Do AI Reviewers Converge? Diversity Collapse in Model-Generated Peer Review

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used to generate or assist reviews. This raises the risk of generative monoculture, where models converge on homogeneous judgments and language patterns, limiting the novelty of scientific progress. We study this phenomenon over 1,943 human reviews from 509 ICLR papers with 10,172 reviews generated by five state-of-the-art LLMs under multiple prompting strategies. We evaluate human and LLM review diversity across several dimensions, including praise-to-critique balance and semantic similarity. Across metrics, LLM-generated reviews exhibit systematic convergence: for example, they are substantially more praise-heavy, and more semantically similar across unrelated papers, suggesting a lack of paper-specific feedback. Specialized review agents reduce—but do not eliminate—these effects. Our work demonstrates the risk of introducing epistemic monoculture into peer review and develops methods to measure it. If left unchecked, this monoculture could narrow the range of perspectives that shape scientific evaluation.

1 Introduction

The rapid progress of large language models (LLMs) has prompted growing interest in their use for scientific knowledge production and evaluation, including the automation of peer review. Journals, conferences, and funding agencies are beginning to explore AI-assisted review pipelines to cope with increasing submission volumes. While such tools promise efficiency and consistency, they also introduce risks that are not yet well understood. One critical concern is *generative monoculture*: the tendency of LLMs to produce outputs that are less diverse than the human-generated distributions they are trained on (Wu et al., 2024). Originally studied in domains such as book reviews and code generation, generative monoculture manifests when models collapse onto a narrow set of outputs—positive sentiment, canonical algorithms, or standardized phrasing—despite the heterogeneity present in their training data.

In the context of scientific peer review, monoculture threatens the very epistemic diversity on which science depends. Effective review is not merely about correctness but about offering heterogeneous perspectives on novelty, rigor, and impact. Yet current LLMs often display *consensus-seeking behavior*, where they converge on "safe" judgments and language patterns rather than exercising genuine independent reasoning. This tendency reflects well-documented side effects of alignment training, where reinforcement learning from human feedback (RLHF) encourages models to produce outputs perceived as broadly agreeable (Christiano et al., 2017; Ouyang et al., 2022). As a result, the model is not truly evaluating a work for itself, but instead reproducing an echo of dominant patterns in its training or prompting context. In practice, such convergence may lead to homogenized reviews that favor mainstream methodologies, undervalue unconventional contributions, and perpetuate existing disciplinary biases.

Recent developments underscore the urgency of this inquiry. Major venues are already experimenting with AI-assisted review: AAAI-26 is piloting the use of LLMs to generate supplementary reviews and discussion summaries (Association for the Advancement of Artificial Intelligence, 2025), ICLR 2025 deployed a large-scale randomized study of a review-feedback agent (Thakkar et al., 2025), and open-source efforts such as OpenReviewer seek to fine-tune models directly on conference review corpora (Idahl and Ahmadi, 2025). Beyond these incremental assistive roles, new initiatives such as *Agents for Science* (Stanford University, 2024) envision autonomous scientific agents that could contribute not only reviews but also literature synthesis and research planning. While these efforts high-

light the potential of AI to reduce reviewer burden and improve clarity, they also foreshadow systemic risks. If such systems converge on homogenized judgments, the very processes that govern knowledge production could become brittle, biased, and resistant to novelty. Our work aims to provide a systematic framework for identifying and measuring these risks before AI review becomes deeply embedded in scientific governance.

This work extends the notion of generative monoculture (Wu et al., 2024) to the domain of scientific review. In particular, our contributions are: (1) empirically demonstrating the potential for review automation to induce monoculture effects across model types, prompting techniques, and across AI tools designed for peer review, (§ 5); (2) propose evaluation paradigms to quantify diversity collapse in AI-generated reviews (§ 3), and (3) discuss whether and how to responsibly integrate LLMs into peer review (§ 6). We argue that as AI systems enter decision-making pipelines in science, it is essential to understand not only their accuracy or efficiency but also their effects on diversity of perspective. By framing review automation through the lens of generative monoculture, and by foregrounding the risks of consensus-seeking convergence, we highlight the broader challenge of preserving epistemic pluralism in an era of increasingly automated scientific governance, and suggest mechanisms to navigate this challenge.

2 Background and Related Work

Generative monoculture and consensus. *Generative monoculture* refers to the tendency of LLMs to produce outputs that are less diverse than the distributions reflected in their training data (Wu et al., 2024). Empirical evidence shows this collapse across domains, from overwhelmingly positive book reviews to template-driven code generation. This collapse can affect human output when working with LLMs (Padmakumar and He, 2024), as well as lead to worsening LLM performance when LLMs are recursively trained on this less diverse data (Shumailov et al., 2024). In this work, we focus on generative monoculture in textual output, e.g. decreasing diversity in sentiment, topic coverage, semantic meaning. Other works have paid closer attention to monoculture in model classification predictions (Jo et al., 2026; Kleinberg and Raghavan, 2021; Creel and Hellman, 2022), demonstrating, e.g., the possibility for correlated errors in LLMs (Kim et al., 2025), or systematic exclusion due to repeating prediction patterns (Jiang and Ng, 2025).

Generative monoculture is closely linked to alignment-induced *consensus-seeking behavior* (or sycophancy) (Christiano et al., 2017; Ouyang et al., 2022). Recent works suggest that techniques such as RLHF incentivize outputs perceived as agreeable, promoting deferential convergence at the potential expense of independent evaluative reasoning. While beneficial for safety and politeness, such dynamics may suppress contrarian perspectives and attenuate critical judgment—limitations that are particularly consequential in evaluative contexts such as grading, hiring, and peer review.

AI in scientific review. AI systems are increasingly integrated into peer-review workflows. Deployments range from assistive mechanisms that generate supplementary reviews and summarize discussions to agents designed to enhance review clarity and structure (Thakkar et al., 2025; Robertson, 2023; PaperReview.ai, 2024). Journals similarly adopt AI tools for language refinement, statistical checks, and preliminary screening (Association for the Advancement of Artificial Intelligence, 2025). Beyond these incremental uses, recent initiatives envision autonomous scientific agents capable of literature synthesis, experimental planning, and review itself (Stanford University, 2024).

Risks and open questions. A growing literature has begun to question the suitability of AI systems for use in peer review, with several works pointing to bias in AI reviews across demographic groups (Vasu et al., 2025), low overlap with human review and decisions (Akella et al., 2025), and generally insufficiently critical feedback (Huang et al., 2025). However, comparatively little attention has been paid to epistemic risks arising from diversity collapse. If AI-generated judgments become widespread, homogenized reasoning may systematically narrow evaluative perspectives. Our work addresses this gap by extending the framework of generative monoculture to scientific review.

3 Problem Formulation

Let $\mathcal{P}$ denote the space of submitted scientific papers and $\mathcal{R}$ the space of possible reviews. For a given paper $p \in \mathcal{P}$, let D_p^{human} represent the empirical distribution of reviews written by human reviewers, and let D_p^{model} denote the distribution of reviews generated by an AI system. Reviews may

be represented as structured tuples

$$r = (\text{score}, \text{comments}, \text{recommendations}) \in \mathcal{R},$$

where each component can be further decomposed into measurable attributes (e.g., sentiment, novelty assessment, methodological critique).

Definition 1 (*Generative Monoculture in Peer Review*)**.** We follow Wu et al. (2024)'s definition of generative monoculture. Let D_p^{human} and D_p^{model} denote the distributions of human and model reviews for paper p, and let $\{\phi_k : \mathcal{R} \to \mathbb{R}^{d_k}\}_{k=1}^{K}$ be feature mappings, where each ϕ_k captures a specific review attribute (e.g., sentiment, topical focus, critique structure). Let $\text{Disp}_k(\cdot)$ denote an attribute-specific dispersion functional applied to the induced feature distribution (e.g., entropy, covariance trace, or maximum pairwise distance). We define generative monoculture as the existence of a salient review attribute k for which model-generated reviews exhibit reduced dispersion relative to human reviews:

$$\Delta_k(p) = \text{Disp}_k\big(D_p^{\text{human}}\big) - \text{Disp}_k\big(D_p^{\text{model}}\big) > \epsilon_k.$$

Problem statement: We study the problem of detecting and characterizing generative monoculture in AI-assisted peer review. In particular, we investigate whether model-generated reviews exhibit reduced content variability, specificity of critique, or directional evaluative skew relative to human baselines, and quantify how such effects manifest.

3.1 Operational Measures of Monoculture

Generative monoculture may appear through convergence in several aspects of review behavior. In particular, reviews may become similar in evaluative tone, structural balance between praise and critique, semantic content, or linguistic expression. To capture these dimensions, we select metrics that measure dispersion in sentiment, review structure, semantic similarity, and lexical–syntactic diversity. Each metric corresponds to a feature mapping ϕ_k in Definition 1, and its associated statistic serves as an observable proxy for diversity across reviews. We operationalize generative monoculture through attribute-specific empirical statistics designed to capture dispersion, similarity, and evaluative structure in reviews. These measures instantiate concrete dispersion functionals from Definition 1 and serve as observable indicators of diversity collapse. All statistics are computed at the review level and aggregated by source and prompting condition. More details are presented in Appendix A.

1. *Average Sentiment Score:* Mean sentence-level polarity in [-1,1] using `RoBERTa-VADER`.
2. *Sentiment Entropy:* Entropy over discretized sentiment bins, measuring tonal dispersion.
3. *Praise-to-Critique Ratio:* Ratio of positive to negative clauses using polarity thresholding, measuring directional evaluative bias.
4. *Inter-Review Similarity:* Cosine similarity between review embeddings produced by `google/embeddinggemma-300m`, measuring the degree to which reviews of different papers become semantically alike and hence more homogenized across sources.
5. *Lexical and Syntactic Diversity:* A suite of length-normalized lexical and syntactic diversity metrics measuring linguistic dispersion in wording and structure.

4 Experimental Setup

This section describes our experimental protocol used to analyze dispersion and diversity in human and AI-generated peer reviews. Our design isolates sources of variability arising from model choice, prompting strategy, and inference configuration, enabling controlled comparisons of distributional properties across review populations.

4.1 Dataset

To our knowledge, there exists no publicly available dataset designed for studying diversity and dispersion in peer review. To enable systematic analysis, we construct a new benchmark using the OpenReview API. The resulting dataset, which we refer to as ICLR-REVIEWS, comprises 509 papers from ICLR 2022–2023, including 331 accepted and 269 rejected submissions. The dataset contains 1,943 human peer reviews, averaging 3.81 reviews per paper. Abstracts, manuscript PDFs, and peer reviews are available for *all submissions*. Importantly, review formats differ across years due to changes in conference templates and reviewing guidelines. These structural variations are preserved to reflect realistic review distributions; details are provided in Appendix E. All measures are computed exclusively on (the original and generated) review text.[1]

4.2 Generating AI Reviews

For each paper p, we model human evaluations as an empirical distribution D_p^{human} induced by Open-

[1] All generated review artifacts, prompts, and the ICLR-REVIEWS dataset will be made publicly available upon paper acceptance to support reproducibility and follow-up research.

Review peer reviews. We similarly construct an empirical model distribution D_p^{model} by sampling AI reviews under each experimental condition.

Models. We evaluate five LLMs. These include GPT-4o, o3-mini, Gemini-2.5-Pro, Claude-3.5-Sonnet-Pro, and Claude-3.7-Sonnet with extended reasoning [2].

Prompting conditions. AI critiques are generated under four prompting styles designed to vary reasoning structure and exemplar exposure: (i) Direct Prompting (DP), (ii) One-Shot Direct (1S), (iii) Chain-of-Thought (CoT) (Wei et al., 2023), and (iv) One-Shot CoT (1S-CoT). For all prompting conditions, the complete manuscript PDF is provided to the model to ensure access to both textual and visual content (e.g., figures, diagrams, and tables) during review generation. Detailed prompts (and prompting conditions) are in Appendix B. We evaluate multiple prompting styles to examine whether observed monocultural effects are intrinsic to model behavior or sensitive to inference-time conditioning. Prompting strategies alter reasoning structure and exemplar exposure, both of which influence output variability and distributional spread. Varying prompts therefore serves as a controlled mechanism for probing the stability of dispersion and bias effects. We analyze 10,172 LLM-generated reviews[3] (analysis in Appendix E).

