming that the observed performance gaps reflect genuine differences in review quality rather than artifacts of the specific evaluator model.

F Qualitative Analysis & Case Studies

F.1 Depth of Analysis

Case 1: The Evidentiary Collapse — Why Claim-Heavy Reviewers Fail. *Paper: NV-Embed: Generalist Text Embeddings from Decoder-Only LLMs*

Context. NV-Embed proposes a generalist embedding model built on decoder-only LLMs, introducing (1) a latent attention layer replacing mean-pooling, and (2) a two-stage contrastive instruction-tuning pipeline. The model achieves top-1 performance on the MTEB benchmark.

Table 12: Per-reviewer DoA statistics for NV-Embed. Avg GS is the mean grounding score normalized to $[0, 1]$ (raw scores 0/1/2 divided by 2)

Reviewer	DoA_{HM}	$\text{R}_{\text{premise}}$	Avg GS	Total Args	Premises	Claims
Human	0.574	0.673	0.500	55	37	18
DeepReview	**0.626**	0.733	0.546	15	11	4
Reviewer2	0.178	0.152	0.215	46	7	39

Human: Dense Technical Grounding Across All Aspects. Human reviewers produce 37 premises from 55 total arguments ($R_{\text{premise}} = 0.673$), with strong grounding quality ($\overline{\text{GS}} = 0.500$). Premises are component-specific, naming the paper's actual technical building blocks rather than describing them in generalities:

> *[GS = 1.0, Methodology]* "The techniques used are: 1. latent attention layer that achieves better pooling/combination of the last layer embeddings."

> *[GS = 1.0, Methodology]* "2. a two-stage contrastive instruction tuning method. First step tuning with in-batch negative and hard negative on retrieval datasets, and the second step tuning on non-retrieval datasets."

> *[GS = 2.0, Experiment]* "The model achieves top performance on the MTEB benchmark."

Human reviewers also raise critical, component-targeted observations, such as questioning whether a single innovative algorithmic piece exists beyond the two engineering contributions. This combination of evidential support and critical evaluation defines the human gold standard on this paper.

DeepReview: Exceeds Human DoA via Precision-to-Volume Economy. DeepReview produces only 15 arguments, yet 11 are premises ($R_{\text{premise}} = 0.733$, exceeding Human's 0.673), with the highest average grounding score of any reviewer ($\overline{\text{GS}} = 0.546$). Its premises are architecturally precise:

> *[GS = 1.0, Methodology]* "The authors introduce a novel latent attention layer for pooling embeddings, which outperforms traditional methods like average pooling and <EOS> token embedding."

> *[GS = 2.0, Experiment]* "Ablation experiments in Table 2 confirm the contribution of the latent attention layer over alternative pooling strategies."

Reviewer2: Self-Referential Grounding and Volume Inflation. Reviewer2 generates 46 arguments but only 7 qualify as premises, and **4 of those 7 carry grounding score 0** ($\overline{\text{GS}} = 0.215$, barely above the minimum). The 39 claims are structured section-by-section summaries offering no independent analytical judgment:

> *[Claim, Methodology]* "The paper introduces NV-Embed, a generalist embedding model based on decoder-only large language models (LLMs), aimed at enhancing performance in downstream tasks such as retrieval…"

Even the few substantiated premises restate the paper's own justifications rather than providing independent evaluations:

> *[GS = 0.0, Methodology]* "The introduction of a latent attention layer for sequence pooling is theoretically grounded in dictionary learning concepts and is argued to mitigate information dilution compared to mean pooling or last token pooling."

> *[GS = 1.0, Experiment]* "These results are backed by detailed comparisons and tabulated scores (see Table 14)."

The low grounding quality compounds the low premise ratio: $R_{\text{premise}} = 0.152$ and $\overline{\text{GS}} = 0.215$ together produce $\text{DoA}_{\text{HM}} = 0.178$ — a **69% drop** from Human's 0.574.

Key insight. Reviewer2's failure on this paper illustrates a form of analytical failure beyond pure volume-inflation: even its few "premises" ground claims in the paper's own assertions rather than in independent analytical observations. The DoA metric correctly penalizes this through the joint harmonic mean of R_{premise} and $\overline{\text{GS}}$ — both of which must be high to achieve human-level analytical depth.

Case 2: The Surface-Level Trap in Practice. *Paper: VLAP — Visual-Language Alignment via Pre-trained Word Embeddings*

Context. VLAP proposes a lightweight vision-language alignment method that maps visual representations directly into the pre-trained word embedding space of a frozen LLM using a single trainable linear layer, supervised by an optimal transport-based assignment objective and an image captioning loss. The design is deliberately minimal: the LLM and visual encoder remain frozen, with only the linear projection trained.

