# ARD-REVIEW: GENERATING ACCURATE, REASONABLE, AND DIVERSE AI PEER REVIEWS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Current AI review systems tend to generate generic, unsuited, nitpicky, and invalid (GUNI) feedback, and this failure mode risks homogenizing not only reviews but, through defensive writing, the papers themselves. Many of these approaches also evaluate the paper without relying on previous human reviews, so their scores may not match human ratings. We propose ARD-Review, a split-role pipeline that combines a high-recall harsh critic, an adversarial merger that aggressively filters GUNI content, and a retrieval-based calibration step that anchors scores to human-reviewed reference papers. We also introduce a new evaluation suite, ARD-Bench, to quantify the GUNI tendency in AI-generated reviews. On ARD-Bench, our method achieves the best scoring correlation among non-commercial review systems. The DeepSeek-based ARD-Review is the only non-commercial method that significantly outperforms the human one-vs-one baseline on both correlation metrics and MAE. When paired with the stronger Sonnet 4.6 backbone, it outperforms the one-vs-rest human baseline and even surpasses the commercial CSPaper method. Its review content quality is also significantly better than that of all other methods tested. It wins 85.53% of pairwise validity comparisons while keeping cross-paper overlap close to the direct-review level. Code is at `https://anonymous.4open.science/r/ARD-Review-D725/`.

## 1 INTRODUCTION

As reasoning-enabled Large Language Models (LLMs) grow more capable, LLM-generated and LLM-assisted reviews have attracted growing attention; people are using them both to help authors strengthen papers before submission and to attempt to replace or assist human reviewers in formal review settings, but both uses could raise concerns. One problem is that most work focuses on the accuracy of scores and decisions and the helpfulness of the feedback (Zhu et al., 2025; Weng et al., 2026; Cao et al., 2025; Goyal et al., 2026; Sahu et al., 2025), but overlooks that these systems tend to surface similar weaknesses, reducing the diversity of reviews (Baumann et al., 2026). We further argue that this lack of diversity can propagate to the papers themselves and make papers homogeneous. These agents tend to generate generic, unsuited, nitpicky, or even invalid weaknesses (termed GUNI in this paper), and iterative self-checking with these tools or optimizing for these tools in a formal review setting may not only degrade user experience but also encourage authors to write defensively, adding unnecessary content until papers converge toward homogenized, unengaging prose.

Several design patterns in existing systems contribute to GUNI reviews. In multi-agent pipelines that assign fixed scores to individual agents (Cao et al., 2025), low-scoring agents are forced to inflate minor issues into major weaknesses when no fatal flaws exist; because these critiques appear plausible on the surface, they can survive the merging stage. Gated annotation loops (Weng et al., 2026) create similar pressure by requiring a minimum number of comments before an agent can exit, while small fine-tuned models (Idahl & Ahmadi, 2025; Zhu et al., 2025) produce superficial, repetitive comments because they lack the capacity to reason about paper-specific content. Post-hoc grounding (Li et al., 2026) may anchor a generic weakness to a specific section, but generic reviews like demands for broader experiments or more hyperparameter analysis still apply to virtually any paper regardless of where they are grounded. Rebuttal-based filtering (Sahu et al., 2025; Zou et al., 2026), on the other hand, risks amplifying LLM sycophancy when the filtering agent defers too readily to the author's response. These problems motivate ARD-Review, a split-role agentic pipeline that separates the broad search for potential weaknesses and strengths from the decision about which



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.

points belong in a review. An initial harsh critic identifies candidate points, after which a merger aggressively filters invalid or generic items and uses RAG-like comparisons with human-reviewed papers to calibrate the score and suppress sycophancy. To evaluate whether this design produces accurate, valid, and diverse reviews, we introduce ARD-Bench, an evaluation suite that brings the overlooked problem of review homogenization into the assessment alongside scoring accuracy. Our results show that ARD-Review achieves the best scoring accuracy among non-commercial review systems. With DeepSeek, it performs on par with CSPaper in scoring accuracy while outperforming it in review content quality. With Sonnet 4.6, it surpasses CSPaper in scoring accuracy.

Our contributions are: (1) we examine how current reviewing practices lead to GUNI reviews; (2) we propose an AI review agent for generating accurate, reasonable, and diverse reviews; and (3) we design ARD-Bench, a framework and set of metrics for quantifying GUNI tendencies in AI-generated reviews.

## 2 Related Works

Existing AI review systems differ in how they generate candidate comments and refine them into a final assessment, but many introduce artifacts through their design choices. CSPaper (Cao et al., 2025) assigns five agents with fixed scores (1–10 with a step of 2 points) to review the paper independently and selects the most reasonable assessment, then incorporates the others. However, forcing low-scoring agents to justify their predetermined scores leads them to fabricate or overstate unreasonable criticisms. These unreasonable weaknesses might still sound reasonable superficially, and thus get incorporated into the final review, inflating the amount of GUNI content. ReviewerToo (Sahu et al., 2025) takes a different approach by combining a critic, a lenient reviewer, and a specialist, then passing their outputs through an author rebuttal agent and a meta-reviewer. DIAGPaper (Zou et al., 2026) similarly employs an author rebuttal module to filter out invalid criticisms. While the rebuttal mechanism is intended to catch low-quality comments, in both systems the author rebuttal agents risk amplifying LLMs' sycophancy, ultimately inflating the scores (Sahu et al., 2025). DeepReviewer-V2 (Weng et al., 2026) addresses a different problem: it prevents premature termination through a gated loop that requires a minimum number of PDF annotations. This design, however, forces the model to surface trivial issues and floods authors with a wall of text noise. CSPaper also includes a correctness agent (CSPaper, 2026a) that focuses on validating the paper's claims rather than providing a general-purpose review. It identifies many potential issues upfront and then filters them out afterward.

Before general-purpose LLMs became powerful enough, fine-tuning domain-specific models was a common alternative, including for AI reviewers. OpenReviewer (Idahl & Ahmadi, 2025) trains an 8B model on OpenReview data, while DeepReview (Zhu et al., 2025) trains a 14B model with synthetic chains of thought. Both, however, remain limited in their ability to reason about paper-specific content and tend to produce generic or superficial reviews. ReviewGrounder (Li et al., 2026) attempts to address this by grounding the output of a fine-tuned Phi-4-14B drafter in specific paper sections and then applying rubric-guided synthesis. However, when the underlying model already produces a generic review draft, grounding alone does not make the weaknesses specific. For example, pointing "the paper only tested 5 baselines, more baselines will be appreciated" to the baseline section is valid and grounded, but the criticism itself remains generic. Additionally, the guiding rubric in ReviewGrounder and the evaluation rubric are very similar (the evaluation one is an instantiated version of the guiding rubric), which can artificially inflate the quality metric.

Because LLM-based scoring is often neither well-grounded nor well-calibrated, several works have tried to calibrate the resulting scores. The Stanford Review Agent (Jiang & Ng) has the LLM generate a set of subscores across multiple dimensions, then applies linear regression to estimate the overall score. CSPaper also includes a ranking component (CSPaper, 2026b) that performs semantic analysis on the generated review to infer an advantage score and compare it against other papers.

Across these approaches, evaluation has also focused primarily on scoring accuracy against human ratings, with most systems reporting only this metric. ReviewerToo extends the assessment to constructiveness, depth of engagement, balance, readability, and helpfulness, but several of these criteria are difficult to operationalize consistently, and some axes, like readability, favor LLM-generated reviews by construction. ARD-Bench addresses the remaining gap by directly measuring GUNI tendency, making it possible to assess an aspect of review quality that prior work has not evaluated.

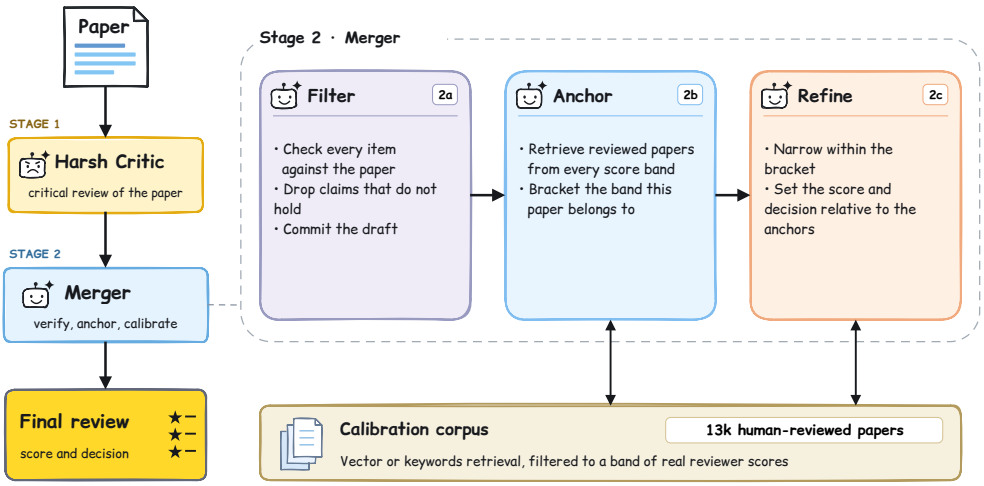



Figure 1: Overview of the ARD-Review pipeline.