4.3 Review Validity Assessment

To verify that AI-generated reviews constitute meaningful evaluations rather than degenerate outputs, we conduct a targeted qualitative validity assessment. We randomly subsample reviews from five papers, yielding a set of 25 reviews spanning AI-generated evaluations across models. Domain experts assessed whether each review was (1) substantively relevant to the paper and (2) evaluated the overall quality of the LLM-generated review, including the accuracy of the contribution summary, novelty assessment, connections to prior work, and coherence of critiques; each review received 2 evaluations. Annotators rated relevance and quality on a 1–5 scale (5 = highest), and could provide free-form comments. Detailed evaluation criteria and annotation protocols are provided in Appendix D.

Results. Overall, the assessment indicates that AI-generated reviews are relevant to paper content: relevance scores ranged from 3.65 (o3-mini) to 4.85 (Gemini-2.5-Pro); with three out of five models tested having an average relevance score above 4. Quality assessments were lower, but still overall seen as acceptable, with average scores largely hovering around 3 (3.75, 3.40, 2.9. 2.85), with the exception of o3-mini at a 2.05. A full table is displayed in Appendix D.1, with detailed diagnosis.

5 Measuring Monoculture

We present our results quantifying generative monoculture in LLM-generated reviews across the metrics introduced in § 3 across a variety of LLMs and prompting techniques, as well as specialized review generators. Due to space limitations, some results are presented in Appendix C.

Key takeaways. Our analysis across five metrics reveals several consistent patterns in LLM-generated peer reviews:

- *Praise-heavy evaluation and higher sentiment: risk for overly positive reviews.* Our work shows that even when specifically prompted to be critical, LLMs produce substantially more positive text than humans. Human reviews exhibit more polarized and varied sentiment, whereas LLM-generated reviews cluster closer to neutral tones (Appendices C.1 and C.2). Even more importantly, LLMs display higher praise-to-critique ratios than human reviewers (often $\geq 2\times$) indicating a systematic tendency toward praise (§ 5.1), running a risk for overly-positive reviews.
- *Lack of specificity to individual paper.* LLM reviews are more semantically similar across different papers than human reviews, and this similarity is less sensitive to topic (§ 5.2). This suggests that LLM reviews are less specific to the individual paper, another major problem.
- *Closer to human performance for specialized LLM review tools.* While many of these patterns continue for specialized review tools tested in § 5.3, they are much less dramatic, and often the distance between human and LLMs are indistinguishable (e.g., on praise-to-critique ratio). This suggests that if LLM-reviews are to be used, higher quality reviews will come from specialized tools which use e.g., agentic methods to consider multiple perspectives on the work (PaperReview.ai, 2024).
- *Lexical richness but syntactic uniformity: human and LLM review writing styles are still distinct.* Stylistically, LLM reviews show higher lexical

[2] All prompting was done between July 2025 and March 2026

[3] 8 reviews contained incomplete information (or errors) and were discarded.

diversity but lower syntactic diversity than humans, suggesting broader vocabulary use alongside more stereotyped grammatical structures and thus that review writing styles remain distinct (see Appendix C.5).

5.1 Sentiment and Praise-to-Critique Ratios

Our analysis compares the sentiment composition of LLM-generated reviews and human reviews using three complementary statistics: (1) the praise-to-critique ratio (PCR), defined as the ratio of positive to negative sentences, (2) the mean sentiment score, and (3) sentiment entropy, which measures diversity of sentiment. We treat PCR as the primary indicator of review balance: human reviews exhibit a praise-to-critique ratio of 1.918 (32,746 sentences: 65.73% positive, 34.27% negative). All retained LLM models produce significantly higher ratios (Figure 1): O3-MINI scores 4.272 ($t = -23.10$, $p < 0.0001$), GPT-4O scores 3.616 ($t = -22.28$, $p < 0.0001$), CLAUDE-3.5-SONNET-PRO scores 3.150 ($t = -16.05$, $p < 0.0001$), CLAUDE-3.7-SONNET scores 2.790 ($t = -13.69$, $p < 0.0001$), and GEMINI-2.5-PRO scores 2.123 ($t = -3.47$, $p = 0.0005$). Thus, all models produce substantially more praise relative to critique than human reviewers. One notable exception, GEMINI-2.5-PRO, is close to human praise-to-critique ratio (Table 3 in Appendix C.3).

Effects on sentiment. Results (Tables 1 and 2, Figures 6 and 7; see Appendices C.1 and C.2) show that base LLMs consistently produce more positive reviews than humans, even when explicitly prompted to be discerning (Appendix B). Mean sentiment scores show a consistent but weaker trend: LLM reviews are closer to neutral on average, while human reviews exhibit more polarized sentiment. Sentiment entropy further indicates that human reviews contain greater variability in tone, whereas LLM-generated reviews are more uniform.

Prompting effects. Prompting style affects sentiment statistics within each model but does not eliminate between-model differences (Table 3). CoT prompting generally produces the highest praise-to-critique ratios (e.g., GPT-4O CoT 3.956, CLAUDE-3.5-SONNET-PRO CoT 3.588), while single-shot prompting yields lower ratios (e.g., GPT-4O 3.097, GEMINI-2.5-PRO 1.697), yet all settings remain above the human baseline.

Takeaway: Across all models and prompting styles, LLMs produce reviews with substantially higher praise-to-critique ratios than humans, even when instructed to be critical. Supporting metrics (mean sentiment and entropy) indicate that LLM reviews are also more uniform and less emotionally varied, suggesting a systematic bias toward positive and non-confrontational feedback.

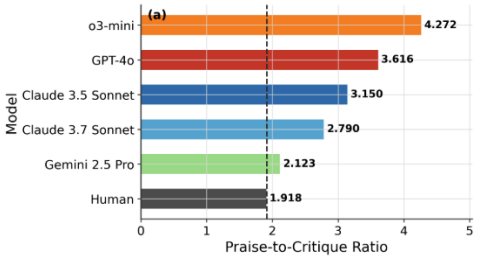

Figure 1: Praise-to-critique ratio comparison. The values are sorted in ascending order, and the dashed line marks the human baseline.

Takeaway: Overall, LLMs produce much more positive reviews than humans, *even when given explicit prompting to "be discriminating"* (see Appendix B for prompts). In particular, LLMs show substantially higher praise-to-critique ratios than human reviewers (often over twice as high) indicating a systematic tendency toward praise.

5.2 Inter-Review Similarity

Overall, we find that LLM-generated reviews are less semantically diverse across papers. This suggests that base LLM reviews bring up fewer unique, paper-specific ideas and critiques in their reviews, but have a boilerplate template across papers that goes beyond structure to meaning.

We conduct two experiments to characterize semantic similarity in reviews. Experiment 1 evaluates whether LLM-generated reviews exhibit stronger semantic similarity than human reviews by randomly across *unrelated* papers by selecting random paper pairs and comparing the similarity of human and LLM reviews across these paper pairs. Intuitively, these reviews should *not* display high semantic similarity as the papers they were generated for are likely unrelated. Experiment 2 examines whether this similarity is conditioned on topic, by following the same procedure but conditioning by paper topic, and testing both within and between topic similarity. Here, intuitively, within-topic paper reviews should display *more* similarity as the ideas discussed are likely more similar, while between-topic papers should be more diverse. We find that LLM reviews are more similar across randomly selected papers, *and* display similar semantic similarity within-topics and across-topics,

suggesting the reviews are less specialized to the specific paper.

Experimental details. We sample $N = 100$ random paper pairs (p_1, p_2) from our dataset. For each pair, we randomly select one LLM-generated review and one human review for each paper. To measure similarity between reviews, we embed each review using a sentence embedding model $E(\cdot)$ based on `google/embeddinggemma-300m`, and compute cosine similarity between the resulting vectors. For each sampled pair (p_1, p_2), we compute

$$\mathrm{Sim}_t(p_1, p_2) = \cos\big(E(r_{p_1}^t), E(r_{p_2}^t)\big) \quad (1)$$

where r_p^t denotes the sampled review of type t (LLM or human) for paper p. We repeat this for all sampled pairs and compare the resulting similarity distributions for LLM and human reviews.

To compare across topics, we group papers using the `subtopic` field provided in the ICLR-Reviews dataset, merging related labels into seven coarse areas: COMPUTER VISION, GRAPH NEURAL NETWORKS, GENERATIVE MODELS, REINFORCEMENT LEARNING, NLP, THEORY, and OTHER (with smaller categories merged into OTHER). We consider all pairs of subtopic groups, including both within-topic pairs and between-topic pairs. For each pair of groups (A, B), where A and B denote subtopic categories (with $A = B$ corresponding to within-topic comparisons), we sample $N = 100$ random paper pairs with one paper drawn from group A and the other from group B. For each sampled paper pair, we select one LLM review and one human review per paper using the same one-review-per-paper procedure as in Experiment 1, and compute cosine similarity between the resulting vectors.

Results. All LLM-generated reviews are *more similar* than human reviews on the same 100 random paper pairs, indicating stronger cross-paper convergence under this semantic similarity metric. These differences are summarized in Table 4 and visualized in Figure 2. This result supports the monoculture hypothesis in the conference-level setting, showing that LLM reviews exhibit higher similarity *across reviews for unrelated papers*.

Further, analyzing the results on Experiment 2, showing within-and-across topic similarity, LLM reviews remain more similar than human reviews for both within-topic and between-topic comparisons. In particular, human reviews exhibit stronger topic-based similarity: similarity is higher for

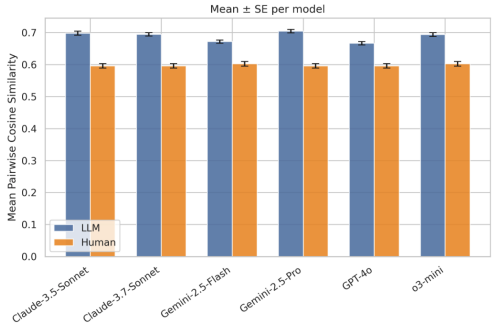

Figure 2: Distribution and means of cosine similarity over 100 random paper pairs (for experiment 1). LLM reviews are consistently more similar than human reviews across all five models.

human reviews within the same subtopic than across different subtopics, whereas LLM similarity decreases much less between within-topic and between-topic pairs, indicating more topic-agnostic convergence. Aggregated statistics are reported in Table 5, and Figure 3 visualizes the within-topic versus between-topic comparisons across models.

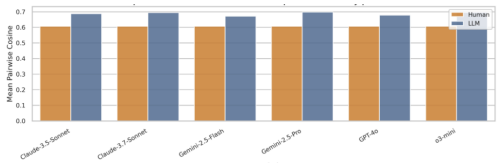

Figure 3: Mean cosine similarity aggregated over within-topic and between-topic paper pairs (for experiment 2). Humans show a clearer within>between pattern, while LLM similarity remains comparatively high across both settings.

Takeaway: These results suggest that human review similarity is more sensitive to research area and specific paper, while LLM review similarity is more uniform across subtopics, suggesting that the LLM takes fewer paper-specific details into account in its review; this is a severe drawback.

5.3 Specialized Review Platforms

While many reviewers submitting LLM-generated (or heavily LLM-modified) reviews (Liang et al., 2024) are likely using base LLMs such at ChatGPT, there also exist specialized tools for generating peer reviews, such as those that have been integrated into conference review processes at e.g., ICLR (Thakkar et al., 2025). To more completely study the impact of generative monoculture on peer review, we test two such open-source tools with the same methods from the above experiments: the Stanford Agentic Reviewer (SAR) (PaperRe-

view.ai, 2024) and the Bibby Agentic Reviewer (BAR) (Bibby, 2024).

We find that while both platforms demonstrate some evidence of monoculture, they have much lower gaps from human reviews. SAR in particular demonstrates less monoculture according to our metrics than other models, potentially due to its agentic design encouraging it to consider "multiple perspectives" (Jiang and Ng, 2025) on the work during review creation. While we believe further study is necessary, this suggests that if LLM-generated reviews are to be used, it is imperative that the LLMs be specially crafted for the the task.