Table 13: Per-reviewer DoA statistics for VLAP. Human row is the mean across 5 individual reviewers. Avg GS normalized to $[0, 1]$; $\text{DoA}_{\text{HM}} = \text{HM}(R_{\text{premise}}, \overline{\text{GS}})$.

Reviewer	DoA_{HM}	$\text{R}_{\text{premise}}$	Avg GS	Total	Premises	Claims	% Clarity
Human (mean, n=5)	0.567	0.614	0.527	14	8	6	2%
Reviewer2	**0.551**	0.569	0.534	51	29	22	**0%**
CycleReviewer	0.483	0.529	0.444	17	9	8	**0%**
SEA	0.417	0.500	0.357	14	7	7	29%
DeepReview	0.263	0.759	0.159	29	22	7	**0%**
TreeReview	0.252	0.194	0.357	36	7	29	**29%**

Reviewer consensus: Clarity is irrelevant on a paper with a simple design. This is a particularly instructive paper for the surface-level trap because the simplicity of VLAP's design — a single linear layer, two losses, frozen backbones — leaves almost no room for legitimate reproducibility criticism. Accordingly, **DeepReview, CycleReviewer, and Reviewer2 allocate 0%** of their premise budget to Clarity, and **Human** allocates a negligible **2%** (Table 13). Instead, they focus exclusively on the paper's technical formulation and experimental comparisons:

> *[GS = 2.0, Novelty]* "Contrastive alignment in ALBEF, BLIP, and the first-stage alignment by BLIP2 includes image-text matching and image-grounded text generation." *(Human_2)*

> *[GS = 2.0, Methodology]* "An optimal transport-based training objective is proposed to enforce the consistency of word assignments for paired multimodal data. This allows frozen LLMs to ground their word embedding space in visual data." *(Human_3)*

> *[GS = 2.0, Methodology]* "Using this method, experiments are done on 3 tasks — image captioning, VQA, image-text retrieval — showing the method outperforms existing methods." *(Human_5)*

Reviewer2, despite its typical tendency toward low R_{premise}, here achieves $\text{DoA}_{\text{HM}} = 0.551$ with 29 premises from 51 arguments ($R_{\text{premise}} = 0.569$, $\overline{\text{GS}} = 0.534$) — all directed at Methodology and

Experimental Design. CycleReviewer produces 9 premises from 17 arguments with 0% Clarity and an average $\overline{\text{GS}} = 0.444$.

SEA: Clarity premises praising, not criticizing. SEA allocates 29% to Clarity, matching TreeReview's proportion. However, its two Clarity premises are *positive quality affirmations* about the paper, not complaints:

> *[GS = 0.0, Clarity]* "The methodology is clearly explained, making it accessible and understandable, which is crucial for reproducibility and further research."

> *[GS = 0.0, Clarity]* "The paper is well-structured, with comprehensive experiments and detailed analyses."

Although these also carry GS = 0 (generic, no specific section cited), they function as strength notes rather than reproducibility criticisms. SEA then pivots to three substantive Experimental premises, maintaining $\text{DoA}_{\text{HM}} = 0.417$.

TreeReview: reproducibility boilerplate displaces technical analysis. TreeReview produces only 7 premises from 36 total arguments ($R_{\text{premise}} = 0.194$). Of these 7, **2 (29%) are Clarity premises**, both carrying GS = 0 and targeting the same generic reproducibility axis:

> *[GS = 0.0, Clarity]* "This omission hinders reproducibility and limits the ability of other researchers to build upon the work."

> *[GS = 0.0, Clarity]* "This would greatly enhance the reproducibility of the method."

Neither statement names a specific omission. The first is a consequence claim ("hinders reproducibility") with no antecedent in the review — it is evidently the continuation of a claim made in a preceding *claim* argument, not a self-contained premise. The second is a one-sentence recommendation that names no missing artifact. On a paper whose entire contribution is a single linear layer with two objectives, these statements convey essentially no analytical information.

By contrast, TreeReview's five non-Clarity premises do engage with the paper's mechanism (optimal transport alignment, the single linear layer, efficiency), but all at GS = 1 with no external literature anchoring. This yields $\overline{\text{GS}} = 0.357$ and, combined with $R_{\text{premise}} = 0.194$, $\text{DoA}_{\text{HM}} = \mathbf{0.252}$ — a **56% drop** from Human's 0.567.

Key insight. The surface-level trap on this paper is not caused by a genuinely unclear manuscript: four of the six reviewers independently assess that VLAP requires no Clarity criticism at all. The trap is therefore *triggered internally* by TreeReview's reviewing heuristic — a tendency to produce generic reproducibility premises ("hinders reproducibility", "greatly enhance reproducibility") regardless of whether the paper's design warrants them. These boilerplate premises consume 2 of TreeReview's 7 available premise slots and contribute nothing to the analytical depth of the review.

