From Replication to Redesign: Exploring Pairwise Comparisons for LLM-Based Peer Review

Yaohui Zhang[1] **Haijing Zhang**[1] **Wenlong Ji**[1]
Tianyu Hua[1] **Nick Haber**[1] **Hancheng Cao**[2] **Weixin Liang**[1]
[1] Stanford University [2]Emory University

Abstract

The advent of large language models (LLMs) offers unprecedented opportunities to reimagine peer review beyond the constraints of traditional workflows. Despite these opportunities, prior efforts have largely focused on replicating traditional review workflows with LLMs serving as direct substitutes for human reviewers, while limited attention has been given to exploring new paradigms that fundamentally rethink how LLMs can participate in the academic review process. In this paper, we introduce and explore a novel mechanism that employs LLM agents to perform pairwise comparisons among manuscripts instead of individual scoring. By aggregating outcomes from substantial pairwise evaluations, this approach enables a more accurate and robust measure of relative manuscript quality. Our experiments demonstrate that this comparative approach significantly outperforms traditional rating-based methods in identifying high-impact papers. However, our analysis also reveals emergent biases in the selection process, notably a reduced novelty in research topics and an increased institutional imbalance. These findings highlight both the transformative potential of rethinking peer review with LLMs and critical challenges that future systems must address to ensure equity and diversity.

1 Introduction

The peer review system serves as a cornerstone for validating and disseminating scientific ideas [2]. Yet, despite its widespread adoption, the peer review system has become increasingly strained because of the rapid growth of research output and a persistent shortage of qualified reviewers [13, 30, 40].

Recent advances in large language models (LLMs) [1, 16, 37] have sparked growing interest in using AI to assist or even automate parts of the review process [17, 22, 39]. Unlike human reviewers, whose availability and productivity are inherently constrained, LLM-based agents offer the potential for greater scalability and consistency in academic evaluation.

However, existing LLM-based paper review efforts [18, 27, 49, 55, 61] largely mirror traditional workflows: For each paper, multiple LLM agents first independently generate structured reviews (e.g., summary, strengths, weaknesses, suggestions for improvement, numeric ratings). Subsequently, an LLM meta-reviewer synthesizes these individual LLM reviews and composes a cohesive meta-review summarizing the collective assessment. Finally, a decision (e.g., accept or reject) is generated either directly by this LLM meta-reviewer agent or based on its synthesized ratings.

While LLMs offer new possibilities for peer review, simply replicating existing workflows may overlook a deeper opportunity. Historical analysis suggests that the conventional structure of peer review was not carefully engineered for optimal evaluation, but rather emerged as a pragmatic response to resource constraints. Studying the development of peer review in sociology, Merriman [32] observes that "the evolution of peer review is best understood as the product of continuous efforts to steward editors' scarce attention while preserving an open submission policy that favors authors' interests."

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-0394

This perspective highlights that many features of today's review systems reflect historical compromises rather than principled design choices. As scholarly publishing expands globally and across disciplines [10], these inherited structures have come under increasing scrutiny. Issues like bias [20, 44, 46] and noisy ratings [28] have spurred debates on how to improve or reform peer review [14, 45, 48, 52].

Given their exceptional scalability, LLMs present an opportunity not merely to automate existing practices, but to fundamentally rethink the design of manuscript evaluation systems. Rather than replicating workflows optimized for human reviewers, it is crucial to rethink the foundations of effective evaluation and to design AI systems that complement and address the persistent limitations of traditional peer review.