Sentiment and Critique Ratios. We run all experiments across a random sample of 20 papers from our original dataset, a sample size constrained by the strict upload quotas and usage limits of the specialized review platforms.

As throughout the earlier experiments, both SAR and BAR exhibit more positive reviews overall than human reviews, shown in Figure 4: SAR is the most neutral of the three source types (mean polarity -0.049), compared to BAR (-0.064) and the human panel (-0.094). However, unlike base models, aggregate sentiment *entropy* is similar across specialized LLM reviewers and humans (SAR 0.530, BAR 0.522 human 0.513, Figure 16 in Appendix F), suggesting broadly comparable tonal balance at the corpus level. Interestingly, we find slightly lower praise-to-critique ratios between specialized reviewers and humans (LLMs more negative), see Figure 17 in Appendix F. However, there is no statistical significance. This result may be due to the small sample, which is also a highly positive subsample for human reviews as seen by the difference in human praise-to-critique ratio from the overall average (3.22 vs 1.918). Given the small sample size, we believe that this result should be interpreted to mean there is much less difference in the amount of critique given in specialized LLM review tools and humans than base models and humans, however, the overall issue of higher sentiment remains.

Semantic Similarity. To test for semantic similarity, we follow the procedure for Experiment 1 from § 5.2. We do not perform Experiment 2 on within- and cross-topic convergence due to small sample size. We find that both SAR and BAR have higher similarity across papers, however, SAR's difference is not statistically significant, $p = 0.27$). BAR's is higher ($p = 0.012$); yet still,

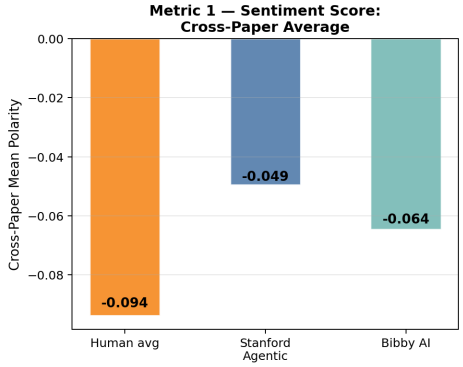

Figure 4: Distribution of average sentiment scores across human baselines and review platforms.

effect sizes (+0.019 (BAR) and +0.045 (SAR)) are substantially smaller than those observed for any base LLM in Experiment 1 (+0.07–+0.09, all $p < 10^{-10}$, see Figure 5). Interestingly, however, BAR and SAR reviews are more semantically similar to each other (mean 0.784) than either is to the human reviews (BAR 0.587; SAR 0.661), establishing a clear AI-versus-human cluster boundary across all 20 papers (see Appendix F).

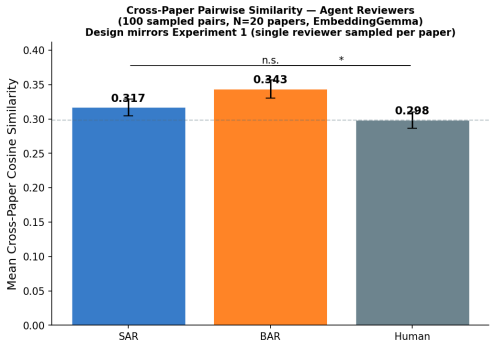

Figure 5: **Left:** Mean cross-paper pairwise cosine similarity ($\pm$SE) for SAR, BAR, and humans (100 sampled pairs, Experiment 1 design, single reviewer per paper, `EmbeddingGemma`). SAR is not significantly different from human ($p = 0.27$); BAR is marginally higher ($p = 0.012$). Both effect sizes are 2–4$\times$ smaller than any base LLM in Experiment 1.

Takeaway(s): Specialized LLM reviews conform with prior results in some ways– higher sentiment for reviews overall, some evidence for lower semantic differences across reviews, and higher lexical diversity: in some, displaying more varied vocabulary but lower concentration of unique ideas overall. However, these signals are weaker: in cross-paper comparisons, BAR's elevation above the human baseline is an order of magnitude smaller than base LLM monoculture; SAR shows no significant difference at all. A stronger monoculture signal for both agents is therefore inter-agent stylistic conver-

gence, where SAR and BAR demonstrate a shared AI semantic space different from human reviews.

6 Discussion

Our results demonstrate a clear risk of generative monoculture in LLM-generated reviews, most prominently for base LLM prompting. Curated LLM-based reviewing tools reduce this effect, but do not fully eliminate it. These findings suggest that monoculture arises from both model-level tendencies and evaluation dynamics, and have implications for how LLMs should be integrated into peer review pipelines.

Prompting vs. model-level effects in generative monoculture. Prompting choices can substantially influence the severity of generative monoculture. Naive attempts to elicit more critical feedback (e.g., instructing the model to be "discerning" or to focus on weaknesses) had limited effect, but some prompting strategies produced consistent shifts across models. In particular, chain-of-thought prompting systematically changed sentiment polarity within the same model, indicating that review tone is partially controllable through prompting rather than fixed by the model alone.

More promising results arise from explicitly incorporating multiple perspectives. Agentic or persona-based prompting, as used by SAR (Stanford University, 2024), and related multi-perspective approaches shown to reduce monoculture in other contexts (Wang et al., 2025; Sorensen et al., 2024), tend to produce reviews that are less uniform and closer to human behavior. These results suggest that monoculture is not inevitable, but depends strongly on how LLMs are used.

Instruction-following vs. independent evaluation. Another possible mechanism underlying monoculture is that LLM reviews reflect instruction-following and text conditioning rather than independent evaluation. Because the model is prompted with the paper itself, the generated review may inherit the framing of the input instead of critically reassessing it. For example, when a paper describes its contribution as novel, the model may repeat or lightly rephrase this claim rather than evaluate novelty independently. In this sense, LLM reviewing can resemble structured summarization with critique-like language, rather than fully independent peer review. This behavior is consistent with known forms of sycophancy and epistemic deference in LLMs, where the model treats the provided text as authoritative rather than as a claim to be challenged (Li et al., 2025).

Qualitative evidence from annotator feedback. We also examined free-form annotator comments on review quality. Across all 50 generated reviews (25 papers $\times$ 2 systems), the most common negative feedback was that the reviews were too vague or generic (28.57% of comments), that they failed to identify limitations or weaknesses (17.14%), and that the review read more like a summary than a critique (11.43%). These observations are consistent with the quantitative results, and suggest that monoculture manifests not only in sentiment statistics, but also in reduced specificity and critical depth. A full list of themes is provided in Appendix D.2.

Implications for the use of LLMs in peer review. Given the measurable risk of generative monoculture, we believe that venues integrating LLM-generated reviews into their pipelines should explicitly test for uniformity and lack of critique. Our results further suggest that allowing reviewers to rely on arbitrary base LLM prompting is likely to produce lower-quality and more homogeneous feedback. Instead, reviewers should be encouraged to use curated tools that are designed to promote critique, incorporate multiple perspectives, and generate paper-specific feedback. Even with such tools, however, careful evaluation remains necessary, as current systems reduce but do not fully eliminate monoculture effects.

7 Conclusion

We investigated the risk of generative monoculture in LLM-generated peer reviews and found that base LLM prompting produces reviews that are more uniform, more positive, and less critical than human reviews across multiple metrics, including praise-to-critique ratio, sentiment distribution, and cross-paper similarity. These effects persist even when models are instructed to be discerning, suggesting that monoculture arises from both model-level tendencies and instruction-following behavior. Curated and agent-based reviewing systems reduce these effects, but do not fully eliminate them. Our results therefore suggest that integrating LLMs into peer review without careful design and evaluation may introduce systematic bias toward uniform and overly positive feedback. We recommend that LLM-based reviewing tools be evaluated not only for correctness, but also for diversity of judgment.

8 Limitations

Our study has several limitations. First, our experiments are conducted on a fixed set of papers and models, and results may differ for other domains, review formats, or future model versions. We focus on conference-style machine learning papers written in English, and the degree of monoculture may vary across other scientific fields, languages, or review cultures. In addition, due to the cost of running large-scale experiments (approximately $1,000 USD in API usage), our evaluation is limited to a single conference dataset, and broader coverage may reveal different patterns.

Second, our analysis relies on automated metrics such as sentiment, praise-to-critique ratio, and embedding similarity. While these metrics capture important aspects of review diversity, they do not fully measure review quality or correctness. A review could be diverse but incorrect, or uniform but accurate, and our work does not attempt to evaluate the scientific validity of the feedback itself.

Third, the behavior of LLM-based reviewers depends strongly on prompting, system design, and model version. Although we evaluate several prompting styles and curated tools, we cannot exhaust the full space of possible reviewer configurations. Moreover, LLMs evolve rapidly. Our experiments were conducted using model versions available between July 2025 and March 2026, and future models or updated system prompts may exhibit different behavior.

Finally, our study focuses on textual outputs rather than the full peer-review process. Real-world reviewing involves interaction, rebuttal, discussion, and program committee deliberation, which may change how monoculture manifests in practice.

9 Ethical Considerations

This work studies the use of LLMs in peer review, a setting where errors or bias may influence scientific evaluation. Our findings suggest that naive use of LLM-generated reviews may introduce systematic bias toward uniform and overly positive feedback, which could affect acceptance decisions or reduce the effectiveness of the review process. We therefore caution against deploying LLM-based reviewing systems without careful validation.

At the same time, LLMs may provide benefits when used appropriately, for example by assisting human reviewers, generating drafts, or providing additional perspectives. Our results indicate that curated or multi-agent systems can reduce, though not eliminate, monoculture effects, highlighting the importance of careful system design.

We also emphasize that our goal is not to discourage research on LLM-assisted reviewing, but to identify potential failure modes. Understanding these risks is necessary to ensure that such systems are used responsibly and do not undermine the diversity of opinions that peer review is intended to provide.

References

Akhil Pandey Akella, Harish Varma Siravuri, and Shaurya Rohatgi. 2025. Pre-review to peer review: Pitfalls of automating reviews using large language models. *arXiv preprint arXiv:2512.22145*.

Association for the Advancement of Artificial Intelligence. 2025. Aaai launches ai-powered peer-review assessment system. https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/. Accessed: 2026-03-15.

Bibby. 2024. Bibby agentic reviewer (bar). Accessed: 2026-03-15.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Kathleen Creel and Deborah Hellman. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43.

Shengzhi Huang, Yong Huang, Yinpeng Liu, Zhuoran Luo, and Wei Lu. 2025. Are large language models qualified reviewers in originality evaluation? *Information Processing & Management*, 62(3):103973.

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

Yixing Jiang and Andrew Ng. 2025. Tech overview. https://paperreview.ai/tech-overview. PaperReview.ai Agentic Reviewer technical overview.

Nathanael Jo, Nikhil Garg, and Manish Raghavan. 2026. The subjectivity of monoculture. *arXiv preprint arXiv:2602.24086*.

David S. Kerby. 2014. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11.IT.3.1.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated errors in large language models. *ICML*.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Jinzhe Li, Gengxu Li, Yi Chang, and Yuan Wu. 2025. Don't take the premise for granted: Evaluating the premise critique ability of large language models. *EMNLP*.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? *ICLR*.

PaperReview.ai. 2024. Paperreview.ai technical overview. https://paperreview.ai/tech-overview. Accessed: 2026-03-15.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *ICML*.

Stanford University. 2024. Agents for science. https://agents4science.stanford.edu. Accessed: 2026-03-15.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*.

Sai Suresh Macharla Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz. 2025. Justice in judgment: Unveiling (hidden) bias in llm-assisted peer reviews. *arXiv preprint arXiv:2509.13400*.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual prompting for improving llm generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6378–6400.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*.

Appendix

A Methodology

Preliminaries. A review R is segmented into n sentences,

$$R = \{s_1, \ldots, s_n\},$$

and contains T tokens in total. At the sentence level, our sentiment scorer provides: (i) a scalar sentiment score $\sigma(s_i) \in [-1, 1]$, and (ii) when needed, a categorical sentiment distribution $\pi(s_i) \in \Delta^{K-1}$ over K sentiment classes. Unless otherwise noted, we normalize quantities by the number of sentences n or the number of tokens T so that metrics are comparable across reviews of different lengths.