F.2 Novelty Assessment

Case 1: The Speculative Critique Trap We illustrate the Novelty Score (NS) metric through a representative paper from ICLR, which proposes **CONPAIR**, a contrastive compositional dataset and **EVOGEN**, a curriculum contrastive learning framework for improving compositional text-to-image (T2I) generation in diffusion models. Table 14 summarises key statistics for the Human and SEA reviewers on this paper.

Table 14: Per-reviewer novelty statistics for paper **CONPAIR**. Stances: **N**=*novel*, **NN**=*not_novel*, **SW**=*somewhat_novel*, **U**=*unclear*. $\bar{s}$ denotes the raw mean per-claim score on the $[-2, +2]$ scale; the normalized review-level metric is $NS = (\bar{s} + 2)/4$.

Reviewer	$\bar{s}$	#Claims	%N	%NN	%SW	%U
Human	1.20	10	50%	30%	0%	20%
SEA	1.73	5	80%	0%	20%	0%

Both reviewers correctly identify the paper's core contributions, specifically, the novel use of contrastive learning in the denoising encoder and the CONPAIR dataset's value as a hard-negative compositional benchmark. The divergence in $\bar{s}$ arises not from a disagreement on *what* is novel, but from how reviewers handle *uncertain* assessments.

Human reviewer produces 10 claims with a nuanced mixture of stances. Claim C8 are labelled *unclear* and carry speculative criticism that the metric system cannot verify against any paper in the related-work pool:

> *[C8, unclear]* "The ContraFusion model is compared against other methods using the T2I-CompBench dataset. However, it is trained on the Com-Diff dataset, which **likely overlaps** noticeably with the T2I-CompBench test set."

This concern about training/test overlap is plausible but entirely speculative: no related paper in the retrieved pool provides evidence for or against dataset overlap between Com-Diff and T2I-CompBench. Consequently, 7 of 11 related-paper comparisons for C8 return UNSUPPORTED or INSUFFICIENT, and the claim's per-pair score averages to -0.18, substantially dragging down the overall Human $\bar{s}$. By contrast, the Human's other well-evidenced claims score positively:

> *[C4, novel]* "Introducing a contrastive loss in the denoising encoder representation is an interesting idea." $(s_k = +2.0)$

> *[C5, not_novel]* "A few other datasets with hard-negative compositional images exist, albeit small. Two examples are COLA and Winoground." $(s_k = +2.0)$

C5 exemplifies *calibrated not_novel* assessment: the reviewer correctly names prior datasets (COLA, Winoground) that are confirmed by related-work retrieval, earning a full $+2.0$ score even though the stance is non-endorsing.

SEA reviewer produces only 5 claims, all either *novel* or *somewhat_novel*, with no speculative or *unclear* claims:

> *[C3, novel]* "The proposed dataset, CONPAIR, is well-designed with a clear progression from simple to complex compositional scenarios." $(s_k = +2.0)$

> *[C4, novel]* "The multi-stage fine-tuning strategy is innovative and addresses the issue of models being overwhelmed by mixed-difficulty data during training." $(s_k = +2.0)$

> *[C5, somewhat_novel]* "The use of a VQA model to revise generated captions and ensure text-image alignment is an effective approach." $(s_k = +2.0)$

Each SEA claim targets a specific, verifiable contribution (curriculum structure, fine-tuning strategy, VQA-based alignment), and none raises unverifiable speculation. As a result, SEA achieves a higher raw mean claim score ($\bar{s} = 1.73$ vs. 1.20) despite generating only half the number of claims.

Insight — Speculative Assessment Penalty. This case illustrates a systematic pattern: Human reviewers raise *unclear*-stance concerns (e.g., suspected data contamination, alignment-quality trade-offs) that are reasonable in context but impossible to corroborate through paper-pool evidence. The NS metric penalises such claims because they lack verifiable grounding—exactly the behaviour the metric is designed to detect. SEA tends to avoid this failure mode: it focuses on claims that are directly grounded in the methods described in the paper, thus achieving higher evidence-calibrated novelty scores. This finding does *not* imply SEA is a "better reviewer"; rather, it shows that SEA's positively-biased and evidence-anchored claim style is systematically rewarded by the evidence-grounded NS metric, while the richer, more skeptical human review style can be penalized when speculative concerns appear.

Case 2: Coverage Without Paper `3TGUvHmZ2v` from ICML, examines a theoretical paper that characterises the expressivity of fixed-precision Transformer decoders using formal language theory. The paper establishes three main results: **(R1)** without positional encoding (NoPE), fixed-precision

Transformers can recognise only finite and co-finite languages; **(R2)** adding absolute positional encoding (APE) extends expressibility to cyclic languages; **(R3)** relaxing parameter bounds further allows recognition of letter-set languages. Table 15 summarises reviewer statistics.

Table 15: Per-reviewer novelty statistics for paper `3TGUvHmZ2v`. Stances: **N**=*novel*, **NN**=*not_novel*, **SW**=*somewhat_novel*. $\bar{s}$ denotes the raw mean per-claim score on the $[-2, +2]$ scale; the normalized review-level metric is $NS = (\bar{s} + 2)/4$.