## 3 ARD-REVIEW

ARD-Review pursues accuracy, reasonableness, and diversity by dividing the review process between a Harsh Critic ($\mathcal{H}$), a Filter ($\mathcal{F}$), and a Calibrator ($\mathcal{K}$) as illustrated in Figure 1. $\mathcal{F}$ and $\mathcal{K}$ together form the merger ($\mathcal{M}$). This division allows the critic to search broadly for potential issues while the merger decides which points are justified and how they should affect the score. Given a paper $P$ and a calibration dataset $\mathcal{D}$, the critic produces an initial assessment that the merger filters and calibrates into the final review $R$ and score $s$.

$$R_0 = \mathcal{F}\big(\mathcal{H}(P), P\big) \tag{1}$$

$$(R, s) = \mathcal{K}(R_0, P, \mathcal{D}). \tag{2}$$

### 3.1 HARSH CRITIC

The Harsh Critic ($\mathcal{H}$) begins with a stringent assessment that prioritizes recall over precision, leaving the subsequent merger to determine which candidate weaknesses warrant inclusion. Its output therefore combines a broad set of potential weaknesses $\mathcal{W}$ with reasonable strengths $\mathcal{S}$ and other comments $\mathcal{C}$:

$$(\mathcal{W}, \mathcal{S}, \mathcal{C}) = \mathcal{H}(P).$$

To form this assessment, $\mathcal{H}$ examines methodological soundness, evaluation validity, comparison fairness, strength of evidence, internal coherence, and significance, prioritizing fatal flaws when present. It also provides additional suggestions that are directed toward improving the paper on its own terms, with attention to the paper's existing direction and positioning.

Keeping strengths and weaknesses within the same first-pass assessment is deliberate. In a "good-cop-bad-cop" design with separate agents for each, we found that a dedicated strength-finding agent triggered sycophancy downstream: later agents became reluctant to assign low scores to poor papers because they "still have some merits." We therefore use a single critic for both and retain the alternative design as the Split Agents ablation.

### 3.2 FILTER

The broad search performed by $\mathcal{H}$ makes filtering central to the merger's role. Using the paper as context, $\mathcal{F}$ applies an adversarial filter to the candidate strengths and weaknesses, aggressively scrutinizing each point and removing or downgrading GUNI content. For each $x \in \mathcal{W} \cup \mathcal{S}$, it decides

whether a strength should be kept or removed and assigns weaknesses to categories that reflect their severity:

$$b_P^s(x) \in \{\text{strength},\ \text{removed}\}.$$
$$b_P^w(x) \in \{\text{fatal},\ \text{major},\ \text{minor},\ \text{trivial},\ \text{nice-to-have},\ \text{removed}\}.$$

These assignments follow a rubric that removes generic criticisms without concrete anchors in the paper, downgrades unsuited demands that exceed its scope or relevant community standards, discards presentation nitpicks, and rejects invalid points caused by factual errors or misreadings. The rubric is shown in the Appendix. The merger also weakens concerns that the paper has already reasonably addressed, distinguishes substantive weaknesses by severity, and subjects strengths to the same evidence check. Points assigned to `removed` remain available to authors as auxiliary feedback, together with brief explanations of why they were excluded; the filtered review state $R_0$ then forms the basis for score calibration.

We conduct this filtering without the author rebuttal agent used in ReviewerToo (Sahu et al., 2025) and DIAGPaper (Zou et al., 2026), because introducing an author's response may lead $\mathcal{M}$ to give that response excessive weight and exacerbate sycophancy.

### 3.3 Calibration Scoring

Once the merger has determined which points are justified, Calibration Scoring ($\mathcal{K}$) translates the filtered assessment into a score by comparing the target paper and $R_0$ with human-reviewed reference papers. This step uses the same merger agent and its existing context, drawing its comparisons from a reference dataset

$$\mathcal{D} = \{(P_i, R_i, s_i)\}_{i=1}^{N}$$

in which each entry contains a paper $P_i$, its human reviews $R_i$, and its associated score $s_i$. We construct $\mathcal{D}$ from 2024–2025 ICLR submissions in the DeepReview-13k dataset so that the calibration has access to both reference scores and the assessments behind them, and there is no direct paper overlap with our dataset (see Appendix G for details).

Starting from the target paper $P$ and $R_0$, $\mathcal{K}$ generates topic-based queries to retrieve an initial set of relevant papers:

$$\mathcal{A}_1 = \text{Retrieve}(\mathcal{D}, P, R_0).$$

This retrieval is score-stratified, giving $\mathcal{A}_1$ anchors from different score regions as well as topical relevance, and $\mathcal{K}$ can inspect both the aggregate score and the individual human reviews for each retrieved paper. We express the resulting comparison as a heuristic function $h(P, R_0, \mathcal{A})$ that jointly considers the target paper, its filtered review, and these human-reviewed anchors. With the initial anchor set, the function estimates a coarse score interval:

$$I_1 = [\ell_1, u_1] = h_{\text{range}}(P, R_0, \mathcal{A}_1).$$

The estimated interval narrows the comparison for a second retrieval, in which $\mathcal{K}$ searches for reference papers within that range:

$$\mathcal{A}_2 = \text{Retrieve}\left(\{(P_i, R_i, s_i) \in \mathcal{D} : s_i \in I_1\}, P, R_0\right)$$

The resulting anchors provide denser reference points around the plausible score region, allowing the same comparison procedure to operate at finer granularity and produce the final score:

$$s = h_{\text{score}}(P, R_0, \mathcal{A}_2), \qquad s \in \left\{\frac{k}{2} : k \in \{1, \ldots, 20\}\right\}.$$

If the evidence remains ambiguous, $\mathcal{K}$ may retrieve anchors once more, for a maximum of three rounds in total, and it reports how the comparisons narrow the initial range toward the final estimate. After calibration, $\mathcal{K}$ produces the final review $R$ and could make further adjustments to $R_0$ with more information from the calibration process. Through this process, real human-reviewed examples anchor the score in comparative human judgments and reduce reliance on the model's uncalibrated scoring priors.

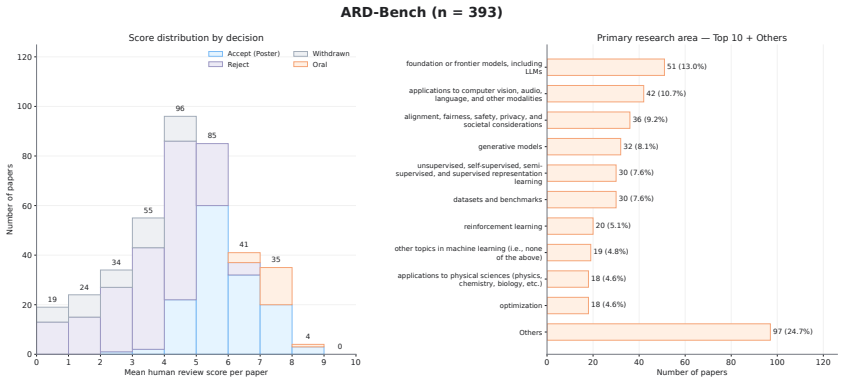



Figure 2: ARD-Bench Score, Decision, and Area Distribution

# 4 Evaluation

## 4.1 Dataset

ICLR's fully open-review process makes its papers and reviews publicly accessible, which has made it a common data source for AI reviewer systems. For our evaluation, we collected ICLR 2026 papers and their pre-rebuttal reviewer scores rather than using an existing benchmark such as DeepReview-13k (Zhu et al., 2025), because both the recency of the papers and the stage at which their scores were recorded matter for the comparison.

The first consideration is data leakage. DeepReview-13k draws its test set from ICLR 2025, whose reviews became public in late 2024 and discussions about them were posted online throughout 2025; these materials are likely to appear in the training data of many base models, allowing memorized information to inflate apparent reviewing performance.

The second consideration is the independence of the reference judgments. A cybersecurity incident led the PC to freeze ICLR 2026 reviewer scores, giving us access to the original pre-rebuttal ratings. These ratings more directly reflect independent judgments than post-rebuttal scores, which may become increasingly correlated as reviewers read authors' responses and one another's reviews, inflating the human baseline numbers.

From these submissions, we constructed a slightly flattened distribution by combining the original data with stratified sampling to include more extreme-case papers. We downloaded the PDFs through the official OpenReview SDK and processed them with Marker, while also collecting the human reviews and associated scores. During preprocessing, we removed the string "Published in ICLR 2026" from accepted papers because it directly reveals acceptance, a source of leakage that, to the best of our knowledge, prior work has not considered. The resulting evaluation dataset, which we call ARD-Bench, is summarized in Figure 2.

## 4.2 Metrics

**Objective Metrics** Our evaluation covers both the accuracy of the assigned scores and the validity and diversity of the reviews that accompany them. For scoring accuracy, we report mean absolute error (MAE), mean signed error (ME, labeled Bias in the tables), Spearman and Pearson correlation coefficients, and AUROC.

**Validity Measurements** Assessing review validity requires a judge that can determine whether a stated weakness is justified by the paper. ReviewCritique (Du et al., 2024) argued that LLMs cannot perform this meta-reviewing task, citing their low agreement with humans on weakness reliability; however, our inspection attributes this low agreement to substantial noise in the dataset, as detailed in Appendix A. We therefore use the dataset from TADDLE (Duan & Li, 2026) to assess our validity judge. TADDLE includes seven types of invalid reviews alongside a normal review for each paper.