We fix a neutrality threshold $\tau > 0$ and define positive and negative sentence indicators:

$$\mathbb{I}^+(s_i) = \mathbb{I}[\sigma(s_i) \geq \tau], \qquad \mathbb{I}^-(s_i) = \mathbb{I}[\sigma(s_i) \leq -\tau].$$

When $\tau = 0$, these reduce to simple sign-based indicators.

For field selection, for ICLR 2022 papers we analyze the `main_review` field, while for ICLR 2023 papers we analyze `strengths_and_weaknesses`. In addition, we include `summary` for both years wherever applicable.

A.1 Sentiment Score

We quantify review tone using a `RoBERTa` sentiment classifier[4] applied at the sentence level. For each sentence s_i, the model outputs class probabilities

$$(p_{\text{neg}}, p_{\text{neu}}, p_{\text{pos}}).$$

We convert these probabilities into a signed polarity score by selecting the most confident *non-neutral* class. Concretely, if the maximum-probability class is neutral, we instead take the second-highest class so that the resulting label is always either positive or negative. Let

$$c^* \in \{\text{neg}, \text{pos}\}$$

denote the selected class and let $s = p_{c^*}$ be its probability. We then define the sentence-level polarity as

$$\sigma(s_i) = \begin{cases} +s, & \text{if } c^* = \text{pos}, \\ -s, & \text{if } c^* = \text{neg}. \end{cases}$$

Thus $\sigma(s_i) \in [-1, 1]$, with positive values indicating praise, negative values indicating criticism, and larger magnitude indicating higher confidence.

[4] `j-hartmann/sentiment-roberta-large-english-3-classes`

The review-level sentiment score is defined as the mean sentence polarity:

$$\mathrm{Sent}(R) = \frac{1}{n}\sum_{i=1}^{n}\sigma(s_i).$$

This gives a length-normalized measure of the overall tone of the review.

To reduce the effect of neutral filler or weakly valenced sentences, we also optionally compute a gated version:

$$\mathrm{Sent}_\tau(R) = \frac{\sum_{i=1}^{n}\sigma(s_i)\cdot\mathbb{I}[|\sigma(s_i)|\geq\tau]}{\sum_{i=1}^{n}\mathbb{I}[|\sigma(s_i)|\geq\tau]+\epsilon},$$

where $\tau > 0$ is a neutrality threshold and $\epsilon > 0$ is a small constant for numerical stability. This variant averages only over sentences whose absolute sentiment exceeds the threshold, thereby focusing the metric on more clearly evaluative content.

A.2 Sentiment Entropy

To quantify how much a review mixes praise and critique across its sentences, we compute normalized sentiment entropy within each review. This metric captures tonal heterogeneity: reviews that contain a balanced mixture of positive and negative statements will have higher entropy, whereas tonally uniform reviews will have lower entropy.

As in the sentiment-score computation, each sentence is assigned a sentiment label using the same `RoBERTa` classifier, with the same non-neutral decision rule: if the model's top class is `neutral`, we instead take the second-highest class so that the selected sentiment is positive or negative. This keeps the sentence-labeling procedure consistent with the polarity definition above.

There are two equivalent ways to express the computation.

Label-fraction formulation. Let

$$(p_+, p_0, p_-)$$

denote the fractions of sentences in review R labeled positive, neutral, and negative, respectively. Under our non-neutral reassignment rule, p_0 is typically zero, but we keep the three-class notation for completeness and consistency. We compute the Shannon entropy

$$H(R) = -\sum_{c\in\{+,0,-\}} p_c \log_2 p_c,$$

with the convention $0\log 0 := 0$. We then normalize by the maximum entropy for three classes:

$$\mathrm{SEnt}(R) = \frac{H(R)}{\log_2 3} \in [0,1].$$

Higher values indicate greater tonal diversity within the review, while lower values indicate a more uniform tone.

Distribution-aggregation formulation. More generally, if the classifier provides a sentence-level categorical distribution

$$\pi(s_i) \in \Delta^{K-1},$$

we can aggregate these distributions across sentences:

$$\bar{\pi}(R) = \frac{1}{n}\sum_{i=1}^{n}\pi(s_i),$$

and compute normalized entropy as

$$\mathrm{SEnt}(R) = \frac{H(\bar{\pi}(R))}{\log K}, \qquad H(p) = -\sum_{k=1}^{K} p_k \log p_k.$$

This formulation uses the full probabilistic output of the classifier rather than only hard sentence labels. In our three-class sentiment setting, it reduces to the same general idea of quantifying within-review sentiment heterogeneity on a normalized $[0,1]$ scale.

A.3 Praise-to-Critique Ratio

We measure the balance between positive and negative feedback by classifying sentences as praise or critique according to their sentiment polarity. In the simplest version, a sentence is treated as praise if its polarity is positive and as critique if its polarity is negative; sentences with exactly zero polarity are excluded. That is, we distinguish praise and critique by polarity > 0 versus < 0.

A binary count-based praise-to-critique ratio is already length-invariant by construction, because both the numerator and denominator scale with the number of sentences. However, binary counts ignore sentiment intensity. To incorporate both frequency and strength, we define intensity-weighted praise and criticism masses:

$$P(R) = \frac{1}{n}\sum_{i=1}^{n}\max(\sigma(s_i),0)\cdot\mathbb{I}^+(s_i),$$

$$C(R) = \frac{1}{n}\sum_{i=1}^{n}\max(-\sigma(s_i),0)\cdot\mathbb{I}^-(s_i).$$

Here, $P(R)$ measures the average positive sentiment mass per sentence, while $C(R)$ measures the average negative sentiment mass per sentence. Because both are normalized by n, they are comparable across reviews of different lengths.

We report the praise-to-critique ratio as a smoothed log-ratio:

$$\mathrm{PCR}(R) = \log \frac{P(R) + \epsilon}{C(R) + \epsilon},$$

where $\epsilon > 0$ is a small constant that prevents division by zero and stabilizes the metric when one side is very small. Positive values of $\mathrm{PCR}(R)$ indicate relatively more praise than critique, negative values indicate relatively more critique, and values near zero indicate a more balanced tone.

When useful, we also report $(P(R), C(R))$ themselves alongside $\mathrm{PCR}(R)$ so that readers can distinguish genuinely balanced reviews from ratios driven by small denominators.

A.4 Inter-review Similarity

We measure semantic inter-review similarity across different papers using embeddings produced by `google/embeddinggemma-300m`. Higher similarity values indicate that review texts written for different papers are more semantically alike, which can reveal how generic or repetitive review content is across papers.

Let $R = \{s_i\}_{i=1}^{n}$ and $R' = \{s'_j\}_{j=1}^{n'}$ be two reviews.

Document-level similarity. When the encoder can process the full review text without truncation, we compute a single document embedding

$$E(R) \in \mathbb{R}^d$$

for each review and define cosine similarity as

$$\mathrm{Sim}_{\mathrm{doc}}(R, R') = \frac{\langle E(R), E(R') \rangle}{\|E(R)\| \cdot \|E(R')\|}.$$

This provides a holistic, length-agnostic representation of review similarity and is the most direct document-level formulation.

Sentence-level distributional similarity. To avoid information loss from collapsing a heterogeneous review into a single vector, we can instead represent each review as an empirical distribution over sentence embeddings. Let

$$e(s_i) \in \mathbb{R}^d$$

denote the embedding of sentence s_i, and let δ_x denote a Dirac measure (unit point mass) at x. We define the empirical measure associated with review R as

$$\mu_R = \frac{1}{n} \sum_{i=1}^{n} \delta_{e(s_i)}, \qquad \mu_{R'} = \frac{1}{n'} \sum_{j=1}^{n'} \delta_{e(s'_j)}.$$

We then define inter-review similarity through optimal transport:

$$\mathrm{Sim}_{\mathrm{OT}}(R, R') = 1 - W(\mu_R, \mu_{R'}),$$

where $W(\cdot, \cdot)$ denotes a Wasserstein distance between probability measures. This formulation preserves sentence-level heterogeneity while remaining invariant to review length. By contrast, mean-pooled document embeddings retain only the first moment of the sentence-embedding distribution.

In the main experiments, we measure cosine similarity between review embeddings, so higher cosine values indicate that reviews for different papers are more semantically similar.

A.5 Lexical and Syntactic Diversity

We also quantify variation in wording and structure to characterize how diverse reviews are beyond sentiment alone.

Lexical diversity. Raw type–token ratio (TTR) is highly sensitive to length, so we use a length-robust segmented formulation such as mean segmental TTR (MSTTR). Let the review contain T tokens. We partition the token sequence into

$$M = \lfloor T/L \rfloor$$

non-overlapping segments of fixed length L (e.g., $L = 50$). If $\mathrm{TTR}(b)$ denotes the type–token ratio of segment b, then lexical diversity is defined as

$$\mathrm{LexDiv}(R) = \frac{1}{M} \sum_{b=1}^{M} \mathrm{TTR}(b).$$

This reduces the strong length bias of raw TTR while still capturing vocabulary variety.

Syntactic diversity. Let

$$\phi(s_i) \in \Delta^{J-1}$$

denote a normalized distribution over syntactic events in sentence s_i, such as part-of-speech tags or

dependency relations. We aggregate these sentence-level syntactic distributions:

$$\bar{\phi}(R) = \frac{1}{n} \sum_{i=1}^{n} \phi(s_i),$$

and compute normalized entropy:

$$\mathrm{SynDiv}(R) = \frac{H(\bar{\phi}(R))}{\log J}.$$

This produces a length-normalized measure of structural variety in the review.

B Prompts Used for Review Generation

To systematically assess the effect of prompting design on review homogenization, we construct a 2×2 factorial prompt matrix along two independent axes: *reasoning strategy* (Direct vs. Chain-of-Thought) and *exemplar count* (0-shot vs. 1-shot), yielding four distinct prompt variants applied uniformly across all five models and all evaluated papers.

Across all four variants, the base persona and core directive remain identical, instructing the LLM to act as a rigorous, evidence-grounded evaluator:

> *"You are a neutral, evidence-driven conference reviewer for ICLR. Your judgments are based only on claims that are explicitly supported by the manuscript (and cited artifacts, if provided). You value validity over vibe: clear claims, sound methods, adequate evidence, and reproducibility. Novelty is neither favored nor penalized a priori—credit it only when it is well-substantiated and meaningfully advances the state of the art."*

Mitigating Score Inflation. To counteract known LLM score-inflation tendencies and sycophancy, *all four* prompt variants explicitly enforce a strict grading curve with the following critical instruction:

> *"CRITICAL: Be discriminating! Not all papers deserve high scores. Look for genuine flaws, limitations, missing comparisons, insufficient evidence, poor experimental design, overclaimed results, or unclear methodology. Many papers have significant weaknesses that warrant scores of 4–6."*

Output Schema and Year-Conditional Handling. All four prompts share an identical, strict JSON output schema to ensure parsing reliability. The requested JSON fields are: `rating` (1–10), `confidence` (1–5), `summary`, `main_review`, `strengths_and_weaknesses`, `summary_of_review`, `correctness` (1–4), `technical_novelty` (1–4), `empirical_novelty` (1–4), `recommendation`, and `confidence_label`.

ICLR year-specific formatting differences are handled via an explicit in-prompt conditional rule:

- **For 2022 papers:** Fill the `main_review` field and set `strengths_and_weaknesses` to an empty string.
- **For 2023 papers:** Set `main_review` to an empty string and fill the `strengths_and_weaknesses` field.

This mirrors the schema change introduced by the OpenReview platform between the two submission cycles. The prompts are strictly model-agnostic; model identity is tracked exclusively through experimental logging and output filename conventions.

Pre-screening. In addition to the four review-generation prompts, a separate filtering prompt comprising four sub-variants (binary classification, multi-criteria scoring, issue detection, and comparative assessment) is used for pre-screening submissions prior to full review generation. These filtering prompts are applied in a 0-shot regime and produce structured JSON outputs encoding quality flags, severity scores, and venue-relative quality tiers; their outputs are used for dataset curation rather than the stylometric analysis reported here.