Reviewer	$\bar{s}$	#Claims	%N	%NN	%SW	Overlap w/ Human
Human	1.467	5	80%	0%	20%	—
DeepReview	1.500	12	58%	42%	0%	5/5 + 7 new

Human reviewer makes 5 well-targeted claims, all centred on the paper's three core theoretical results. Three claims earn the maximum per-claim score $s_k = +2.0$ by directly naming the language classes established:

> *[H-C1, novel]* "This paper demonstrates that fixed-precision Transformer decoders without positional encoding are limited to recognizing only finite or co-finite languages. . ." $(s_k = +2.0)$

> *[H-C5, novel]* "The paper explores the expressive capabilities of Transformer decoders constrained by a fixed-precision setting, such as specific floating-point arithmetic. . .utilising formal language theory." $(s_k = +2.0)$

Two claims are penalised to $s_k = +0.667$: H-C2 makes a vague generality ("limited to *finite memorization*") rather than naming the class precisely, and H-C4 (*somewhat_novel*: "adding positional encoding improves expressivity but does not alleviate the main limitations") lacks the concreteness of specifying *cyclic* languages. The result is a compact but calibrated review: every claim is evidenced, but some are imprecise enough that the related-work pool only partially supports them.

DeepReview produces 12 claims, $2.4\times$ more, at a nearly identical raw mean claim score ($\bar{s} = 1.500$ vs. 1.467). Its first 7 claims cover *all* of Human's 5 novelty dimensions, often with greater precision, while claims DR8–DR12 open an entirely new dimension absent from the human review: well-evidenced critical analysis of the paper's *scope limitations*.

Covering Human's novelty dimensions with increased precision:

> *[DR2, novel]* "Introducing absolute positional encoding extends their capabilities to recognizing **cyclic languages**, while allowing non-finite floating-point values further expands their expressivity to **letter-set languages**." $(s_k = +2.0)$

> *[DR5, novel]* "The paper's focus on **constant precision**, as opposed to *logarithmic* precision, is a significant strength, as it reflects the practical constraints of real-world implementations." $(s_k = +2.0)$

DR2 subsumes H-C4 (*somewhat_novel*) by naming both upgraded classes precisely, earning $s_k = +2.0$ vs. H-C4's $+0.667$. DR5 makes an explicit constant-vs-log-precision distinction that Human's C5 only gestures towards.

Identifying additional gaps through well-evidenced limitation claims:

> *[DR10, not_novel]* "The paper's analysis of positional encoding is limited to absolute positional encoding (APE) and no positional encoding (NoPE). It does not explore **relative positional encodings**, which are commonly used in modern Transformer architectures." $(s_k = +2.0)$

> *[DR12, not_novel]* "Finally, the paper's analysis is primarily **theoretical**, and it lacks empirical validation of its findings." $(s_k = +2.0)$

Crucially, these limitation claims are *not_novel* in stance but still earn $s_k = +2.0$. This is because the novelty metric rewards *calibration*: the claim that "relative PE is not studied here" is verifiable

against the related-work pool (papers using RoPE, ALiBi, etc. do exist, confirming the gap), so the claim is evidence-grounded. By contrast, three of DeepReview's limitation claims (DR6: APE/cyclic redundancy; DR8: multi-layer not explored; DR11: sinusoidal vs. learned PE) earn $s_k = 0.0$, because either they are internally redundant with other DeepReview claims or the related-work pool cannot confirm they are genuine gaps.

Table 16: Semantic overlap between Human and DeepReview novelty claims for `3TGUvHmZ2v`. DeepReview DR1–DR7 cover all Human claims; DR8–DR12 add new gap analysis.

Human claim	**DeepReview equivalent(s)**	s_k **(DR)**
H-C1: finite/co-finite (NoPE)	DR1, DR4	$+2.0, +2.0$
H-C2: "finite memorization" (vague)	DR3 (rigorous framework)	$+2.0$
H-C3: regular language subclass (NoPE)	DR1, DR7	$+2.0, +2.0$
H-C4: PE improves but limits remain (vague)	DR2 (APE → cyclic/letter-set)	$+2.0$
H-C5: fixed-precision setting (practical)	DR5 (constant vs. log precision)	$+2.0$
— (not in Human)	DR8: multi-layer/multi-head gap	0.0
— (not in Human)	DR9: softmax→hardmax gap	$+2.0$
— (not in Human)	DR10: relative PE gap (RoPE)	$+2.0$
— (not in Human)	DR11: sinusoidal vs. learned PE	0.0
— (not in Human)	DR12: no empirical validation	$+2.0$