Because the original trained TADDLE judge is not publicly available, we use a training-free judge validated on the TADDLE dataset. The judge prompt lists six deficiency dimensions: information error, careless omission, unsubstantiated claims, constructiveness, bias, and professionalism, but the

judge outputs only an overall validity comparison. We implement this assessment as a pairwise comparison with GLM-5.3-Flash (GLM-5-Team et al., 2026), which receives two reviews and the paper itself and predicts which review is more valid. The order of the two reviews given to the judge is randomized to prevent position bias.

To measure the judge's reliability, we evaluated it on all 90 reviews in the TADDLE test split, randomly sampling pairs whose validity scores differed by more than 1. Its pairwise accuracy was 85.526%, which we use to account for judge error when reporting comparisons between review systems: for an observed win rate $w$, we report the range $[\max(0, w - e), \min(1, w + e)]$, where $e = 1 - 0.85526$.

**Diversity Metrics** Alongside validity, we assess whether a system repeatedly raises similar weaknesses across different papers, estimating cross-paper review overlap with an approach similar to Baumann et al. (2026). We first use GLM-5.3-Flash to merge near-duplicate weaknesses within each review. This mainly reduces duplicate items in settings with multiple reviewers, such as human review and DeepReviewer-14B. For paper $i$, $R_i = (W_{i,1}, W_{i,2}, \ldots, W_{i,n_i})$ denotes the resulting list of $n_i$ weaknesses. We then compare papers $i$ and $j$ through a separate GLM-5.3-Flash call that receives $R_i$ as the prompt and the weaknesses in $R_j$ as options. For each $W_{j,k}$, the model returns a binary indicator $u_{i,j,k}$ of whether the same point appears in $R_i$, allowing us to compute the Dice-like overlap score

$$\text{Overlap Score} = 2 \cdot \sum_{k=1}^{n_j} u_{i,j,k} / (n_i + n_j)$$

, where lower values indicate greater diversity. We deliberately avoided embedding-based similarity as in Baumann et al. (2026), as it can misleadingly raise this metric: LLM-produced text frequently exhibits shared surface-level properties such as style and vocabulary, which do not necessarily correspond to deeper content similarity.

### 4.3 Baselines

We compare ARD-Review with systems representing the main approaches discussed above: the fine-tuned DeepReviewer-14B (Zhu et al., 2025) (meta review output), the agent-based DeepReviewer-V2 (Weng et al., 2026), and the commercial CSPaper (Cao et al., 2025). The backbone model of CSPaper is GPT-5.4. We run both DeepReviewer-V2 and ARD-Review with DeepSeek-V4-Flash (DeepSeek-AI, 2026a), choosing it for its balance of price and quality[1]. We also included a stronger model backbone, Claude Sonnet 4.6 (Anthropic, 2026a), for our method. To place these systems alongside direct use of an LLM, we also include two prompting baselines. The DeepSeek version is implemented in the OpenAI Agent SDK and the Claude version is implemented in the Claude Agent SDK, with the global default Claude Code config applying. Direct Review uses a detailed prompt that requests sections similar to those in our pipeline and supplies a past score distribution and the official ICLR reviewer guidelines, whereas Simple Review uses only "Review this paper for ICLR" to approximate a typical user's chat request.

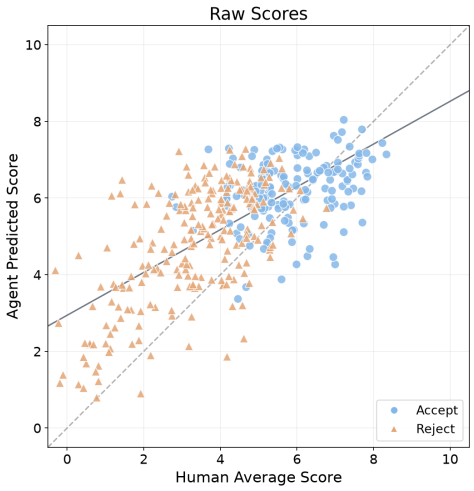



Figure 3: Score visualization of ARD-Review using Sonnet as backbone model.

For CSPaper, we tested two versions: an official public version with internet access and a limited-access offline version, as our results show that the internet-enabled version produced implausibly high correlation with human scores, likely because it directly accessed the data online. We confirmed this in Appendix Section F. We report the online result for transparency, but it should not be treated as valid. CSPaper also offers a paper ranking feature that uses semantic analysis on the generated

[1]The DeepSeek-V4-Flash version we used is not the later 0731 release but the original release.

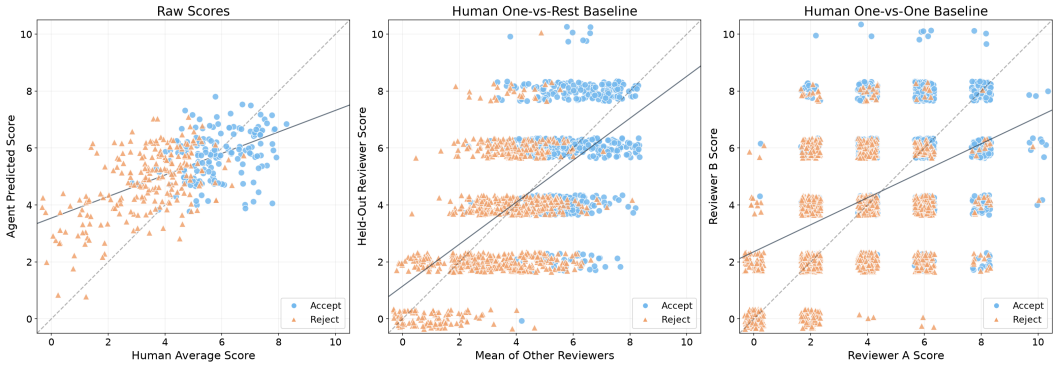



Figure 4: Visualization of the scoring accuracy of ARD-Review (DeepSeek) compared with humans. The three plots (starting from the left) are: (1) a scatter plot of human scores versus ARD-Review predictions; (2) a human one-vs-rest correlation; (3) a human one-vs-one correlation.

review to produce a score. We included this as CSPaper Ranking. Since it does not produce a new review and its score is not normalized, we do not report its quality metrics and MAE/Bias.

Due to the same leakage concern, we also disable DeepReviewer-V2 retrieval components, including DeepXiv access and related-work search, because the evaluation papers and their ground-truth reviews are publicly indexed. Since DeepReviewer-V2 does not require retrieval for scoring and it is only used for review, omitting it is a fair comparison with our method. However, data leakage at the model level is less concerning. We have explained in detail in Appendix G the possibility that the model may have encountered these papers during training.

The availability of human reviews also allows us to compare AI scores with human judgments. We used two human baselines: a human one-vs-rest baseline that uses one reviewer's score to predict the mean of the remaining reviewers for the same paper, and a human one-vs-one baseline that uses one reviewer's score to predict another reviewer's score. We also included a naive leave-one-out mean (LOO Mean) baseline that predicts each paper's score using the mean score of all other papers. The comparison excludes the Stanford Review Agent, ReviewerToo (Sahu et al., 2025), and ScholarPeer (Goyal et al., 2026), since these systems either lack batch access or are not publicly available.

# 5 Results

## 5.1 Main Results

Table 1 and Figures 4 and 3 present the scoring results, with three ablations showing how the pipeline's design relates to its performance: Direct Review, Split Agents with separate agents for strengths and weaknesses, and No-Calibration, which retains the multi-stage review process but omits calibration.

ARD-Review achieves the best Spearman, Pearson, and MAE among non-commercial review systems, exceeding the human one-vs-one baseline and closely matching human one-vs-rest correlation. Excluding the unreliable online CSPaper result, it also has the lowest bias and least sycophancy among agentic methods, where positive bias indicates sycophancy and negative bias indicates overly harsh scoring. Bias needs to be read alongside the other metrics: Simple Review's low bias reflects mode collapse, just as the naive mean baseline has zero bias.

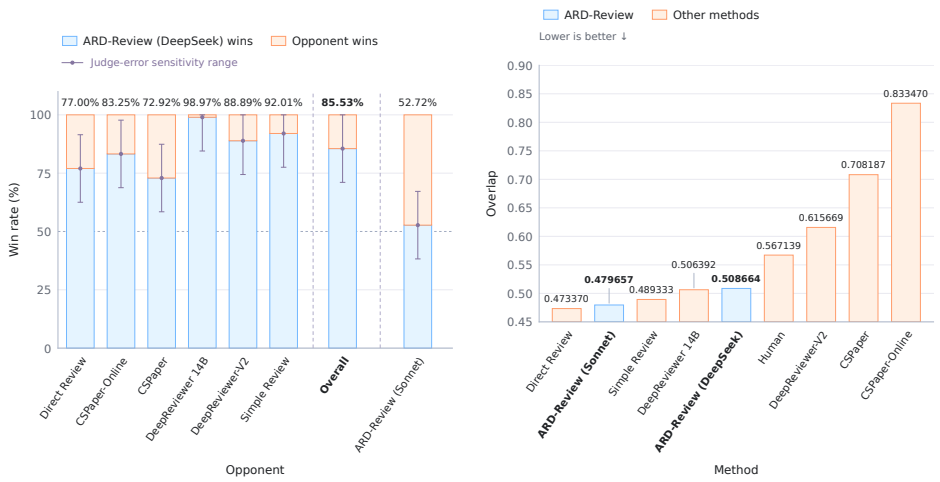
Statistical comparisons with the competing methods support the accuracy gains. ARD-Review's MAE is significantly lower than that of every other non-commercial review system, although its difference from the naive LOO Mean baseline is not significant ($p$=.073); its Pearson's $r$ is also significantly higher than that of the other non-commercial systems except DeepReviewer-V2 ($p$=.091) and Split Agents, against which it is only marginally significant ($p$=.051). Against commercial CSPaper, the scoring performance of the DeepSeek version is on par, with no significant differences in Spearman ($p$=.680), Pearson ($p$=.252), or MAE ($p$=.099), while the review-content assessment in Section 5.2 shows an advantage for ARD-Review. With Sonnet 4.6, ARD-Review significantly outperforms CSPaper in both Spearman ($p$=.041) and Pearson ($p$<.001) correlations, while its AUROC advantage is only marginal ($p$=.056).