B.1 Prompt 1: Direct 0-Shot (DP)

The *Direct 0-Shot* prompt serves as our baseline. It instructs the model to immediately evaluate the paper and produce the structured JSON object without any prior reasoning steps or demonstrations, relying solely on the base persona, the zero-shot instructions, and the shared grading constraints.

B.2 Prompt 2: Direct 1-Shot (1S)

The *Direct 1-Shot* prompt utilizes the same base persona and direct evaluation format as the 0-shot variant, but prepends a complete, idealized worked example to the prompt. We use the seminal paper *Attention Is All You Need* as the demonstration. The example provides a fully populated review (Rating: 8, Confidence: 4) with realistic, evidence-backed justifications. This establishes a stylistic and structural prior for the model's output.

B.3 Prompt 3: Chain-of-Thought (CoT) 0-Shot

The *Chain-of-Thought (CoT) 0-Shot* variant augments the base directive with a mandatory six-step reasoning scaffold. The model must complete this scaffold in natural language *before* producing the final JSON output. The structured scaffold is intended to reduce anchoring on superficial features and encourage domain-grounded reasoning. The exact required steps are:

1. **Claim Analysis**: What specific claims does the paper make? Are they explicitly supported?
2. **Methodological Assessment**: Are the methods clearly described and sound?
3. **Evidence Evaluation**: Is the experimental evidence adequate and convincing?
4. **Reproducibility Check**: Are sufficient details provided for reproduction?
5. **Novelty Assessment**: If novelty is claimed, is it well-substantiated and meaningful?
6. **Validity Judgment**: Overall assessment of evidence quality and methodological rigor.

B.4 Prompt 4: Chain-of-Thought (1S-CoT)

The *Chain-of-Thought (CoT) 1-Shot* variant combines the mandatory six-step reasoning scaffold with the *Attention Is All You Need* exemplar. For this prompt, the exemplar is expanded to include not only the idealized final JSON review, but also the idealized 6-step reasoning process preceding the review. The identical exemplar paper is used across all models, ensuring that cross-condition differences are attributable to the reasoning scaffold rather than the choice of demonstration.

B.5 Prompting Protocol

For each paper, model, and prompting condition, we generate a single model-produced review using a temperature $T \in \{0.2, 0.7\}$. We apply the default nucleus sampling parameters for each provider: top-$p = 1.0$ for GPT and Claude models, and top-$p = 0.95$ for Gemini models. Because OpenAI reasoning models (e.g., O3-MINI) do not expose sampling temperature controls, those reviews are generated using the model's default inference settings. Although only one review is generated per specific configuration, overall variability arises from differences across papers, models, prompting styles, and inference settings. Dispersion and diversity metrics therefore capture cross-conditioning variability rather than within-prompt stochastic variation.

C Additional Results

C.1 Average Sentiment Score

A clear divergence emerges in sentiment distributions. Human reviews exhibit more negative polarity on average than GPT- and Claude-generated reviews, while GEMINI-2.5-PRO remains the most negative model among the retained systems. Table 1 reports summary statistics of review-level sentiment polarity across human and LLM-generated reviews.

Model	n	μ	σ	Median
Human	1943	-0.088	0.152	-0.078
GPT-4O	2031	-0.020	0.048	0.000
O3-MINI	2036	-0.024	0.059	0.000
CLAUDE-3.5-SONNET-PRO	2036	-0.036	0.071	-0.012
CLAUDE-3.7-SONNET	2035	-0.046	0.078	-0.027
GEMINI-2.5-PRO	2034	-0.114	0.140	-0.064

Table 1: Summary statistics of per-review sentiment polarity (mean μ, standard deviation σ). Polarity is computed from sentence-level RoBERTa sentiment probabilities averaged within each review. All models evaluated on the same 509 common papers.

Specifically, human reviews have mean polarity $\mu = -0.088$ ($\sigma = 0.152$), compared to GPT-4O ($\mu = -0.020$, $\sigma = 0.048$) and CLAUDE-3.5-SONNET-PRO ($\mu = -0.036$, $\sigma = 0.071$). GEMINI-2.5-PRO is still notably more negative ($\mu = -0.114$, $\sigma = 0.140$). These distributional differences are illustrated in Figure 6, which shows the polarity histograms for human reviews and all five LLM systems on the dataset.

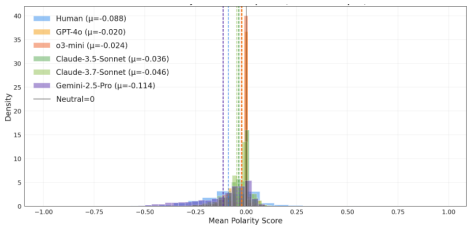

Figure 6: Sentiment polarity histograms (per-review mean polarity) for human reviews and all 5 retained LLM models (CLAUDE-3.5-SONNET-PRO +CLAUDE-3.7-SONNET, GPT-4O + O3-MINI, GEMINI-2.5-PRO). Polarity is computed from RoBERTa sentence-level sentiment probabilities and averaged within each review; values closer to 0 indicate more neutral tone, while more negative values indicate more critical tone.

Decision outcomes. Both human and model-generated reviews become more negative for rejected papers. Human reviews shift from $\mu = -0.055$ (accepted) to $\mu = -0.130$ (rejected), while GPT-4O exhibits a more modest separation (-0.014 to -0.028). GEMINI-2.5-PRO also shows a comparatively large shift, consistent with a stronger critique-oriented tendency.

Prompting effects. Prompting effects are model-dependent rather than uniform. For GPT-4O, CoT prompting yields the most neutral polarity ($\mu = -0.005$), whereas one-shot direct prompting produces more negative evaluations ($\mu = -0.044$), indicating that review sentiment varies systematically with prompt design. This suggests that the framing of the prompt can influence the tone of generated reviews, even when the underlying paper is unchanged.

2022 vs. 2023 reviews. We additionally observe a year-dependent shift in LLM sentiment (2022: $\mu = -0.057$; 2023: $\mu = -0.089$). This pattern is consistent with structural differences in review templates rather than intrinsic model behavior. The 2023 schema explicitly solicits strengths and weaknesses, encouraging more critique-oriented language. Figure 6 highlights these distributional differences, revealing pronounced clustering of GPT and Claude reviews near neutral polarity and broader, more negative dispersion in human and GEMINI-2.5-PRO reviews.

Takeaway(s). Human reviews show more polarized and varied sentiment, while LLM-generated reviews cluster closer to neutral and more uniform, indicating reduced "emotional" nuance compared to human feedback.

C.2 Sentiment Entropy (Uniformity)

Sentiment entropy reveals clear differences in tonal dispersion across review populations. Human reviews exhibit substantial within-review tonal diversity, with mean entropy $\mu_{\tilde{H}} = 0.513$ ($\sigma_{\tilde{H}} = 0.154$). Among the retained LLMs, GEMINI-2.5-PRO produces the most tonally diverse reviews ($\mu_{\tilde{H}} = 0.531$, $\sigma_{\tilde{H}} = 0.098$), whereas GPT-family models generate comparatively more uniform sentiment profiles (GPT-4O: $\mu_{\tilde{H}} = 0.442$, $\sigma_{\tilde{H}} = 0.141$; o3-mini: $\mu_{\tilde{H}} = 0.378$, $\sigma_{\tilde{H}} = 0.201$). Claude models fall in between (CLAUDE-3.5-SONNET-PRO: $\mu_{\tilde{H}} = 0.479$, $\sigma_{\tilde{H}} = 0.115$; CLAUDE-3.7-SONNET: $\mu_{\tilde{H}} = 0.489$, $\sigma_{\tilde{H}} = 0.126$). This pattern indicates model-dependent variation in tonal dispersion rather than uniform diversity collapse. These distributional differences are visualized in Figure 7, which shows the normalized sentiment entropy distributions.

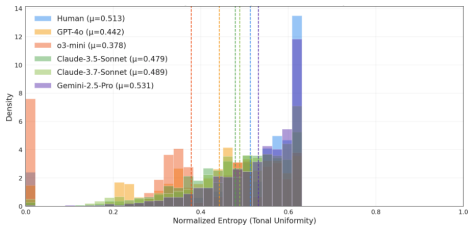

Figure 7: Distribution of normalized sentiment entropy $\tilde{H}$ (computed within each review) for human reviews and all 5 LLMs. Higher values indicate greater tonal diversity (less uniform tone) across sentences in a review.

Decision outcomes. Entropy generally increases for rejected papers. Human reviews shift from $\mu_{\tilde{H}} = 0.498$ (accepted) to $\mu_{\tilde{H}} = 0.532$ (rejected), while GPT-4o and GEMINI-2.5-PRO exhibit similar directional trends ($0.429 \rightarrow 0.459$ and $0.510 \rightarrow 0.559$, respectively). This increase suggests that reviews of rejected papers more frequently mix positive and negative sentiment signals, reflecting greater evaluative tension.

Prompting effects. We further examined *within-model* prompting effects by comparing the four prompting styles; results are in Table 2. Prompting measurably shifts review tone, but the direction is not universal across models. For GPT-4o, 1S prompting produces substantially more negative polarity ($\mu = -0.044$) compared to DP ($\mu = -0.007$) or CoT ($\mu = -0.005$). For CLAUDE-3.7-SONNET, a similar negativity shift is observed under 1S ($\mu = -0.068$ vs. DP $\mu = -0.027$). For GEMINI-2.5-PRO, polarity is already relatively negative and becomes more negative under 1S (DP $= -0.083$, 1S $= -0.161$). Overall, prompting acts as a secondary but non-negligible factor: it can shift sentiment within a given model, yet its effect size and direction depend on the model family. Table 2 reports the effects of different prompting strategies on review sentiment polarity across models. While prompting style produces small within-model shifts, overall differences across models remain larger, with Gemini generating consistently more negative reviews and GPT/Mini remaining closer to neutral.

We believe these entropy differences emerge from model-specific generation biases rather than decoding stochasticity alone: the alignment objectives and instruction tuning implicitly shape how models balance positive and negative evaluative language. Models optimized for perceived agreeability may favor sentiment-consistent phrasing, yielding lower entropy, whereas models exhibiting stronger critique-oriented tendencies more frequently mix polarity signals.One possible explanation is that these entropy differences reflect model-specific generation biases shaped by alignment and instruction tuning rather than decoding stochasticity alone (Christiano et al., 2017; Ouyang et al., 2022; Wu et al., 2024). Under this view, models optimized to produce more consistent and helpful responses may favor sentiment-consistent phrasing, whereas models with stronger critique-oriented tendencies may mix positive and negative evaluative language more often. Taken together, these findings suggest that tonal dispersion is governed not only by sampling variability but by learned generation priors that influence variance structure in model-generated evaluations.

Model	Method	Avg polarity	Median polarity
GPT-4o	DP	-0.007	0.001
GPT-4o	CoT	-0.005	0.001
GPT-4o	1S	-0.044	-0.038
GPT-4o	1S-CoT	-0.025	-0.001
o3-MINI	DP	-0.025	0.000
o3-MINI	CoT	-0.026	0.001
o3-MINI	1S	-0.020	0.000
o3-MINI	1S-CoT	-0.026	0.000
CLAUDE-3.5-SONNET-PRO	DP	-0.018	-0.001
CLAUDE-3.5-SONNET-PRO	CoT	-0.025	-0.002
CLAUDE-3.5-SONNET-PRO	1S	-0.041	-0.023
CLAUDE-3.5-SONNET-PRO	1S-CoT	-0.060	-0.037
CLAUDE-3.7-SONNET	DP	-0.027	-0.003
CLAUDE-3.7-SONNET	CoT	-0.027	-0.004
CLAUDE-3.7-SONNET	1S	-0.068	-0.054
CLAUDE-3.7-SONNET	1S-CoT	-0.060	-0.037
GEMINI-2.5-PRO	DP	-0.083	-0.037
GEMINI-2.5-PRO	CoT	-0.069	-0.029
GEMINI-2.5-PRO	1S	-0.161	-0.124
GEMINI-2.5-PRO	1S-CoT	-0.142	-0.090

Table 2: Prompting effects within the retained models. Avg/median polarity are computed from sentence-level sentiment polarity aggregated per review.