Insight — Coverage Expansion with Preserved Calibration. This case illustrates a second systematic pattern distinct from Case 1: DeepReview does not achieve a higher raw mean claim score $\bar{s}$ by avoiding criticism, but by *expanding coverage* while maintaining the same calibration quality as human reviewers. DeepReview's extra volume comes in two flavours: (i) *precision elaboration*—naming specific language classes and precision regimes more exactly than Human; and (ii) *gap enumeration*—identifying well-evidenced scope limitations (no relative PE, no empirical validation) that human reviewers do not consider. The NS metric rewards both, because both types of claim can be verified against the related-work pool. Thus DeepReview achieves $2.4\times$ the claim count with only $\Delta\bar{s} = +0.033$, demonstrating that LLM reviewers can generate *denser* novelty assessments that are *equally* well-calibrated to the literature, not merely more verbose.

F.3 Flaw Identification & Major Issues Prioritization

Case 1: Complementary Blind Spots — The Equation-Level Scanner vs. the Practical Assessor. *Paper* `Tv2JDGw920` (ICML)

Context. This paper proposes a preconditioning-based optimizer for Domain Generalization that leverages the One-Step Generalization Ratio (OSGR) to dynamically balance parameter-wise gradient updates. The paper provides three theoretical analyses (OSGR equalization, PAC-Bayes bound, convergence proof) and extensive experiments across five DG benchmarks. The canonical flaw bank contains 52 entries (after de-duplication: $\sim$24 unique flaws), of which 20 are valid upon independent verification (9 identified only by the LLM reviewer, 9 only by human reviewers, and 2 by both) — all Minor in severity.

Table 17: Flaw coverage comparison for `Tv2JDGw920`. After de-duplication, the LLM and Human reviewers each independently identify $\sim$9 unique valid flaws, with only **2 flaws shared** between them, demonstrating largely complementary diagnostic profiles.

Reviewer	**Valid Flaws**	**LLM-only**	**Human-only**	**Shared**
Human (3 reviewers)	11	—	9	2
Reviewer2 (LLM)	11	9	—	2
Union	20	9	9	2

What the LLM catches that all Humans miss: systematic equation-level scrutiny. Reviewer2 identifies 9 unique valid flaws that none of the three human reviewers raise. These flaws share

a distinctive pattern: they arise from systematically walking through the paper's mathematical derivations and cross-referencing theoretical claims against their practical implementation.

> *[LLM-only, Methodology]* "In the PAC-Bayes analysis, the prior π is approximated using all data except the current batch. Why is this approximation valid, and what is the impact of this choice on the tightness of the generalization bound?"

This flaw targets the data-dependent prior in Theorem 3.6 (Appendix C.3.1), where the paper states the prior is "approximated with stochastic gradient descent using all data excluding the current mini-batch." Classical PAC-Bayes theory requires the prior to be chosen *independently* of training data. While modern PAC-Bayes extensions accommodate data-dependent priors, the paper does not formally invoke such extensions, leaving a gap in the theoretical justification.

> *[LLM-only, Methodology]* "In Corollary 3.2, the preconditioning factor p_j is derived under the assumption that gradients are independent across parameters. However, in practice, gradients are often highly correlated."

Verifying this requires tracing from Equation (48) through (50) in Appendix C.2, where the factorization $\mathbb{E}(\mathbf{p} \odot \boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}') = \sum_j p_j \cdot \sigma^2(\theta_j)$ implicitly assumes zero cross-parameter covariance — a standard but unverified assumption in deep networks with shared activations across layers.

> *[LLM-only, Clarity]* "The paper frames the use of OSGR in the optimizer as novel, [but] it is not entirely clear how this differs from existing adaptive learning rate methods. For instance, Adam already adapts learning rates based on gradient magnitudes."

Table 1 in the paper decomposes the proposed optimizer and Adam into convergence and alignment terms, but the structural similarity between the two requires parsing dense notation to appreciate the distinction — precisely the kind of close reading that time-constrained reviewers may skip.

What Humans catch that the LLM misses: claim–evidence calibration and field norms. Human reviewers identify 9 unique valid flaws that Reviewer2 entirely overlooks. These flaws target gaps between what the paper *claims* and what the *evidence* supports — a form of practical wisdom rooted in familiarity with community expectations.

> *[Human-only, Methodology]* "The paper claims [the optimizer] promotes domain-invariant features, but it doesn't directly evaluate this claim by examining feature representations. For example, some DG papers use metrics like center divergence between domain features." *(Human_2)*

The paper provides qualitative visualizations showing class separation across domains, but no quantitative domain-invariance metric (e.g., MMD, center divergence) — a standard evaluation in the DG literature that the LLM does not demand.

> *[Human-only, Methodology]* "The claim that uniformly distributed OSGR across parameters indicates better generalization. . . is stated as a conjecture rather than a theorem, and while intuitively supported, it's not rigorously demonstrated." *(Human_2)*

The paper transparently labels this as a "Conjecture" (§3.2) and provides partial support via Jensen's inequality (Appendix C.3.2). However, only a human reviewer flags the intellectual-honesty gap between a conjecture and a theorem — the LLM accepts the conjecture's supporting evidence without questioning its formal status.