The comparison with human reviewers provides a further reference point. In the bootstrap comparisons against human one-vs-one performance, ARD-Review (DeepSeek) is the only non-commercial method that significantly improves both correlations and MAE, and ARD-Review (Sonnet) is the only method significantly better than human one-vs-rest on both correlation and MAE. The human AUROC is computed from individual reviewer scores. Because human scores directly contribute to the accept/reject decisions themselves, their AUROC involves a circular comparison and is inflated. However, our pipeline, when using Sonnet 4.6, still reaches an AUROC close to this inflated number.

### 5.2 GUNI Assessment

The content evaluation examines whether the scoring performance is accompanied by valid, paper-specific feedback. Figure 5a reports pairwise validity comparisons against each baseline using the GLM judge described above, which prefers ARD-Review (DeepSeek) in 85.53% of all comparisons, indicating a substantial advantage in review content quality. The DeepSeek version and the Sonnet version have similar validity. We exclude human reviews from this assessment because the post-rebuttal manuscripts given to the judge may already address many of their comments, leaving those comments no longer valid for the version of the paper being assessed.

Validity alone does not capture the homogenization concern that motivates our method, so Figure 5b also reports the cross-paper overlap metric from Section 4.2. The overlap of ARD-Review (DeepSeek) remains close to the direct-review baseline and below DeepReviewer-V2, whereas CSPaper-Online and CSPaper exhibit by far the highest homogenization. Since lower overlap indicates more diverse, paper-specific reviews, these results, together with Figure 5a, show that aggressive filtering improves validity while preserving diversity.




(a) Validity win rate.

(b) Cross-paper overlap.

Figure 5: Review quality assessment. (a) Pairwise validity win rate of our method against each baseline. The range is the judge-error-considered range, not confidence intervals. (b) Cross-paper review overlap (lower is more diverse), estimated by an LLM judge on randomly sampled paper pairs. In both figures, our method uses DeepSeek.

### 5.3 Different Backbones

The main results use DeepSeek-V4-Flash and Claude Sonnet 4.6, but we evaluate the pipeline with five additional models to evaluate how well it transfers: Claude Opus 4.6 (Anthropic, 2026b), Qwen-3.6-27B (Qwen Team, 2026), DeepSeek-V4.1-Flash (DeepSeek-AI, 2026b), Qwen-3.8-Flash-Next (Qiu et al., 2026), and GLM-5.3-Flash (GLM-5-Team et al., 2026). Qwen-3.6-27B is a small model that can be run locally on many consumer-grade or professional-grade GPUs, offering options for labs to review their papers without sending them to a cloud LLM provider. DeepSeek-V4.1-Flash, Qwen-3.8-Flash-Next, and GLM-5.3-Flash are affordable open-source models, allowing groups with limited budgets to use the tool. Claude Opus 4.6 is a stronger commercial model whose training

Table 1: Score prediction on ICLR 2026. Boldface marks the highest correlations and AUROC and the lowest MAE among review systems. * marks results significantly better than the human one-vs-one baseline, and † marks results significantly better than the human one-vs-rest baseline ($p < .05$, paired paper-level bootstrap). None of these marks applies to CSPaper-Online, whose results are unreliable (see Section F), or to the human and naive baselines.

| Method | Spearman $\rho$ | Pearson $r$ | MAE | Bias | AUROC |
|---|---|---|---|---|---|
| Simple Review | 0.3928 | 0.4227 | 1.7905 | +0.5910 | 0.6708 |
| Direct Review | 0.4644 | 0.5117 | 2.0198 | +1.6588 | 0.7186 |
| Naive (LOO Mean) | - | - | 1.4471 | 0.0000 | - |
| Split Agents | 0.5100 | 0.5549 | 1.6329 | +1.2629 | 0.7051 |
| No-Calibration | 0.5052 | 0.5512 | 1.7358 | +1.4724 | 0.7031 |
| CSPaper-Online | 0.7725 | 0.7588 | 0.9824 | +0.4236 | 0.9275 |
| CSPaper | 0.5650* | 0.5713* | 1.4157† | +0.9120 | 0.7455 |
| CSPaper Ranking | 0.5102 | 0.5262 | - | - | 0.7517 |
| DeepReviewer-V2 | 0.5167 | 0.5587 | 1.6959 | +1.4053 | 0.7545 |
| DeepReviewer-14B | 0.4547 | 0.5223 | 1.7046 | +1.3471 | 0.6966 |
| Human (one-vs-rest) | 0.5612 | 0.6025 | 1.4986 | - | - |
| Human (one-vs-one) | 0.4625 | 0.4831 | 1.7064 | - | 0.8096 |
| Ours (DeepSeek) | 0.5487* | 0.6137* | **1.3335**† | +0.8336 | 0.7417 |
| Ours (Sonnet 4.6) | **0.6376**† | **0.7000**† | 1.3636† | +1.0187 | **0.7879** |

data cutoff predates the release of the ICLR 2026 reviews (Appendix G). Table 2 reports the results of these backbone comparisons.

Our pipeline shows strong robustness across different LLMs. For all models, the correlation remains on par with the human one-vs-rest baseline, and the AUROC stays above 0.7 and close to the human AUROC. Because the human AUROC is inflated by circularity, coming close to it is already meaningful. Even with the smaller Qwen-3.6-27B backbone, score correlation remains comparable to human one-vs-rest performance and is slightly above human one-vs-one performance, significantly so for Pearson ($p$=.0040) but not for Spearman ($p$=.2060). With Claude Opus 4.6, ARD-Review significantly exceeds the human one-vs-rest baseline in both Spearman ($p$=.0010) and Pearson ($p$<.001), and its AUROC of 0.8105 reaches the human AUROC.

Table 2: Performance of our method under different backbones. The $p$-values are computed against the human one-vs-rest baseline. The AUROC from individual human reviewer scores is 0.8096.

| Model | Spearman | Pearson | AUROC |
|---|---|---|---|
| DeepSeek-V4-Flash | 0.5487 ($p$=.7690) | 0.6137 ($p$=.7020) | 0.7417 |
| Claude Opus 4.6 | 0.6611 ($p$=.0010) | 0.7176 ($p$<.001) | 0.8105 |
| Qwen-3.6-27B | 0.5115 ($p$=.2270) | 0.5814 ($p$=.5570) | 0.7488 |
| DeepSeek-V4.1-Flash | 0.5684 ($p$=.8150) | 0.5953 ($p$=.8660) | 0.7549 |
| GLM-5.3-Flash | 0.5400 ($p$=.7070) | 0.5868 ($p$=.7860) | 0.7317 |
| Qwen-3.8-Flash-Next | 0.5136 ($p$ = .2790) | 0.5843 ($p$ = .6800) | 0.7227 |

## 6 CONCLUSION

AI review systems that repeatedly generate generic, unsuited, nitpicky, and invalid (GUNI) feedback risk homogenizing both reviews and, through defensive writing, the papers revised in response to them. ARD-Review addresses this problem by separating a high-recall search for potential issues from adversarial filtering, then calibrating the resulting assessment against human-reviewed reference papers. On ARD-Bench, this design achieves the best scoring correlation among non-commercial review systems and the least sycophancy bias among agentic methods; when using DeepSeek, it is also the only non-commercial method that significantly exceeds the human one-vs-one baseline in both correlations and MAE. With DeepSeek, its scoring accuracy is on par with commercial CSPaper, while its review content quality is significantly better, winning 85.53% of pairwise validity comparisons and keeping cross-paper overlap close to the direct-review level. With a stronger backbone, Sonnet 4.6, its accuracy is significantly higher than a stronger human one-vs-rest baseline, and its score correlation is even higher than that of the commercial method CSPaper, also with statistical significance.

## 7 AI USAGE DISCLOSURE

LLMs were used to search for related works, brainstorm ideas, and execute the ideas by writing code and monitoring the experiments. The paper was also reviewed by ARD-Review itself, CSPaper, and other tools, and revised accordingly. A human remained in the loop, exercised critical judgment, and made the final decisions on LLM outputs.

## REFERENCES

Anthropic. System Card: Claude Sonnet 4.6. Technical report, February 2026a. URL https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf.

Anthropic. System Card: Claude Opus 4.6. Technical report, February 2026b. URL https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Position: Stop automating peer review without rigorous evaluation. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=128yyxUoRk.

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, June 2023.

Lele Cao, Lei You, and R&D Team. CSPaper Review: Fast, Rubric-Faithful Conference Feedback. In Lucie Flek, Shashi Narayan, Lê Hòng Phuong, and Jiahuan Pei (eds.), *Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations*, pp. 3–7, Hanoi, Vietnam, October 2025. Association for Computational Linguistics.