Takeaway(s). Sentiment entropy reveals systematic model-dependent differences in tonal dispersion: while human reviews and some models (e.g., Gemini) exhibit higher within-review sentiment diversity, GPT-family models tend to produce more tonally uniform evaluations, indicating that generation priors shape the balance of positive and negative critique signals.

C.3 Praise-to-Critique Ratio

Table 3 shows praise-to-critique ratios for human and LLM reviews overall and across prompting styles. All LLM models exhibit substantially higher ratios than human reviewers, indicating a systematic tendency toward praise despite moderate variation across prompting methods.

Model	Overall	DP	CoT	1S	1S-CoT
Human	1.918	–	–	–	–
O3-MINI	4.272	4.081	4.422	4.269	4.325
GPT-4O	3.616	3.953	3.956	3.097	3.638
CLAUDE-3.5-SONNET-PRO	3.150	3.237	3.588	3.040	2.754
CLAUDE-3.7-SONNET	2.790	2.940	3.167	2.506	2.602
GEMINI-2.5-PRO	2.123	2.366	2.528	1.697	1.815

Table 3: Praise-to-critique ratios for human reviews and five retained LLM models overall and by prompting style.

C.4 Inter-review Similarity

C.4.1 Experiment 1

Table 4 reports the cosine similarity between reviews of randomly paired papers. Across all models, LLM reviews show significantly higher inter-review similarity than human reviews, indicating more homogeneous review language across different papers.

C.4.2 Experiment 2

Table 5 reports topic-conditioned inter-review similarity for human and LLM reviews. Across all models, LLM reviews exhibit consistently higher similarity both within and between topics, suggesting that model-generated reviews are less sensitive to topic variation than human reviews.

C.5 Lexical and Syntactic Diversity

We evaluated lexical diversity using MTLD (Measure of Textual Lexical Diversity) (McCarthy and Jarvis, 2010), HD-D (Hypergeometric Distribution D) (McCarthy and Jarvis, 2010), and MATTR (Moving Average Type-Token Ratio) (Covington and McFall, 2010). Syntactic diversity was measured via POS entropy, dependency-relation entropy, mean dependency distance (MDD), and mean parse-tree depth—all computed with spaCy. All metrics were normalised for text-length effects using regression residuals. For each model, the four prompting styles are treated as four independent reviewers, mirroring the multiple human reviewers per paper. Table 6 contains the mean (& standard deviation) values for all six groups. All between-group differences are highly significant (Mann-Whitney $p < 0.001$; Kruskal-Wallis $p < 0.001$).

Lexical Diversity. All LLMs score significantly higher than humans on MTLD and MATTR, indicating richer vocabulary usage. The gap is largest for O3-MINI (ΔMTLD $= +43.6$, $r = 0.71$[5]) and the Claude family (ΔMTLD $\approx +33$–38, $r \approx 0.70$–0.74), and smallest for GPT-4O (ΔMTLD $= +7.6$, $r = 0.22$). HD-D, which controls for sample size, reverses the ordering for GPT-4O: human reviewers score higher (0.836 vs. 0.823, $r = -0.32$), while all other models remain above humans. This suggests that GPT-4O generates fluent prose but cycles through a narrower effective vocabulary than the other models. These distributional patterns are illustrated in Figure 8, which shows the MTLD and POS entropy distributions across human and LLM-generated reviews.

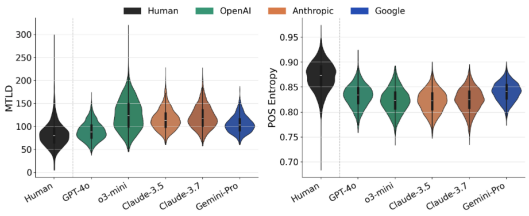

Figure 8: Violin plots of MTLD (left) and POS entropy (right) for all groups. A dashed vertical line separates human from LLM groups. The two panels capture the central trade-off: all LLMs exceed human lexical diversity (MTLD), while humans the highest grammatical variety (POS entropy). Colours indicate company: dark grey = human; green = OpenAI; orange = Anthropic; blue = Google.

Syntactic Diversity. Human reviews achieve the highest POS entropy (0.871) across all groups, indicating broader use of grammatical categories—more adverbials, pronouns, and discourse markers—than any model (range: 0.822–0.841; Figure 8). Effect sizes are uniformly large ($|r| = 0.63$–0.82; Figure 9), making POS entropy the strongest discriminator between human and LLM writing. Dependency entropy shows a similar pattern, except for O3-MINI, which is statistically indistinguishable from humans ($p = 0.13$).

Syntactic complexity differs by model provider. A striking divergence emerges in MDD and tree depth. GPT models (GPT-4O: MDD $= 4.78$; O3-MINI: 3.75) write shorter-dependency, shallower sentences than humans (6.85), indicating concise, direct prose. By contrast, CLAUDE-3.5-SONNET-PRO and GEMINI-2.5-PRO produce longer-dependency, deeper trees (MDD ≈ 9.1–9.5; depth ≈ 7.8–8.3), reflecting more elaborately nested constructions. CLAUDE-3.7-SONNET

[5] denotes the rank-biserial effect size for the Mann-Whitney U test (Kerby, 2014); larger absolute values indicate stronger separation between human and model distributions, with positive values meaning the model scores higher than humans and negative values meaning humans score higher.

Model	#Pairs	LLM Mean	Human Mean	Gap (LLM−Human)	p (t-test)
CLAUDE-3.5-SONNET-PRO	100	0.6847	0.6028	+0.0819	2.47e-15
CLAUDE-3.7-SONNET	100	0.6950	0.6028	+0.0923	5.61e-20
GEMINI-2.5-PRO	100	0.6972	0.6028	+0.0945	7.38e-20
GPT-4O	100	0.6743	0.6028	+0.0715	1.04e-11
O3-MINI	100	0.6939	0.6028	+0.0911	8.16e-19

Table 4: Experiment 1: Random paper pairs across the conference (EmbeddingGemma cosine). Each score is the cosine similarity between one selected review from paper p_1 and one selected review from paper p_2, averaged across 100 random pairs (shared across all models).

Model	Within LLM	Within Human	Within Gap	Between LLM	Between Human	Between Gap
CLAUDE-3.5-SONNET-PRO	0.6957	0.6135	+0.0822	0.6799	0.6009	+0.0790
CLAUDE-3.7-SONNET	0.7003	0.6134	+0.0869	0.6880	0.6009	+0.0871
Gemini-2.5-Flash	0.6782	0.6135	+0.0647	0.6660	0.6009	+0.0651
GEMINI-2.5-PRO	0.7059	0.6135	+0.0924	0.6908	0.6009	+0.0899
GPT-4O	0.6838	0.6134	+0.0704	0.6735	0.6009	+0.0726
O3-MINI	0.7000	0.6134	+0.0866	0.6917	0.6009	+0.0908

Table 5: Experiment 2: Topic-conditioned similarity (EmbeddingGemma cosine). "Within" averages cosine over all within-topic group cells; "Between" averages over all cross-topic cells. Positive gaps indicate higher LLM similarity than human similarity.

Metric	**Human**	**GPT-4O**	**O3-MINI**	**CLAUDE 3.5**	**CLAUDE 3.7**	**GEMINI**	**Sig.**
Lexical Diversity							
MTLD	82.1 (25.9)	89.7 (19.6)	125.6 (33.5)	115.4 (22.4)	120.1 (24.8)	105.1 (18.5)	***
HD-D	0.836 (0.035)	0.823 (0.022)	0.854 (0.024)	0.862 (0.018)	0.864 (0.019)	0.850 (0.015)	***
MATTR	0.804 (0.042)	0.816 (0.027)	0.848 (0.029)	0.843 (0.021)	0.849 (0.022)	0.826 (0.018)	***
Syntactic Diversity							
POS Entropy	0.871 (0.029)	0.833 (0.023)	0.823 (0.025)	0.822 (0.023)	0.824 (0.024)	0.841 (0.019)	***
Dep. Entropy	0.864 (0.029)	0.852 (0.020)	0.861 (0.021)	0.832 (0.018)	0.840 (0.018)	0.834 (0.017)	***
Mean Dep. Dist.	6.85 (4.76)	4.78 (1.09)	3.75 (0.95)	9.51 (4.03)	7.04 (1.61)	9.07 (2.06)	***
Tree Depth	5.83 (3.27)	3.96 (1.21)	3.56 (0.80)	7.79 (2.49)	6.25 (1.90)	8.28 (2.79)	***
N reviews	1937	2031	2036	2036	2035	2034	

Table 6: Lexical and syntactic diversity metrics across review sources on 509 papers where all models have complete data after excluding `Gemini-2.5-Flash`. Values show mean (SD). **Sig.**: Mann-Whitney U test significance level; *** $p < 0.001$ for all comparisons except Dep. Entropy for O3-MINI vs. Human ($p = 0.13$, ns). GPT-4O and O3-MINI produce *syntactically simpler* sentences than humans; Claude and GEMINI-2.5-PRO produce *more complex* ones, while *all* LLMs exhibit higher lexical diversity.

(MDD $= 7.04$; depth $= 6.25$) is closest to human syntactic complexity. All differences remain significant after length normalisation, confirming they reflect genuine structural choices rather than length artifacts.

Takeaway(s). The results reveal a consistent *lexical–syntactic trade-off*: LLM reviews exhibit higher lexical diversity than human reviews, meaning they use a wider range of vocabulary, but lower syntactic diversity, meaning their sentence structures rely on a narrower set of grammatical patterns. Within LLM, there is a large stylistic difference demonstrated by differences in syntactic metrics. Figure 9 summarizes these differences using effect-size estimates (rank-biserial r) across all seven diversity metrics.

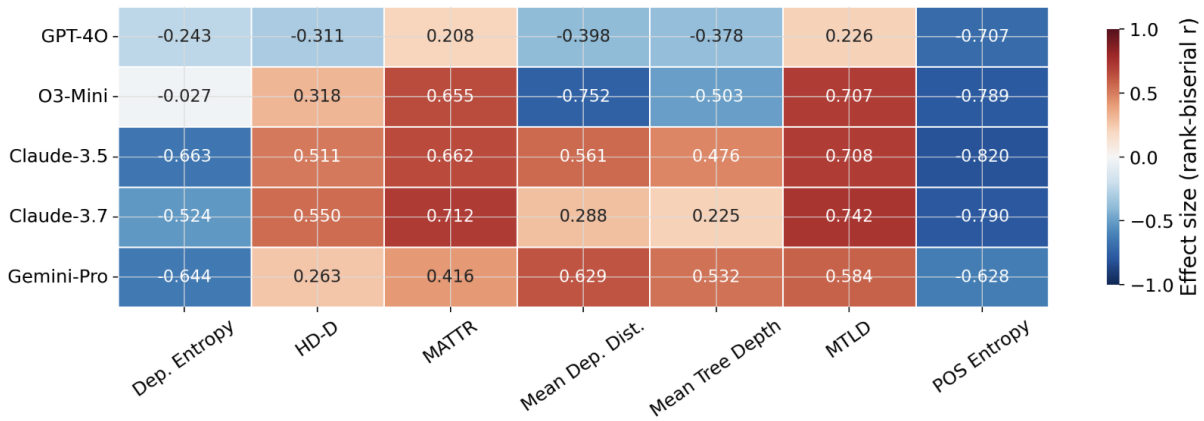

Figure 9: Effect-size heatmap (rank-biserial r) for Mann-Whitney U tests comparing each LLM model against human reviews across all seven metrics. Red cells indicate the model scores higher than humans ($r < 0$); blue cells indicate humans score higher ($r > 0$). The POS entropy column (uniformly red) shows the strongest and most consistent human advantage. The Mean Dep. Dist. column reveals the OpenAI–Anthropic/Google split in syntactic complexity.