Where the LLM hallucinates: asserting absence of content that exists. Reviewer2 also generates several flaws that are *directly contradicted* by the paper. The most striking: the LLM claims "the paper does not adequately explain the computational cost relative to simpler optimizers," despite the paper explicitly reporting training times across multiple iteration budgets (e.g., proposed method: 4,292 s vs. Adam: 5,443 s at 5K iterations). Similarly, the LLM asserts the paper "lacks qualitative insights (e.g., feature maps, attention weights)," overlooking multiple visualizations already present in the manuscript. These fabrications follow a consistent pattern: the LLM evaluates flaw claims in isolation without cross-referencing the manuscript's actual content.

Key insight. This case demonstrates that LLM and human reviewers operate as *complementary diagnostic instruments* with near-zero overlap. The LLM excels at systematic, equation-level verification — surfacing implicit assumptions (gradient independence, data-dependent priors) and theory-practice gaps that require cross-referencing proofs against algorithms. Human reviewers excel at claim–evidence calibration — demanding quantitative backing for qualitative claims and recognizing when a conjecture substitutes for a theorem. Neither perspective subsumes the other: the union of their flaw sets produces substantially broader diagnostic coverage than either alone, while both are susceptible to distinct failure modes (LLM: content hallucination; Human: limited equation-level scrutiny under time pressure).

Case 2: The Diagnostic Volume Advantage — Broader Coverage Through Exhaustive Scanning. *Paper* `aapUBU9U0D` *(ICLR)*

Context. This paper proposes an iterative data augmentation pipeline for fine-tuning LLMs on Operations Research (OR) tasks. The method generates synthetic optimization problems via an evolutionary strategy, validates them through four LLM-based checkers, and fine-tunes LLaMA-3-8B on the resulting corpus. Performance is evaluated on three OR benchmarks. The canonical flaw bank contains 28 entries, all valid upon verification, contributed equally by human reviewers (14 flaws) and Reviewer2 (14 flaws).

Table 18: Flaw coverage and aspect distribution for `aapUBU9U0D`. Despite equal flaw counts, the two reviewer groups cover substantially different *aspects* of the paper, with Reviewer2 providing broader topical coverage across all four macro-categories.

Reviewer	Valid	Methodology	Experimental	Clarity	Applicability
Human (4 rev.)	14	3	7	2	2
Reviewer2 (LLM)	14	4	4	1	5
Union	28	7	11	3	7

Human reviewers: deep but concentrated coverage. The four human reviewers collectively produce 14 valid flaws, with a strong concentration on Experimental Design (7 of 14, 50%). Their critiques are precise and field-specific, reflecting deep familiarity with OR benchmarking conventions:

> *[Human, Experimental Design]* "No comparison with traditional OR solvers is provided. The paper claims to advance 'automation of decision-making' but never shows whether LLM-based modeling is competitive with established OR methods."

> *[Human, Experimental Design]* "The results should clarify whether COPT or GUROBI is used for evaluation, and whether baselines were given equal solver access."

These flaws target the paper's core experimental validity — precisely the type of domain-expert criticism that requires knowledge of solver ecosystems and OR benchmarking norms. However, human coverage leaves notable gaps: only 2 of 14 flaws address Applicability and Limitations, and no human reviewer questions the pipeline's scalability to large-scale industrial problems or its dependence on English-only data.

Reviewer2: broader aspect coverage with systematic gap enumeration. Reviewer2 produces an equal number of valid flaws (14), but distributes them more evenly across aspect categories. Notably, 5 of its 14 flaws target Applicability and Limitations — a category that human reviewers largely neglect:

> *[LLM-only, Applicability]* "The scalability of the pipeline to large-scale optimization problems with thousands of variables and constraints is not evaluated."

> *[LLM-only, Applicability]* "The pipeline's applicability beyond linear and mixed-integer programming — for example, to nonlinear or stochastic optimization — remains unverified."

> *[LLM-only, Methodology]* "The four validation checkers are entirely LLM-prompt-based. No formal algorithmic specification, error bounds, or coverage guarantees are provided for the checking pipeline."

These flaws arise from the LLM's systematic scanning pattern: rather than evaluating only the paper's explicit claims, Reviewer2 probes the *boundaries* of the contribution — what is not tested, not formalized, and not generalized. This "boundary probing" behavior is precisely what drives Reviewer2's elevated recall at the macro level: by exhaustively questioning scope and applicability, the LLM surfaces valid concerns that time-constrained human reviewers deprioritize in favor of core experimental scrutiny.