Zijin Chen, Lesui Yu, Xiaofei Liao, Hai Jin, and Qinbin Li. Position: Peer Review Should Be Calibrated via LLM Scoring. In *Forty-Third International Conference on Machine Learning Position Paper Track*, June 2026.

CSPaper. Correctness Check: Claim Verification and Error Discovery for Scientific Papers. https://cspaper.org/articles/correctness-check, March 2026a.

CSPaper. Introducing Review Rank: Calibrated Scores and Relative Standing, February 2026b. URL https://cspaper.org/articles/introducing-review-rank.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026a.

DeepSeek-AI. DeepSeek-V4.1-flash: Pushing the limits of KV cache compression, 2026b.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5081–5099, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.292.

Hanqi Duan and Xiang Li. TADDLE: A Tool-Augmented Agent for Detecting Deficient LLM-Generated Peer Reviews. *arXiv preprint arXiv:2605.26911*, 2026.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao,

Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763. arXiv: 2602.15763 [cs.LG].

Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review, May 2026.

Wenqi Marshall Guo, Yiyang Du, Heidi J. S. Tworek, and Shan Du. Position: The Pitfalls of Over-Alignment: Overly Caution Health-Related Responses From LLMs are Unethical and Dangerous, August 2025a.

Wenqi Marshall Guo, Qingyun Qian, Khalad Hasan, and Shan Du. Aesthetic Alignment Risks Assimilation: How Image Generation and Reward Models Reinforce Beauty Bias and Ideological "Censorship". *arXiv preprint arXiv:2512.11883*, 2025b.

Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A Specialized Large Language Model for Generating Critical Scientific Paper Reviews. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.44.

Yixing Jiang and Andrew Ng. Tech Overview - Stanford Agentic Reviewer. https://paperreview.ai/tech-overview.

Enkelejda Kasneci and Gjergji Kasneci. Sycophancy is an Educational Safety Risk: Why LLM Tutors Need Sycophancy Benchmarks, May 2026.

Marius Köppel, Alexander Segner, Martin Wagener, Lukas Pensel, Andreas Karwath, and Stefan Kramer. Pairwise Learning to Rank by Neural Networks Revisited: Reconstruction, Theoretical Analysis and Practical Performance, September 2019.

Zhuofeng Li, Yi Lu, Dongfu Jiang, Haoxiang Zhang, Yuyang Bai, Chuan Li, Yu Wang, Shuiwang Ji, Jianwen Xie, and Yu Zhang. ReviewGrounder: Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents, April 2026.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, Fan Zhou, Hao Luo, Haofeng Huang, Haoran Lian, Haoyan Huang, Hongqing Chen, Jianwei Zhang, Jing Xu, Junjie Wang, Langshi Chen, Liangyu Wang, Linlang Jiang, Man Yuan, Minmin Sun, Peng Jin, Siqi Zhang, Siyu Wang, Xingzhang Ren, Yakai Wang, Yi Zhang, Yiming Dong, Yizhong Cao, Yubo Ma, Yunfei Mao, Bo Zheng, and Dayiheng Liu. On the Design of Qwen3.8-Next Architecture: Evaluation, Efficiency, and Training Stability,

August 2026. URL http://arxiv.org/abs/2608.30320. arXiv:2608.30320 [cs.CL] version: 1.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. ReviewerToo: Should AI Join The Program Committee? A Look At The Future of Peer Review. October 2025.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Zhiyuan Ning, Shichen Li, Panzhong Lu, Zhen Lin, Enhao Gu, Qiyao Sun, and Yue Zhang. DeepReviewer 2.0: A Traceable Agentic System for Auditable Scientific Peer Review. https://arxiv.org/abs/2604.09590v1, March 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, June 2025.

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.

Zhuoyang Zou, Abolfazl Ansari, Delvin Ce Zhang, Dongwon Lee, and Wenpeng Yin. DIAGPaper: Diagnosing Valid and Specific Weaknesses in Scientific Papers via Multi-Agent Reasoning, February 2026.

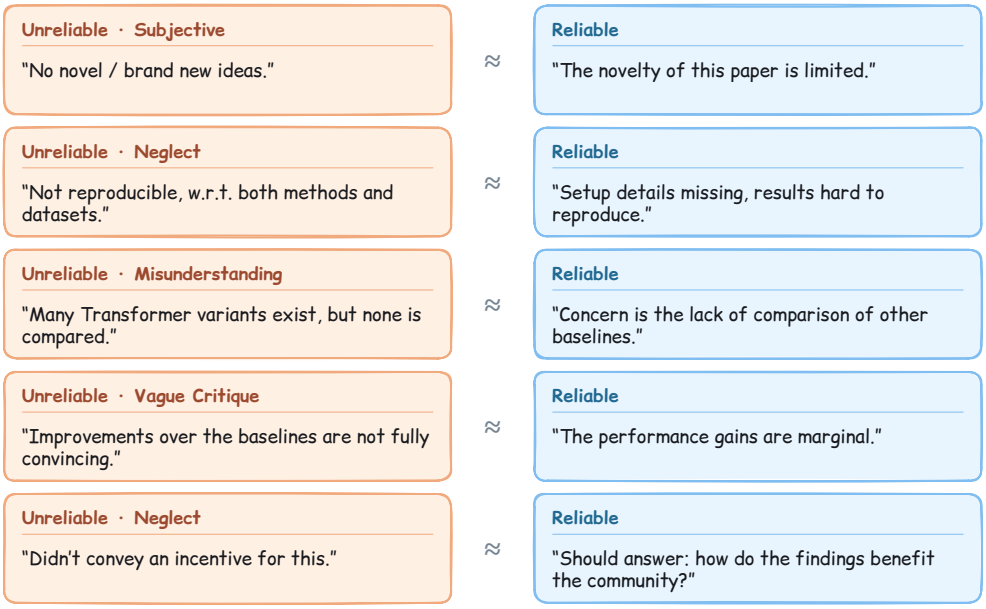



Figure 6: Near-duplicate weakness segments from reviews of the same paper receive opposite labels in ReviewCritique.

## A Problems with the ReviewCritique Dataset

We observed that ReviewCritique does not report inter-rater agreement. This raises the possibility that the low performance of LLMs stems from label noise rather than an underlying capability limitation. Because ReviewCritique lacks multi-rater annotations, we adopted a proxy strategy. We compared segments from reviews of the same paper, identified near-duplicates, and inspected their labels. These labels often diverged, as illustrated in Figure 6. When we compute inter-segment agreement, the F1 score is only approximately 0.3. We therefore contend that the low performance reported on ReviewCritique is not simply because "LLMs are not good enough," but rather reflects a combination of the inherent subjectivity of meta-reviews and limitations in the dataset's quality control.

This result invalidates the claim in ReviewCritique that LLMs are unable to perform meta-reviews, even for the earlier-generation LLMs evaluated in that work.

## B Qualitative Comparison of Reviews on SEA

Figure 7: Full weakness sections for *SEA: Sparse Linear Attention with Estimated Attention Mask*. The review text is reproduced without rephrasing. RAG scoring was disabled when producing ARD-Review because SEA is a 2024 paper.

**CSPaper**

1. **Efficiency vs other linear methods is not clearly favorable.**
   - While SEA is much more accurate than Performer/Reformer at comparable budgets, it is also *considerably* more expensive than most linear baselines in both latency and memory, especially at moderate sequence lengths.
   - From Table A.2, at $T = 4096$ the latency for SEA with $k = 64$ is 1.77 ms/iter versus 0.25 ms (Performer), 0.36 ms (Cosformer), and 0.21 ms (Sinkhorn); at $T = 16384$, SEA (k=64) is 7.40 ms vs 0.90 / 1.49 / 1.00 ms respectively. So for these lengths SEA is ∼3–8× slower than other efficient attentions.

**CSPaper (continued)**

- Table A.1 shows the same trend for memory: at $T = 16384$, SEA (k=64) uses 198.76 MB versus 46.54 MB (Performer) and 24 MB (FlashAttention). Even though it is much better than quadratic attention, "competitive" with other linear methods (Section 5, Fig. 8 top) is a bit overstated.
- In Fig. 7a (GLUE trade-offs), SEA points often sit to the right of Performer/Sinkhorn in the latency plots (higher ms) while only modestly better in accuracy, especially on smaller sequence lengths (256 tokens). The paper would be stronger with a clearer narrative on when the extra overhead is justified and concrete break-even analyses.