D Analysis of Human Quality Assessments of AI Reviews

To evaluate the relevance and quality of LLM generated reviews at a high level, we use a quality assessment with human annotators. We randomly sample reviews from five papers, giving 25 AI-generated reviews across models. Each review was annotated by two different annotators. Seven domain experts evaluated whether reviews were substantively relevant to the paper and assessed overall review quality. These annotators were provided with access to the original paper, abstract, and an LLM review. They were instructed to rate both relevance and quality on a 1–5 scale with the option to provide free-form comments. Detailed annotation criteria and procedures are described below, along with results of analysis of their scores and comments.

D.1 Relevance and Quality Scores By Model

We report average relevance and quality as assigned by human annotators for each model, as well as the standard deviation across scores. Table 7 provides the raw average scores and standard deviations, and Figure 10 provides a bar chart for score comparison across models.

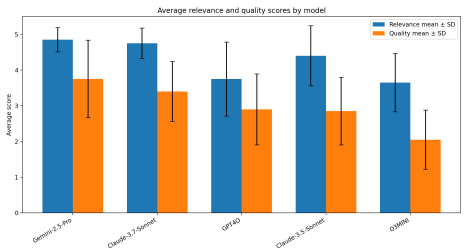

Figure 10: Bar chart of average relevance (blue) and quality (orange) scores by model across all reviewed papers. Error bars denote variance (standard deviation).

D.2 Theme Development and Categorization

Common themes in human quality evaluation of LLM generated reviews were identified by manual review of all 50 evaluations. Quality evaluations varied in length, but each primarily concerned at least one of nine themes: summarizing instead of reviewing, missing limitations or weaknesses, endorsing instead of reviewing, lack of mentioning related work, being too vague or generic, containing factual errors, not discussing novelty, being sufficiently detailed and relevant (positive), and/or providing a factually accurate understanding of the paper (positive).

Once themes were identified, common phrases in the human quality evaluations were manually categorized into each theme. Any human annotation containing at least one phrase related to a specific theme was categorized under that theme in the ultimate breakdown of theme matches. Evaluations were categorized under all matched themes if they contained phrases that matched to multiple themes. Table 8 provides a breakdown of all included themes and their respective matched phrases.

D.3 Theme Distribution Across All Evaluations

The top three identified themes were: "too vague or generic", at 20 total annotator comments, "missing limitations/weaknesses" at 12 evaluations, and "detailed and relevant" at 9 evaluations. Figure 11 provides counts for all 9 themes.

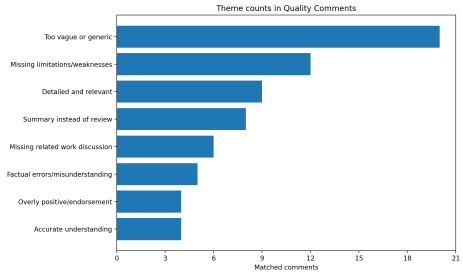

Figure 11: Bar chart showing the number of matched comments for each theme across all models and papers (50 total quality evaluations).

Model	Relevance (mean ± SD)	Quality (mean ± SD)
GEMINI-2.5-PRO	4.85 ± 0.34	3.75 ± 1.09
CLAUDE-3.7-SONNET	4.75 ± 0.42	3.40 ± 0.84
CLAUDE-3.5-SONNET-PRO	4.40 ± 0.84	2.85 ± 0.94
GPT-4O	3.75 ± 1.03	2.90 ± 0.99
O3-MINI	3.65 ± 0.82	2.05 ± 0.83

Table 7: Average relevance and quality scores assigned by human annotators to LLM-generated paper reviews. Scores are reported as mean ± standard deviation across 10 scored reviews per model (5 papers, 7 annotators).

Theme	Matched Phrases
Summary instead of review	"summary rather than a review"; "more like a summary"; "like an abstract instead of a review"; "rewrite of the original abstract"; "summarization instead of a review"; "only summarizes what the paper did"; "more descriptive of the paper content rather than subjective comment"
Missing limitations or weaknesses	"missing limitations"; "lack of limitations"; "no concerns or questions"; "does not provide concerns or suggestions"; "does not point out shortcomings"; "unjustified aspects not addressed"; "limited discussion of the paper's shortcomings or limitations"
Overly positive or endorsement	"too positive"; "more like an endorsement"; "mostly praise the paper"; "exaggerated praise"; "friendly tone that avoids criticism"
Missing related work discussion	"need related-work engagement"; "lack of novelty"; "does not discuss relation to related work"; "limited connection to existing work"; "does not explain how this work fits into the literature"; "does not mention relevant citations"
Too vague or generic	"too short or brief"; "too generic"; "too high-level"; "vague"; "too general to be useful"; "lacks detail"; "minimal content"; "could go into further depth"; "does not express any opinion"
Factual errors or misunderstanding	"factual error"; "hallucination"; "clear mistake"; "fundamental misunderstanding"; "reviewing an incomplete excerpt"; "experimental section absent"; "not true of the actual paper"; "ungrounded claims"
Novelty not discussed	"does not discuss novelty"; "does not discuss originality"; "limited discussion of novelty or originality"; "evaluates methodology more than novelty"; "no evaluation of motivation or novelty"
Detailed and relevant (positive)	"detailed and relevant"; "very relevant"; "relevant to the paper"; "covers the entirety of the paper"; "all aspects are covered"; "technically sound"; "thorough review"; "well explained"
Accurate understanding (positive)	"accurate"; "factually correct"; "grounded in the paper"; "demonstrates understanding of the paper"; "identifies real strengths and weaknesses"; "raises reasonable technical concerns"

Table 8: Theme classification rules used to categorize reviewer comments. Each theme is identified by matching comment text against a set of trigger phrases or expressions.

E Review Length and Structural Monoculture

To quantify structural monoculture, we examine the absolute review length and its variance. Let $T(R)$ denote the total word count for a review R. We quantify structural homogenization using the coefficient of variation (CV) for word length:

$$\mathrm{CV}(T) = \frac{\sigma(T)}{\mu(T)},$$

where a lower CV indicates a collapse in the natural, heavy-tailed distribution of reviewer effort.

Results. Table 9 summarizes the review length statistics pooled across all four prompting styles and the full 509 paper corpus, detailing the mean length (μ), standard deviation (σ), and coefficient of variation (CV) for each source. Historically, human-authored reviews exhibit high length variance, reflecting a diverse mixture of concise acceptances, brief dismissals, and exhaustive critiques. In our dataset, human reviews have a mean word count of $\mu_{\text{Human}} = 306.6$ with a coefficient of variation of $\mathrm{CV}_{\text{Human}} = 0.732$, spanning a range from single-sentence responses (1 word) to highly detailed assessments exceeding 2,600 words.

Source	Mean (μ)	Std (σ)	Median	CV
Human	306.6	224.4	255	0.732
GPT-4O	208.1	52.1	214	0.250
O3-MINI	148.6	34.8	146	0.234
CLAUDE-3.7-SONNET	323.1	62.7	319	0.194
CLAUDE-3.5-SONNET-PRO	473.7	134.6	444	0.284
GEMINI-2.5-PRO	466.9	128.1	432	0.274

Table 9: Summary statistics of review word count pooled across all four prompting styles and 497–498 papers. AI models exhibit a systematically lower coefficient of variation (CV) than human reviewers ($\mathrm{CV}_H = 0.732$), indicating structural monoculture in length distributions.

In contrast, AI-generated reviews converge to a default algorithmic verbosity. GPT-4O produced reviews with a mean of 208.1 words (CV $= 0.250$); O3-MINI was the most terse at 148.6 words (CV $= 0.234$). The larger reasoning- and context-rich models are substantially more verbose—CLAUDE-3.5-SONNET-PRO averages 473.7 words and GEMINI-2.5-PRO averages 466.9 words—yet still exhibit markedly compressed variance relative to human reviewers (CV ≈ 0.27–0.28, compared to $\mathrm{CV}_{\text{Human}} = 0.732$). Across all five LLMs, CV values fall in the range $[0.194, 0.284]$, a 2.6–$3.8\times$ reduction relative to the human baseline, indicating a systematic collapse of the natural heavy-tailed length distribution.

Prompting effects. Per-style breakdowns (Table 10) reveal that GPT-4O exhibits the largest prompt-sensitivity in length: its mean rises from 182 words (CoT) to 239 words (1S-CoT), yet its within-style CV remains the lowest of all models (0.13–0.28). GEMINI-2.5-PRO shows the opposite pattern: its mean word count nearly halves between 0-shot (≈ 550 words) and 1-shot conditions (≈ 380 words), indicating that example-based prompting substantially suppresses its verbosity.

Visualizations. These structural differences are further illustrated in our visual analyses. Figure 12 provides a grouped bar comparison of absolute re-

Source	DP	1S	CoT	1S-CoT
Human	307 ± 225	306 ± 224	307 ± 225	307 ± 224
GPT-4o	184 ± 50	227 ± 30	182 ± 51	239 ± 48
o3-mini	140 ± 26	146 ± 30	150 ± 42	159 ± 36
Claude-3.7-Sonnet	352 ± 60	332 ± 53	301 ± 62	307 ± 62
Claude-3.5-Sonnet-Pro	485 ± 136	490 ± 113	513 ± 155	407 ± 104
Gemini-2.5-Pro	548 ± 107	388 ± 64	557 ± 133	375 ± 52

Table 10: Mean $\pm$ standard deviation of word count per prompting style. Human variance is stable across all four conditions, whereas LLM lengths are sensitive to prompting strategy—most notably GPT-4o and Gemini-2.5-Pro.

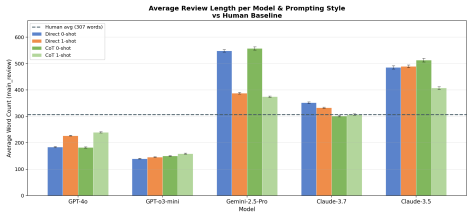

Figure 12: Grouped bar chart showing average review word count per model and prompting style, with the human baseline shown as a dashed line. Error bars denote standard error of the mean.

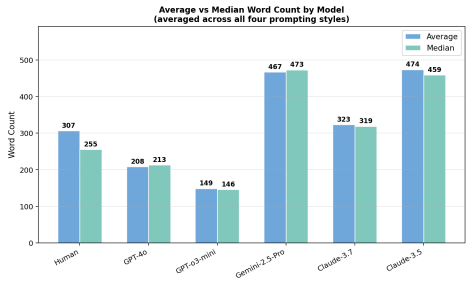

Figure 13: Average vs. median word count per source, averaged across all four prompting styles. Human reviews are right-skewed (mean $>$ median by $\sim$50 words), while most LLM outputs are approximately symmetric, indicating a collapse of the heavy-tailed effort distribution characteristic of human peer review.

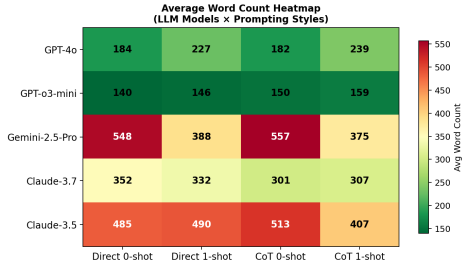

Figure 14: Heatmap of average word count across LLM models and prompting styles. Gemini-2.5-Pro and Claude-3.5-Sonnet-Pro are the most verbose models overall, while both exhibit strong prompt-sensitivity: Gemini-2.5-Pro drops from $\sim$550 words (0-shot) to $\sim$380 words (1-shot), and Claude-3.5-Sonnet-Pro similarly contracts under 1S-CoT. o3-mini remains consistently terse across all conditions (140–159 words).

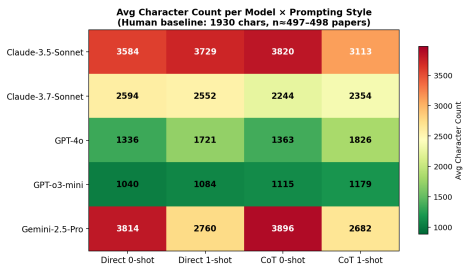

Figure 15: Heatmap of average character count per model and prompting style (human baseline: 1,930 characters). Colour intensity encodes verbosity; darker red indicates longer reviews. Gemini-2.5-Pro and Claude-3.7-Sonnet are the most prompt-sensitive in absolute length, while o3-mini remains consistently terse across all conditions.

view lengths across models and prompting styles.