The aspect complementarity pattern. When the two flaw sets are merged, the union covers all four macro-categories substantially: Methodology (7), Experimental Design (11), Clarity (3), and Applicability (7). Neither reviewer group alone achieves this breadth. Human reviewers anchor the evaluation in domain-specific experimental rigor, while Reviewer2 extends coverage into methodological formalization and scope limitations.

Key insight. This case concretizes the volume-to-coverage advantage observed at the aggregate level. Reviewer2's exhaustive scanning style does not merely produce *more* flaws — it produces flaws in *different diagnostic categories* than human reviewers, systematically covering scope limitations and methodological formalization gaps that humans deprioritize. The resulting union of human and LLM flaw sets achieves broader aspect coverage than either alone, reinforcing the practical value of LLM-augmented peer review as a diagnostic broadening mechanism rather than a replacement for human expertise.

F.4 Multi-dimensional Constructiveness

Case: The Actionability Gap — From Diagnosis to Prescription. *Paper: GenColor: A Diffusion-Based Framework for Color Enhancement in Digital Photography* (NeurIPS)

Context. GenColor proposes a no-reference color enhancement pipeline consisting of three learned components: (1) a diffusion-based Color Generation Module, (2) a Texture Preservation Module, and (3) a post-processing Global Adjustment step. The system is trained on the authors' proprietary ARTISAN dataset and evaluated via user studies and no-reference image quality metrics.

Table 19: Per-system Constructiveness scores for GenColor. All dimensions are on a $[0, 2]$ raw scale.

System	MCS	D1:Act	D2:Spec	D3:Just	D4:Sol	D5:Tone	ARCs
Human	0.488	0.721	1.767	0.372	0.326	1.698	43
DeepReview	**0.724**	**1.588**	**2.000**	0.882	**1.059**	1.706	17
Reviewer2	0.800*	1.000	2.000	**2.000**	1.000	**2.000**	24
CycleReviewer	0.400	0.833	1.833	0.167	0.167	1.000	6
TreeReview	0.434	0.793	1.517	0.241	0.207	1.586	29

*Reviewer2's MCS is inflated by near-perfect D3 scores (verbose observation-only summaries; see text).

Human: Precise Diagnosis, Absent Prescription. Human reviewers produce 43 ARCs with respectable specificity ($\overline{D2} = 1.767$), demonstrating solid familiarity with the paper's technical details. However, D4 (Solution) averages only 0.326: the majority of human ARCs identify a problem but stop short of prescribing a remedy. Representative human weakness ARCs illustrate this gatekeeping pattern:

> *[D1 = 1, D4 = 0]* "The method's near-deterministic nature raises concerns about user control and the ability to capture personalized styles."
>
> *[D1 = 1, D4 = 0]* "The proposed method has a relatively long runtime compared to lightweight comparison models."
>
> *[D1 = 1, D4 = 0]* "Previous methods trained on the proposed dataset perform worse, questioning the dataset's general usefulness."

Each statement correctly identifies a real limitation of GenColor:determinism, efficiency, dataset scope but offers no concrete path forward for the authors to address it. This is the hallmark of *diagnostic gatekeeping*: the review functions as a verdict, not a guide.

DeepReview: Prescriptive Constructiveness at Scale. DeepReview produces only 17 ARCs, yet achieves $\overline{D1} = 1.588$ and $\overline{D4} = 1.059$, both substantially above Human. Critically, **six of its ARCs reach the maximum solution score D4 = 2** — meaning the feedback specifies not only *what* is missing but *how* to address it:

> *[D1 = 2, D4 = 2]* "Include comprehensive ablation studies by systematically removing or modifying components to evaluate their individual contribution."
>
> *[D1 = 2, D4 = 2]* "Provide a detailed analysis of the computational cost of each component and the overall pipeline, including training time, inference time, and memory requirements."
>
> *[D1 = 2, D4 = 2]* "Clearly articulate the novelty of the specific combination and modifications of existing components, detailing unique aspects such as the training regime, degradation scheme, and weight blending strategy."

Compared to the human weakness ARCs above, these comments address the *same underlying concerns* (ablation coverage, efficiency, novelty justification) but close the loop by telling the authors exactly what artifact to produce or section to revise.

Reviewer2: High MCS via Justification Inflation, Not Solutions. Reviewer2 achieves the highest raw MCS on this paper (0.800), driven entirely by near-perfect D3 scores ($\overline{D3} = 2.000$). Its 24 ARCs are detailed, well-grounded observations — but they are *observations*, not directives. D1 averages only 1.000 and D4 averages 1.000, because every ARC points to a general direction at best ("no systematic analysis is provided", "no theoretical basis is provided") without specifying an implementable fix. This reveals a limitation of MCS as a holistic score: high D3 can mask absent D4. The D4 sub-score surfaces this asymmetry directly.