2. **Training-time quadratic cost and scalability of distillation are under-discussed.**
   - Section 3.2 and Eq. (1) define several losses that require *dense* teacher and student attention matrices during training:
   - $\mathcal{L}^{(i)}_{\text{approx}} = \text{KL}(A'_i, \hat{A}_i) + \text{MSE}(A'_i, \hat{A}_i)$ after interpolating $\hat{A}_i$ to a full $T \times T$ matrix;
   - $\mathcal{L}^{(i)}_{\text{prob}} = \text{KL}(A_i, \hat{A}_i) + \text{MSE}(A_i, \hat{A}_i)$ where $A_i = \sigma(Q_i K_i^\top)$;
   - plus MSE losses on context and layer outputs.
   - Both computing and storing the teacher $Q_i K_i^\top$ and the dense student $Q_i K_i^\top$ for each layer/head are $\mathcal{O}(T^2)$ in time and memory. Yet, the paper repeatedly emphasizes linear complexity and "only a small finetuning dataset" without quantifying the actual overhead of distillation (wall-clock vs simply finetuning vanilla attention, peak memory during KD, etc.).
   - For long-context LMs (e.g., $T \geq$ 4k–8k) or large teachers, these quadratic training costs could be prohibitive and partially negate the practical benefit of a linear-time inference module. A careful discussion or empirical measurement of KD cost is needed.
3. **Methodological complexity and a few unclear mathematical choices.**
   - The SEA block combines Performer $(\phi(Q), \phi(K), V_{\text{cat}})$, a two-MLP + CNN decoder to produce $\hat{A}$, grouped top-$\hat{k}$, sparse interpolation (with corner cases when $kK < T$), a FlatCSR-based sparse matmul, and an additional global pooling term mixed via a learned gate $s_{\text{mix}}$. This is quite a heavy and specialized stack compared to standard attention or most linear variants.
   - The sparse attention definition in Section 3.1.1 sets
     $$A^* = s_{\text{prob}} \odot \sigma(\rho(Q, K^\top, M^*)), \quad s_{\text{prob}} = f_{\text{prob}}(Z) \in \mathbb{R}^T.$$
     After softmax, each row of $\sigma(\rho(Q, K^\top, M^*))$ sums to 1; elementwise scaling with $s_{\text{prob}}$ breaks this normalization. The text still refers to $A^*$ as "attention probability," but it is no longer a probability distribution over keys. This is not necessarily wrong, but the semantics are confusing and might matter for interpretability. A short justification (e.g., viewing $s_{\text{prob}}$ as a temperature-like factor or attention dropout) would help.
   - Likewise, for the global component, the paper defines $\hat{i} = \frac{1}{T}\sum_t \hat{A}_{t,:}$ and interpolates it to $i \in \mathbb{R}^T$, then sets $C_{\text{avg}} = i^\top V$. Since each row of $\hat{A}$ is softmax-normalized, $\sum_j \hat{i}_j = T$, so $C_{\text{avg}}$ scales with $T$ unless further normalized; this is left implicit and may interact with the gating $s_{\text{mix}}$.
   - The complexity arguments for the interpolation from $\bar{M} \in \{0,1\}^{T \times K}$ to $M^* \in \{0,1\}^{T \times T}$ are mostly relegated to Appendix A.5.3. The main text briefly notes special handling when $\hat{k} = \max(1, \text{round}(kK/T)) = 1$, but the description is easy to miss. Providing a concise formula for the number of non-zeros $|M^*| = \Theta(Tk)$ and the conditions under which this holds would make the claimed $\mathcal{O}(T)$ more convincing.
4. **Empirical scope and baselines are somewhat limited.**
   - Evaluation focuses on BERT-base on three GLUE tasks (MNLI, MRPC, COLA) and on relatively modest OPT models (125M–1.3B) on Wikitext-2 and OpenWebText. This is a reasonable starting point, but:
   - There is no evaluation on standard long-context benchmarks (e.g. Long Range Arena, document QA, code completion) where linear vs quadratic differences might matter more than on Wikitext-2.
   - The GLUE experiments use a fixed sequence length of 256 (Appendix A.4.2), where quadratic attention is not the main bottleneck; SEA's extra overhead may be less justified here.

## CSPaper (continued)

- While strong baselines like Reformer, Performer, Cosformer, Sinkhorn, Synthesizer are included, other competitive sparse long-range transformers such as BigBird, Longformer, or hybrid methods like Scatterbrain are absent. Since Scatterbrain also combines low-rank and sparse components, a direct comparison would be particularly informative.
- The long-context extension experiment to 4k tokens (Table A.5, A.6) is promising, but uses relatively small budgets and self-distillation; more extensive experiments would strengthen the practical relevance claim.

5. **Dynamic k results raise fairness/compute questions.**
   - In Fig. 6 (Wikitext-2), the authors show that increasing $k$ at test time drives perplexity below the vanilla baseline; however, there is no reporting of FLOPs or effective sparsity for these larger $k$. For high $k$, SEA may approach quadratic compute, making the comparison less informative.
   - Similarly, Table A.6 explores "query skipping" and dynamic $k$ on 4k-context LM. While this is a nice knob, many entries correspond to significantly higher latency and memory than the vanilla model, yet are presented in the same table without normalizing for compute. It would be more transparent to overlay these dynamic-$k$ curves on the FLOPs curves of Fig. A.4 or to report achievable perplexity at fixed FLOPs.
6. **Interpretability and utility of estimated attention are mostly qualitative.**
   - The paper argues that SEA "maintains an interpretable attention matrix" and can be used for token importance analysis or pruning. However, the only support is qualitative visualization (Fig. 9, A.1, A.2, A.3, A.10).
   - There is no quantitative evaluation of how well $\hat{A}$ or $A^*$ approximate the teacher attention (e.g., average KL divergence, top-$k$ overlap), nor any downstream use-case such as token merging/pruning or explanation tasks. As a result, the interpretability advantage remains plausible but unproven.
7. **FlatCSR novelty and generality are somewhat limited.**
   - FlatCSR is essentially a CSR-like structure optimized for the particular grouped top-$\hat{k}$ sparsity pattern (causal-per-batch) and implemented in Triton. Table 3 demonstrates that, for this specific pattern on random matrices, FlatCSR is faster and more memory-efficient than COO; however, comparisons to existing block-sparse or segment-sparse kernels (e.g., those underlying FlashAttention and other libraries) are missing.
   - The paper does not discuss how easily FlatCSR integrates into general-purpose frameworks or whether it supports other sparsity patterns beyond those induced by SEA's grouping. As such, its impact may be narrower than suggested.
8. **Some experimental details are under-specified, affecting reproducibility.**
   - While hyperparameter ranges are sketched (e.g., $k \in \{7, 13, 25\}, K \in \{32, 64, 128\}$ on GLUE, Section A.4.2), it is not always clear how $k, K$ were selected for the main headline results (e.g., in Fig. 5a, 7a) or how much hyperparameter tuning was done for baselines vs SEA.
   - The KD loss weights in Table A.4 look somewhat ad hoc (e.g., $\mathcal{L}_{\text{kd}}$ weight 5.0, $\mathcal{L}_{\text{task}}$ weight 0.1), and there is little sensitivity analysis. In practice, correctly balancing these many terms can be non-trivial and may influence results. More detail, or even partial ablations on loss weighting, would help other researchers adopt the method.

## ReviewGrounder

1. **Adaptation cost not fully quantified** – While Section 3.3 describes replacing the teacher's attention with SEA and fine-tuning via KD, the paper does not report the additional training time, data size, or compute needed for this adaptation, which may limit plug-and-play applicability. (Section 3.3)

## ReviewGrounder (continued)

2. **Limited model and task diversity** – Experiments are confined to OPT-125M/1.3B for causal LM and BERT-base for GLUE. Broader evaluation on larger LLMs (e.g., GPT-2/3) or vision transformers would strengthen claims of generality. (Section 4)
3. **CNN decoder architecture details missing** – The decoder is said to be a 3-layer 2-D CNN with channel-expansion $c_h$ and width-reduction $c_s$, but kernel sizes, strides, padding, and channel counts are omitted, hindering reproducibility. (Section 3.1, "CNN Decoder")
4. **Computational overhead not fully broken down** – The latency breakdown (Fig. exp.figure.complexity bottom) shows percentages for dense, FlatCSR, and other ops, but absolute FLOP counts for the Performer encoder, CNN decoder, and top-$k$ selection are absent, making it hard to assess the true cost of the estimator. (Fig. exp.figure.complexity)
5. **Hyper-parameter selection guidance limited** – The paper provides an ablation for grouping strategies (Table method.table.ablation_k) and dynamic-$k$ results, yet offers no systematic procedure for choosing $K$ or the base $k$ for new tasks, nor analysis of sensitivity to these choices. (Table method.table.ablation_k, Section 4.3)
6. **Scalability beyond $2^{13}$ not explored** – Memory and latency are shown up to sequence length $2^{13}$; experiments on longer sequences (e.g., $2^{15}$ or more) would better demonstrate linear scaling claims. (Fig. exp.figure.complexity)
7. **Comparative positioning could be richer** – Related works such as SLA (hybrid sparse-linear for diffusion models) and SPARSEK (learnable differentiable top-$k$) address similar efficiency goals. Direct empirical comparison or discussion of trade-offs would clarify SEA's niche. (Related-work summaries)