Figure 13 summarizes the overall mean-versus-median distributional shifts.

Figure 14 details how mean character counts vary conditionally across all model–prompting-style combinations.

Figure 15 confirms identical verbosity and prompt-sensitivity patterns at the absolute character level.

F AI Reviewers

To further contextualise the homogenisation observed in base LLMs, we conduct a detailed comparative analysis between human reviewers and two autonomous reviewing agents evaluated across all 20 papers (a sample size constrained by the strict upload quotas and usage limits of the respective platforms): the Stanford Agentic Reviewer (SAR) (PaperReview.ai, 2024) and the Bibby Agentic Reviewer (BAR) (Bibby, 2024). In cross-paper comparisons using the same design as Experiment 1 (100 randomly sampled pairs, `EmbeddingGemma`), SAR shows no significant difference from the human baseline ($p = 0.27$), while BAR is marginally higher ($p = 0.012$). Both effect sizes ($+0.02$–$+0.05$) are 2–4$\times$ smaller than those of any base LLM in Experiment 1. Agentic reviewers therefore show at most a weak cross-paper homogeneity signal—far below the systematic monoculture of base LLMs.

F.1 Stanford Agentic Reviewer (SAR)

Review Length. SAR produces reviews of consistent, moderate length (mean 2,022 characters, $\sigma = 237$), approximately 16% longer than the human mean (1,742 characters) but with substantially lower variance. Human reviews are highly heterogeneous ($\sigma = 1{,}221$), spanning terse 58-character comments to exhaustive 7,178-character assessments, with a right-skewed distribution in which the median (1,416) falls well below the mean. SAR's near-zero skew (mean $\approx$ median $\approx$ 2,000 characters) is consistent with a templated, length-controlled generation process.

Sentiment and Tonal Balance. SAR reviews exhibit mild negative polarity (mean -0.049 across all 20 papers), while human reviewers are more negative on average (mean -0.094) with substantially higher inter-reviewer variance. Aggregate sentiment entropy is similar across sources (SAR 0.530, human 0.513), suggesting broadly comparable tonal balance at the corpus level. Both sources follow the same accept/reject gradient: SAR polarity is -0.016 for accepted papers and -0.111 for rejected ones, mirroring the human pattern (-0.080 vs. -0.119), confirming that SAR's sentiment is sensitive to paper quality rather than fixed by template (Figures 4 and 16).

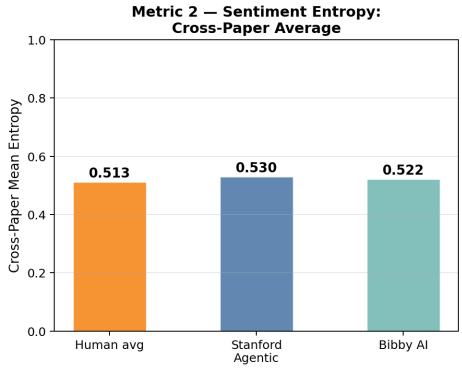

Figure 16: Average sentiment entropy comparison, showing tonal variability within reviews.

Praise-to-Critic Ratio. SAR's mean praise-to-critic ratio across 20 papers is 2.79 (human: 3.22). Stratified by decision, both SAR and humans assign relatively higher praise to accepted papers (SAR: 2.99 accept vs. 2.44 reject; human: 3.34 vs. 2.99)—a directionally consistent pattern across all sources. However, the differences are not statistically significant ($p = 0.67$ for the human accept/reject comparison), reflecting high paper-level variance ($\sigma \approx 1.4$–1.7 across all sources). We note that the praise-to-critic values reported here ($\sim$3.2 for humans) differ from the main-body figure (Table 3: 1.9). This discrepancy is attributable to the small sample size ($N = 20$ vs. $N \approx 500$) rather than a methodological error: even grouping our 20 papers by accept/reject status does not resolve the gap, and the wide confidence interval on the 20-paper estimate (95% CI: $[2.5, 4.0]$) easily encompasses the main-body value. Given this high variance and the small sample, the praise-to-critic metric should be interpreted directionally rather than as a precise point estimate (Figure 17).

Inter-Review Similarity: Within-Paper Alignment. SAR achieves a mean cosine similarity of 0.661 with human reviews for the *same* paper (range: 0.49–0.79), near the human-vs.-human within-paper mean of 0.683. This confirms that SAR captures paper-specific technical content com-

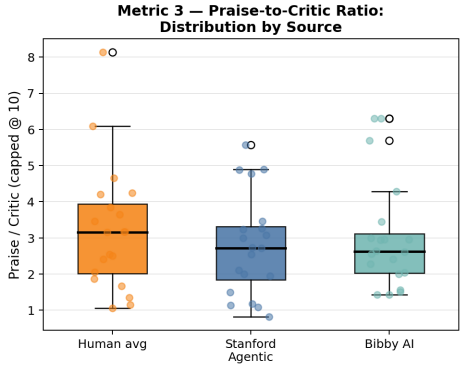

Figure 17: Distribution of Praise-to-Critic ratios for autonomous agents compared to human reviewers. All three sources show higher praise-to-critic for accepted papers, but paper-level variance is large ($\sigma \approx 1.5$) and group differences are not significant ($p = 0.67$, $N = 20$).

parably to human reviewers (Figure 18).

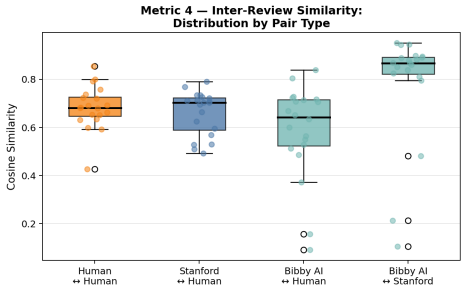

Figure 18: Distribution of inter-review semantic similarity scores (within-paper).

Inter-Review Similarity: Cross-Paper Homogeneity. To test for monoculture directly—using the same design as Experiment 1—we randomly sample 100 pairs from the $\binom{20}{2} = 190$ unique cross-paper pairs, select one human reviewer per paper at random, and compute cosine similarity using `EmbeddingGemma` (Table 11, Figure 20). SAR's cross-paper self-similarity (0.317) is not significantly different from the human baseline (0.298; $p = 0.27$). BAR's (0.343) is marginally higher ($p = 0.012$). Both effect sizes (+0.019 and +0.045) are substantially smaller than those observed for any base LLM in Experiment 1 (+0.07–+0.09, all $p < 10^{-10}$).

Lexical and Syntactic Diversity. SAR exhibits markedly higher MTLD (mean $\approx$ 183) than human reviewers (mean $\approx$ 79), a 2.3$\times$ gap consistent across all 20 papers. Conversely, human reviewers display higher POS and dependency entropy, reflecting greater syntactic diversity despite a narrower vocabulary (Figure 19). **Takeaway.** SAR

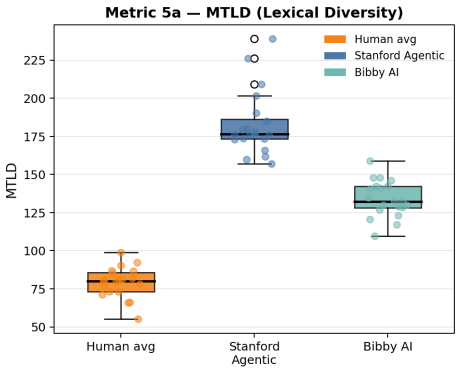

Figure 19: Average Lexical Diversity (MTLD) demonstrating the significantly broader vocabulary of agentic reviewers.

captures paper-specific technical content and exhibits high lexical diversity, but relies on a structurally rigid, syntactically formulaic generation process. Its cross-paper similarity is statistically indistinguishable from the human baseline, placing it firmly outside the monoculture pattern observed in base LLMs.

F.2 Bibby Agentic Reviewer (BAR)

BAR produces substantially longer reviews than both SAR and human reviewers (mean $\approx$14,000 characters per paper), driven by a rigid multi-section template. Length here reflects template design rather than analytical depth.

Sentiment and Tonal Balance. BAR is the most neutral of the three source types (mean polarity -0.064), compared to SAR (-0.049) and the human panel (-0.094). Like SAR, BAR's sentiment tracks paper outcome (-0.052 accepted vs. -0.088 rejected), mirroring the human pattern and confirming that neither agent applies a fixed tonal register regardless of quality (Figures 4 and 16).

Praise-to-Critic Ratio. BAR's mean praise-to-critic ratio (3.02) closely tracks the human mean (3.22), with the same directional accept/reject pattern (BAR: 3.44 accepted vs. 2.24 rejected). As with SAR, differences are not statistically significant given the small sample and high paper-level variance.

Lexical Diversity. BAR's MTLD (mean $\approx$ 134) falls between the human baseline ($\approx$ 79) and SAR ($\approx$ 183). Despite generating far longer text than

SAR, BAR achieves lower MTLD, indicating that additional length is driven by structural boilerplate rather than genuine lexical variety (Figure 19).

AI–Human Similarity: Within-Paper Boundary. Within each paper, BAR and SAR are more semantically similar to each other (mean 0.784) than either is to the human panel (BAR 0.587; SAR 0.661), establishing a clear AI-versus-human cluster boundary across all 20 papers.

AI–Human Similarity: Cross-Paper Design. In the Experiment-1-style cross-paper analysis (Table 11, Figure 20), SAR shows no significant elevation above the human baseline ($p = 0.27$); BAR is marginally higher ($p = 0.012$), with an effect size ($+0.045$) roughly half that of the weakest base LLM in Experiment 1. Stratifying all $\binom{20}{2} = 190$ unique pairs by decision (a separate exhaustive analysis) reveals a notable asymmetry: BAR's within-reject cross-paper similarity (0.404, 21 pairs) is *higher* than within-accept (0.362, 78 pairs)—opposite to SAR (0.272 vs. 0.346). This suggests that BAR's rigid template yields more convergent output when reviewing clearly weak submissions.

Source	Mean	Std	SE	Gap vs. Human
SAR	0.317	0.121	0.012	+0.019
BAR	0.343	0.133	0.013	+0.045
Human	0.298	0.116	0.012	(baseline)

Table 11: Cross-paper pairwise cosine similarity for SAR, BAR, and human reviewers. Design mirrors Experiment 1: 100 randomly sampled pairs (N=20 papers), one human reviewer sampled per paper, `EmbeddingGemma` embeddings. SAR vs. human: $p = 0.27$ (n.s.); BAR vs. human: $p = 0.012$.

Takeaway. BAR's rigid template inflates review length without adding lexical variety and converges with SAR into a shared AI semantic space within each paper. In cross-paper comparisons, BAR's elevation above the human baseline is statistically marginal and an order of magnitude smaller than base LLM monoculture; SAR shows none at all. The dominant monoculture signal for both agents is therefore inter-agent stylistic convergence *within* each paper, not cross-paper repetition. BAR's higher cross-paper similarity on rejected papers further reveals a template-induced failure: structured format suppresses paper-specific argumentation exactly where differentiated critique matters most.

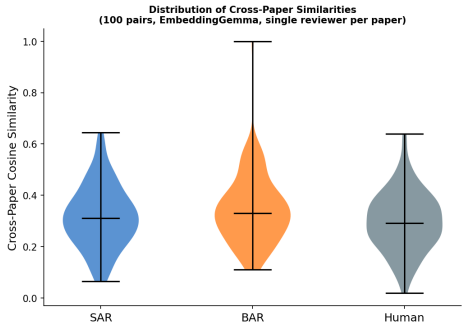

Figure 20: **Cross-paper similarity distribution** for SAR, BAR, and human reviews under the Experiment 1 design (single reviewer per paper, `EmbeddingGemma`). Each point represents cosine similarity between embeddings of reviews written for different papers. Both agent-based reviewers produce similarity distributions comparable to humans, with BAR showing slightly higher similarity, but still far below the levels observed for base LLM reviewers in Experiment 1.