CycleReviewer and TreeReview: Vagueness and Volume. CycleReviewer contributes 6 ARCs with $\overline{D4} = 0.167$ and $\overline{D3} = 0.167$ — vague, unsubstantiated comments such as:

> *[D1 = 1, D4 = 0]* "The method might struggle with images exhibiting complex textures or patterns."

TreeReview produces 29 ARCs but maintains $\overline{D4} = 0.207$; its highest-scoring feedback requests reproducibility details (training parameters, hyperparameter selection) rather than engaging with the paper's core design choices.

Key insight. This case illustrates the central behavioral gap in constructiveness: even when human reviewers are technically perceptive (high D2, moderate D3), they default to *problem identification without resolution*. DeepReview's architectural orientation toward explicit remediation — reflected in D4 exceeding human baseline by $+0.733$ and D1 by $+0.867$ — demonstrates that high constructiveness is not a matter of writing more, but of *closing the feedback loop* from critique to actionable prescription.

G Limitations and Future Works

While PRISM provides a rigorous, multi-dimensional benchmarking framework for automated peer review, we acknowledge several limitations that highlight avenues for future research.

Domain Generalization. Our dataset comprises 600 manuscripts exclusively from premier machine learning and representation learning venues (ICLR, ICML, NeurIPS). The structural norms, citation densities, and evaluation criteria in these venues differ from those in other scientific disciplines (e.g., clinical medicine, humanities, or pure mathematics). Consequently, the current instantiation of PRISM may require recalibration before deployment in non-ML domains.

Temporal Scope. PRISM is anchored to the three most recently concluded venue-years available at the time of dataset construction (ICLR 2026, ICML 2025, and NeurIPS 2025). This temporal choice was made to reduce the risk of benchmark contamination: where cutoff information is available, the selected manuscripts and their human reviews were released after the reported training cutoffs of

the evaluated reviewer systems and the LLM judge. A benchmark constructed from artifacts that a model could plausibly have encountered during training would provide a weaker separation between genuine reviewing capability and memorization. Nevertheless, temporal filtering cannot guarantee the complete absence of contamination. Training corpora and model update schedules, particularly for proprietary models, are not fully observable, and post-cutoff release dates therefore provide only a practical mitigation rather than a guarantee that benchmark instances were never encountered during training. We additionally verified that the selected venue-years do not materially alter our conclusions: across an earlier and broader pool of candidate venue-years considered during dataset construction, the relative ranking of all five automated reviewer systems remains unchanged across PRISM dimensions. However, as scientific manuscripts and peer reviews continue to accumulate publicly, currently uncontaminated benchmark instances may eventually enter the training corpora of future reviewer or judge models. We therefore view PRISM primarily as a reusable evaluation framework rather than a fixed dataset: future benchmark releases can periodically refresh the evaluation corpus with newly released conference cycles while retaining the underlying evaluation methodology.

LLM Dependency: Hallucination, Prompt Sensitivity, and Judge Bias. A foundational premise of PRISM is delegating complex tasks—such as text atomization, fact-finding, and scoring—to frontier LLMs. It is well-documented that LLMs are heavily susceptible to hallucinations (fabricating non-existent critiques or citations) and prompt sensitivity (where minor structural variations in instructions yield divergent outputs). To actively mitigate these vulnerabilities, PRISM strictly departs from monolithic, single-prompt evaluation. By decomposing the framework into constrained, multi-phase pipelines and enforcing deterministic decoding, we significantly restrict the generation space and effectively filter out hallucinated noise. Nevertheless, this dependency introduces residual bottlenecks: atomizing isolated sentences inherently risks context loss, fact-finding remains bounded by the coverage of external retrieval APIs, and models acting as judges may still retain subtle internal priors for specific rhetorical styles. In this work, our primary evaluation pipeline is instantiated using `Gemini 2.5 Flash Lite`. While we conducted preliminary robustness checks with an alternative model (Xiaomi `MiMo V2.5 Pro`) on a data subsample to confirm baseline metric stability, this single-judge dependency means we cannot fully rule out model-specific evaluation biases. Future work must not only develop robust uncertainty quantification to prevent edge-case extraction errors from cascading into downstream metrics, but also conduct comprehensive multi-judge studies across diverse LLM families to fully isolate and eliminate judge-specific priors.

Future Works We identify four priority directions: (1) *Cross-domain generalization*—recalibrating PRISM for clinical medicine, social sciences, and pure mathematics. (2) *Judge robustness*—systematically studying inter-judge agreement across LLM judge families and human raters. (3) *Deeper outcome validation*—while Section 4.2.5 shows that only human review scores differ significantly between accepted and rejected papers, extending this analysis to additional downstream outcomes, such as post-review author satisfaction, would further test whether PRISM metrics capture meaningful review quality rather than superficial fluency. (4) *Benchmark sustainability*—periodically refreshing the evaluation corpus with newly released conference cycles to limit contamination risk while retaining PRISM's underlying evaluation methodology.