## ARD-Review

1. Table 2 does not include a fine-tuned vanilla baseline (same 10k steps, same Wikitext2 data). The caption states "We trained the same number of steps (10k) for each method," but the Vanilla row reports numbers consistent with the pretrained OPT checkpoint's out-of-the-box perplexity rather than a checkpoint after 10k Wikitext2 fine-tuning steps. If vanilla OPT fine-tuned on Wikitext2 for 10k steps also improves from 29.2 / 13.9 PPL, the headline claim that SEA "achieves better perplexity than OPT-1.3B" (26.0 vs. 29.2; 13.5 vs. 13.9 respectively) cannot be cleanly attributed to the SEA architecture rather than the fine-tuning process. A single additional row would resolve the ambiguity and is cheap to run.
2. The comparison against Reformer and Performer in Table 2 is structurally asymmetric in a way the paper partially acknowledges (Section 4.1: "Our method exhibits more latency") but does not fully confront. Reformer and Performer replace attention wholesale; initializing them from a pretrained OPT checkpoint whose weights were trained for quadratic attention is likely to put them far from their own training optima after only 10k steps. SEA, by contrast, preserves most pretrained weights and adds a small set of new parameters around them. This structural asymmetry means the 10k-step comparison partly reflects initialization compatibility rather than architecture quality — a point worth stating explicitly. Section 4.1 gestures at this ("Reformer shows unpredictable performance between the tasks") without fully owning it.
3. SEA is slower than vanilla attention at every sequence length tested in Table 2 (6.76ms vs. 4.88ms for OPT-125M; 9.43ms vs. 6.71ms for 350M; 21.57ms vs. 16.49ms for 1.3B). The linear latency advantage only materializes at very long sequences ($\sim 2^{13}$ and beyond per Fig. 8). The paper's practical framing — "opens the possibility of running large transformers on resource-limited devices" — is accurate for memory but overstated for latency at moderate sequence lengths. Section 4.1 mentions this parenthetically; it deserves a more direct sentence in the efficiency section clarifying the latency crossover threshold.
4. Figure 8 (top-right) shows that quadratic attention runs out of memory above $2^{13}$, and reports 32.72% relative latency at exactly $2^{13}$. The crossover point where SEA's latency matches vanilla (below $2^{13}$) is never reported, leaving the reader unable to identify the practical threshold at which the latency advantage begins. Stating this threshold explicitly would sharpen the efficiency claim.

**ARD-Review (continued)**

5. No ablation for the $V_I$ component in $V_{\text{cat}}$ (Section 3.1). The theoretical motivation — treating FAVOR+ output as an attention matrix when $V = I$ — is approximate because $V_I$ is $d \times d$ identity interpolated to $T \times d$, not $T \times T$. The paper calls this "a natural choice" and does not isolate its contribution to $\hat{A}$ quality. Given the approximate nature of the justification, an empirical ablation carries real weight.
6. No ablation isolating $C_{\text{avg}}$ in the causal output calculation (Section 3.1.1). $C_{\text{avg}}$ injects a global pooled signal bypassing the sparse mask entirely; it contributes to every token's output via the $s_{\text{mix}}$ gate. Table 1 ablates grouped top-$k$ modes but not this component. Its contribution to the causal PPL numbers is unknown.

## C MERGER FILTERING PROMPT

**Filtering Prompt**

### Filtering Discipline (apply aggressively)

The input review sweeps general areas (method soundness, evaluation validity, comparison fairness, evidence strength, internal coherence, significance). This produces some real findings and some category-driven noise. Be aggressive in stripping the noise:

- For every retained weakness, you must be able to point to a specific sentence, equation, figure, or table in the paper that the weakness applies to. If the criticism is framed generally ("the evaluation lacks rigor", "baselines may not be fair", "evidence is weak for the claims") without a concrete anchor, REMOVE it.
- If a weakness reads like an area-of-concern sweep ("could the metric be measuring a proxy?", "are confounders controlled?") rather than a specific identified problem, REMOVE it. The harsh critic was asked to use those areas only as lenses; do not let speculation that surfaced through that sweep enter the final review.
- If the harsh critic asserts something is "fatal" or "structural" but the assertion depends on information not present in the paper (e.g., "the appendix may specify X but…", "assuming Y is the case…"), DEMOTE it to at most Minor or REMOVE it. A fatal flaw must be unambiguous given what is on the page, not a speculative gap.
- If two reviewers raise the same concern in different framings, merge them; do not let duplication inflate the weakness count.
- Default to fewer weaknesses than the inputs contain. The merger's job is to compress, not to union.

Note: For the following rules, REMOVE means moved it to a new section called Removed Points, do not completely remove them from the review

### Hard Rules (absolute, override all other rules)

- REMOVE any criticism that questions the existence, release status, or availability of any model, tool, benchmark, dataset, or reference cited in the paper. If the paper cites it, it exists. This includes phrasing like "not yet released," "does not correspond to currently available systems," "cannot be independently verified," or any reproducibility concern rooted in doubting that a cited entity exists. These reflect reviewer knowledge gaps, not author errors.
- REMOVE criticisms that are factually wrong or misunderstand the paper.
- REMOVE "weaknesses" about unfair comparison with other methods if the asymmetry favors the baseline and not the author's method. This is intentionally asymmetric to prove a stronger point.
- DO NOT mention missing related works, as you do not have external sources to confirm their existence and could be making things up.
- REMOVE pure formatting/style nitpicks.
- REMOVE any criticism about typos, spelling, grammar, punctuation, capitalization, whitespace, line breaks, broken characters, garbled text, missing/extra symbols, or any other formatting artifact. These are parser errors, not author errors — the original submission does not have these issues.

**Filtering Prompt (continued)**

- REMOVE nitpicks about reproducibility such as undisclosed hyperparameters, trivial implementation details, or large artifacts impractical to include in a submission (e.g., complete training logs).
- REMOVE strawman weaknesses that misunderstand the paper content or claiming something the paper already addressed
- REMOVE weaknesses about missing appendix, missing proofs in appendix, or absent references. The parser strips those sections from all papers; they exist in the original submission.
- The harsh reviewer will give weaknesses with grounded paragraph, verify those weaknesses against the paragraph to make sure the weakness is valid
- Many of the harsh reviewer's weaknesses are real but minor (presentation, appendix-deferred proofs, precision nitpicks). Rank by severity, not count: score from the worst flaw that actually threatens the core claim.
- Filter the strengths from the input review. Drop strengths that are generic, superficial, or lack a specific citation or concrete content (examples: this paper addressed an important problem, this paper targeted a interesting question). Drop strengths that conflict with a verified weakness — when a strength and weakness disagree, the weakness wins. Move dropped strengths to Removed Points.
- Be careful with claimed strengths: they can be invalid. Remove strengths that are generic, strengths about whether the problem is important, strengths that are delusional, superficial, sycophantic, and strengths drawn from pure pseudoscience. Only keep strengths that are concrete, specific to this paper, and grounded in real evidence.
- FUNDAMENTAL ISSUES: If any weakness is severe enough to undermine the paper's core claims or it is simply "not even a paper", it overrides all strengths. The overall assessment must reflect this severity rather than averaging strengths and weaknesses or softening the judgment with "could be strong with revisions." However: a weakness only counts as fundamental if it is verifiable from the paper as written — not from speculation about a stripped appendix, missing supplementary, or assumed-but-unverified setup. Speculative-fatal claims (e.g., "if the normalization were X, the reported values would be impossible") should not trigger a score collapse; demote them to Major or Minor and proceed normally.
- Similarly, if the paper made real contributions do not reject just because it has some weaknesses - every paper has some. A strong, well-supported contribution should be scored high — do not pull a clearly strong paper down to the middle out of caution. The same calibration discipline that demands low scores for fatally-flawed papers demands high scores for genuinely strong ones.
- The human finder finds similar weaknesses from other papers, they might not be related to this paper, remove those that are not or barely related.

## Soft Rules (apply judgment)

- WEAKEN criticisms that demand the paper address problems outside its stated scope. A paper about X should be evaluated on whether it does X well, not on whether it also does Y. If the paper explicitly scopes out a direction, criticizing its absence is scope creep. If doing Y would genuinely strengthen the paper, mention it as a nice-to-have.
- WEAKEN weaknesses that are generic or one-size-fits-all and do not harm the core claim. Examples: requesting a larger dataset when the current size is sufficient, adding more models when the model zoo is already adequate.
- WEAKEN weaknesses the authors already address in the paper, even if imperfectly, as long as the addressal is reasonable.
- MOVE TO NICE-TO-HAVE weaknesses that demand methodological practices not standard in the paper's field or setting. Examples: requesting confidence intervals for large-scale benchmarks where single-run evaluation is the norm, demanding theoretical proofs for an empirical systems paper, or requiring user studies for a purely algorithmic contribution. Evaluate the paper against its own community's standards.
- The "Strengthening the Paper on Its Own Terms" section should be considered as minor weakness or similar tier in nice-to-have and not ignored

## Keep Rules

**Filtering Prompt (continued)**

- KEEP criticisms that are factually correct AND substantive, even if only one reviewer raised them.
- KEEP genuine strengths backed by evidence.
- KEEP and EMPHASIZE insightful weaknesses that could help the author improve their paper.
- If the weaknesses identified would, if true, invalidate or severely undermine the paper's core contribution, the review should reflect that clearly. Do not soften the overall tone to appear balanced.

Table 3: Effects of the paper heading on CSPaper scoring accuracy ($n = 200$).

| Metric | With Heading | Without Heading |
|---|---|---|
| Pearson | 0.5245 | 0.5073 |
| Spearman | 0.4829 | 0.4513 |

## D SYCOPHANCY OF LLMS IN REVIEWS

We argue that LLM sycophancy remains a major bottleneck for improving both score correlation and separation in LLM-generated reviews. Our results reflect that on many models, even papers with low human scores sometimes receive inflated ratings from LLMs, suggesting that models are optimizing for encouragement rather than critical accuracy. We tested three models known for sycophancy in Table 4 using a subset of papers with a human average rating of 2.25. The results show that these models sometimes rate the paper above 6 or even 8 when rated directly. Our method can lower their bias.

That models need such harness engineering to stay critical suggests they are optimized for encouragement rather than critical accuracy—a form of over-alignment and toxic positivity that treats success as the norm rather than the result of high-quality work. We borrow the concept of toxic positivity from Guo et al. (2025b). Following Kasneci & Kasneci (2026), we argue that a helpful LLM response should retain a certain degree of criticism, especially in educational and academic contexts.

Meanwhile, the latest Claude and GPT models over-correct, becoming overly harsh and critical, likely due to anti-sycophancy training; this could harm score alignment in the opposite direction and can also force authors to be defensive again. Our current design mainly targets sycophancy (harsh critics, prompts encouraging harshness) matching the behavior of the models available when this paper was developed, and likely will not fix this opposite bias, though prompt and architecture modifications should make that possible. Following Guo et al. (2025b) and Guo et al. (2025a), we argue that the "personality" of LLMs should not be fixed during training but instead be steerable by users or developers through prompts or PEFT, benefiting both calibration-sensitive tasks and value pluralism in domains like creative writing.

Table 4: Sycophancy on the 46 low-rated ICLR papers covered by all nine runs (human mean score 2.25). Acc.: percentage of papers the reviewer recommends accepting; $\geq 6$ / $\geq 8$: percentage of papers rated at or above that score; Bias: mean predicted score minus mean human score.

| | Seed-2.0-Lite | | | | Gemini-3.5-Flash | | | | Gemini-3.5-Flash-Lite | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Acc. | $\geq 6$ | $\geq 8$ | Bias | Acc. | $\geq 6$ | $\geq 8$ | Bias | Acc. | $\geq 6$ | $\geq 8$ | Bias |
| Simple | 47.8 | 32.6 | 6.5 | +2.21 | 39.1 | 37.0 | 6.5 | +2.33 | 76.1 | 76.1 | 30.4 | +4.19 |
| Direct | 100.0 | 100.0 | 26.1 | +4.57 | 58.7 | 58.7 | 15.2 | +3.30 | 100.0 | 100.0 | 78.3 | +5.62 |
| Ours | 10.9 | 21.7 | 2.2 | +1.52 | 2.2 | 2.2 | 0.0 | +0.80 | 58.7 | 58.7 | 2.2 | +3.71 |

## E CALIBRATING HUMAN SCORES

The calibration problem also arises in human reviewing: even when reviewers agree on a paper's strengths and weaknesses, their scoring standards can lead them to assign different scores. This motivates applying our calibration method to human-written reviews, retaining the human assessment

while reducing variation due to individual scoring standards. Chen et al. (2026) propose a related use of LLM scoring through an item-wise subscore method, in which an LLM independently assigns a weight to each strength and weakness and uses the aggregate as a calibration anchor.

**Item-Weight Baseline Implementation** This approach is also similar to the semantic analysis in CSPaper (Cao et al., 2025); however, we do not explicitly use semantics as a proxy for paper quality. In our setup, each item is scored independently, but we do not simply use these scores as the final score; the scores are passed to the LLM as input after the LLM generates the review, and the LLM then reasons and produces the final score. Instead of using an LLM to give the scores directly, we fine-tuned a Qwen embedding model (Qwen3-Embedding-4B) (Zhang et al., 2025) using LoRA, with an added linear head and sigmoid activation, to predict a score for each strength ($S_i$) and weakness ($W_i$). Scores for strengths are forced to be positive and for weaknesses are forced to be negative. The scores are first averaged within the strengths and within the weaknesses, and then these two averages are averaged again. We perform this within-section averaging to prevent the model from exploiting the shortcut of simply counting the number of strengths and weaknesses instead of learning the true contribution (weight) of each item. The final score prediction can be written as follows, where $E$ is the embedding model with a linear layer:

$$s = 5 + \frac{1}{2} \cdot \frac{1}{|S|} \sum_{i=1}^{|S|} 10 \cdot \sigma(E(S_i)) - \frac{1}{2} \cdot \frac{1}{|W|} \sum_{i=1}^{|W|} 10 \cdot \sigma(E(W_i))$$

To train this model, we used both MAE loss for score alignment and RankNet (Köppel et al., 2019) and margin ranking loss for ranking to avoid score collapse. The MAE loss is given a larger weight for extreme scores, also to prevent score collapse.

We compare this item-wise method with our retrieval-based calibration using human reviews from 2026. Given a review $R_{i,j}$ and the score $s_{i,j}$ assigned by reviewer $i$ to paper $j$, each method produces a recalibrated score $s'_{i,j}$. We then compare the original and recalibrated scores through their one-vs-rest correlation and MAE against the mean score of the other reviewers for the same paper, measuring whether calibration brings each individual judgment closer to the rest of the reviewer pool.

As shown in Table 5, our RAG-based approach improves both absolute score agreement and correlation when paired with a sufficiently strong backbone, Claude Sonnet 4.6. The item-weight-based method instead reduces correlation and increases MAE, so the benefit observed here depends on how the calibration is performed.

Table 5: We applied our calibration method to one randomly selected reviewer per paper and compared the result against the rest. Closer, Tie, and Further report the fractions of cases in which the calibrated score is closer to, equally distant from, or farther from the mean score of the remaining reviewers than the original score. The human one-vs-rest here used the same hold-out set as LLMs, so the numbers differ from the main table.

| Model | Method | Pearson | Spearman | MAE | Closer | Tie | Further |
|---|---|---|---|---|---|---|---|
| Human one-vs-rest | - | 0.5902 | 0.5579 | 1.5787 | - | - | - |
| DeepSeek-V4-Flash | RAG (ours) | 0.5403 | 0.4875 | 1.3669 | 0.5473 | 0.0665 | 0.3862 |
| DeepSeek-V4-Flash | Item Weight | 0.4106 | 0.4106 | 2.2124 | 0.3043 | 0.1483 | 0.5473 |
| Claude Sonnet 4.6 | RAG (ours) | 0.6235 | 0.5807 | 1.2258 | 0.5499 | 0.1151 | 0.3350 |
| Claude Sonnet 4.6 | Item Weight | 0.4676 | 0.4585 | 1.8566 | 0.3529 | 0.2174 | 0.4297 |

## F CSPaper Data Leakage Testing

In our initial testing, CSPaper achieved an implausibly high correlation, accompanied by near-perfect separation of accepted and rejected papers: almost all accepted papers receive scores above 5, while almost all rejected papers receive scores below 5. At a cutoff of 5, this yields a decision accuracy of 0.923, well above the 0.859 achieved by the average human score at its optimal threshold of 4.8. Since acceptance decisions are themselves derived from human scores, an external system exceeding that accuracy raises a substantial concern.

Such accuracy is also difficult to reconcile with the inherent randomness of peer review, in which around 50% of papers accepted by one committee would be rejected by another (Beygelzimer et al.,

2023). The most plausible explanation is decision leakage: with all ICLR 2026 papers and decisions publicly available, CSPaper's internet-enabled harness may have retrieved the actual decisions and assigned matching scores. We therefore treat its online result as unreliable.

To confirm this hypothesis, we contacted CSPaper and obtained a limited-access version with tool access disabled, which we evaluated on the same papers and reported as CSPaper; its correlation fell back to a reasonable range. Given that the only change was tool access, we can reasonably confirm that the original CSPaper's high correlation was due to data leakage through its internet access.

Because CSPaper accepts only PDFs, this evaluation uses the original files, including their paper headings. A check on 200 papers confirmed that including the heading does not affect score accuracy (Table 3 in the Appendix), likely because the PDF parser removes it.

## G Data Leakage Clarification

With current LLMs, a benchmark may overlap with a model's pretraining or post-training data. Since labs disclose little about their training data, we cannot rule out that some models have seen the papers and reviews in ARD-Bench. This is especially true because the training data cutoff is unknown for all models except Claude Sonnet 4.6 and Claude Opus 4.6. Sonnet 4.6 has a reliable knowledge cutoff of August 2025, which is before the ICLR 2026 reviews were released, and a complete training data cutoff of January 2026, which is after the reviews were released. Opus 4.6 has a complete training data cutoff of August 2025, which is also before the ICLR 2026 review release.

We consider this a limited concern for several reasons. First, even if some of these papers and reviews appear in the training data, they are unlikely to be included on purpose or given special weight. They would make up a tiny fraction of the corpus, and a model is unlikely to memorize the exact scores attached to a specific paper from such sparse exposure. Second, assuming no meaningful contamination for recent benchmarks is a fairly common practice in LLM evaluation, since full verification is not possible without access to the training data. Third, our results do not look like those of a model recalling answers. In the direct review baseline, its scoring accuracy is low. If the model had memorized the reviews, direct review should already produce accurate scores. Instead, the gains come from our pipeline, which suggests the improvement is due to the method and not to leaked data. Fourth, our baseline, CSPaper, uses similar generation models, so it is exposed to the same potential overlap. Any advantage from leakage would apply to both systems, and the comparison between them remains fair. Finally, Opus 4.6 is completely unaffected by potential leakage, since its training data cutoff predates the release of the ICLR 2026 reviews. Its scoring accuracy with our pipeline is the highest of all backbones we tested (Table 2), which confirms that the gains come from our pipeline.

We also did not deduplicate papers across the DeepReview calibration set and ARD-Bench, which cover different submission years and have no overlapping paper IDs. A small number of papers were revised and resubmitted from 2025 to 2026. Reliably identifying these cases is difficult without introducing false positives because they might have been substantially revised. Given their limited prevalence, the changes between submission versions, and the fact that they usually differ in review scores, we treat these cases as noise rather than data leakage.