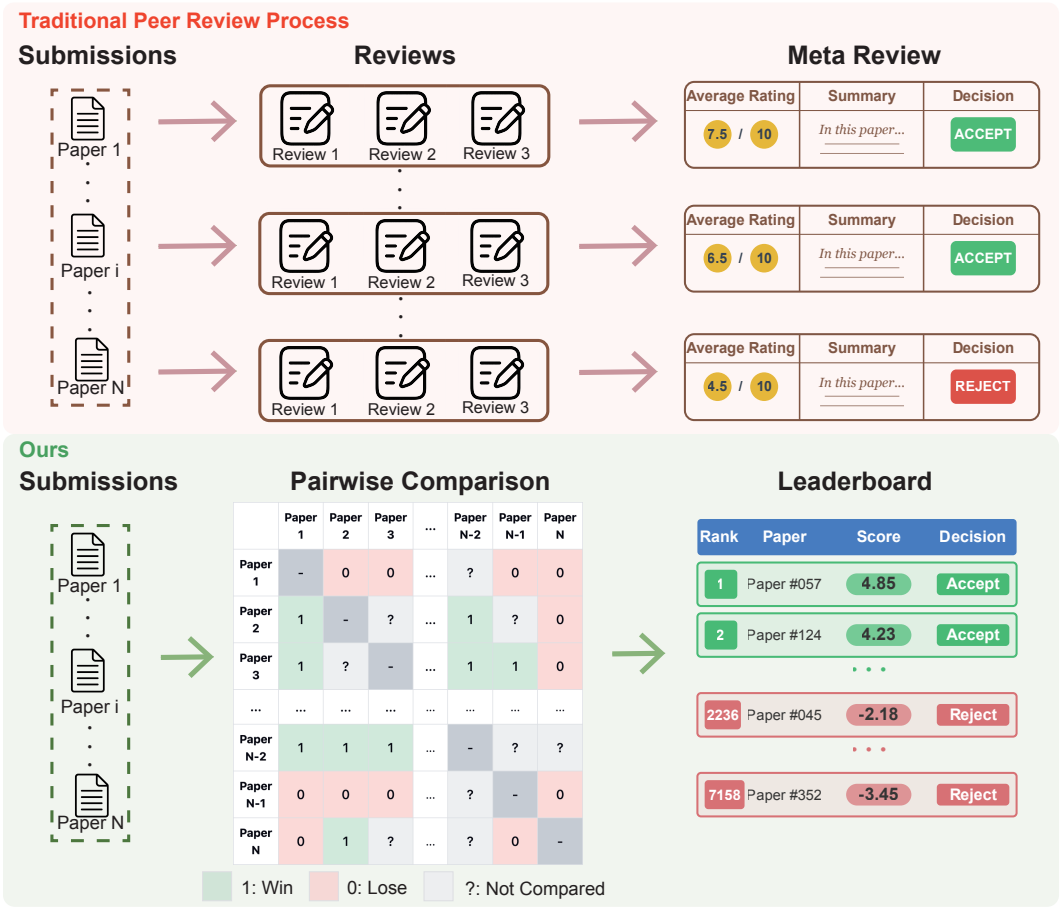

Figure 1: **Comparison of the conventional peer review process (top) and the proposed pairwise evaluation ranking system.** The traditional paradigm (followed by both human reviewers and existing LLM-based review systems) works on paper-level assessment, where each manuscript is first evaluated independently by multiple reviewers who provide detailed comments and scores. These evaluations are integrated by a meta-reviewer to recommend a final decision. Our method adopts pairwise comparisons between papers: each agent is randomly assigned a pair of submissions to evaluate. These comparative judgments are then aggregated using the Bradley–Terry model, where the resulting BT coefficients serve as scoring functions to produce a global ranking of all submissions. Final decisions are derived from the aggregated rankings, enabling a more objective assessment of the relative quality across the entire submission pool.

In this work, we propose a shift from replication to redesign. We focus on the **decision-making layer** of peer review—specifically, mechanisms that can rank submissions and support acceptance decisions. Our approach is intended to complement, not replace, the current peer review process. Specifically, we introduced and explored a novel mechanism that engages LLM agents in pairwise comparisons for more effective manuscript assessment. Rather than assigning absolute scores to

individual papers, our approach directly contrasts pairs of submissions to determine which one better meets the multidimensional evaluation criteria. Each LLM agent is tasked with evaluating two manuscripts at a time, and by integrating the results from numerous pairwise comparisons, we can construct a robust overall assessment that more accurately reflects relative quality.

Our results indicate that with proper scaling, this pairwise mechanism demonstrates potential to identify papers of higher academic impact (as measured by future citation count), significantly outperforming paper-level rating approach. However, our analysis also reveals important limitations: Papers preferred by our LLM-based pairwise ranking system tend to exhibit lower topic novelty and greater concentration among a limited number of high-prestige institutions. While these biases highlight critical challenges for the deployment of LLM-based review systems, we view this work as a first step toward exploring fundamentally new paradigms for scalable, equitable, and robust research evaluation.

It is important to note that while we use proxy metrics—such as future citation counts—to evaluate our proposed mechanism, this represents an exploratory and pragmatic step rather than an ideal. The gold standard would be direct human evaluation of how well a mechanism supports the selection and improvement of high-quality research, but such evaluations are costly and difficult to scale [43]. Proxy metrics offer a tractable, though imperfect, way to approximate long-term impact and anticipate how different mechanisms might perform. Ideally, a strong alternative review mechanism should correlate with—but not fully replicate—the outcomes of the current system, offering complementary perspectives and surfacing different strengths. As we show, our proposed pairwise comparison mechanism achieves this balance, aligning with current human judgments while introducing useful differentiation. Crucially, our goal is not to mimic existing outcomes, but to design mechanisms that add value.

Moreover, we believe the strength of human-led peer review lies not only in final decisions, but also in the human interactions it enables—feedback, discussion, and iterative refinement—that help authors strengthen their work. Through these interactions, reviewers and authors also engage in a shared evaluative practice that reinforces norms, builds trust, and helps individuals identify with the scholarly community [63, 19]. Our intention is not to automate or replace this process, but to broaden how we understand and support scholarly evaluation.

2 Related Work

Challenges in Peer Review Peer review, while essential to scientific progress, faces numerous challenges in today's academic landscape. Reviewers often struggle with overwhelming workloads, further compounded by the absence of meaningful incentives and the inequitable distribution of review requests [7, 31]. In addition, implicit biases can influence evaluation [46, 47, 50], with studies showing that factors like author affiliation or gender may affect acceptance rates [59]. As research output continues to accelerate globally [5], these systemic challenges demand innovative solutions to preserve the integrity and effectiveness of scientific evaluation. Earlier works [48, 57, 56] introduce pairwise comparisons by incorporating author-provided rankings to calibrate noisy reviewer scores. In contrast, we fully replace paper-level scores with large-scale pairwise comparisons performed by LLM agents, without relying on human ratings or author input.

LLMs for Peer Review Recent advancements in Large Language Models (LLMs) have sparked considerable interest in their potential to transform academic peer review processes [21, 62, 24, 23]. While existing studies [9, 22, 38] have demonstrated that LLMs can provide valuable feedback for research papers, exhibiting substantial overlap with reviews written by human reviewers, they often struggle with assigning reasonable ratings and making sound decisions, which raises concerns about their reliability for evaluative tasks [26, 58]. For instance, Niu et al. [35] showed that ChatGPT tends to grant acceptances: out of 1558 full paper evaluations, ChatGPT suggested 1243 acceptances compared to only 315 rejections, resulting in an acceptance rate of approximately 79.8%. Similarly, Zhou et al. [60] highlighted that LLMs could generate reasonable aspect scores (e.g. recommendation, soundness, originality) from human reviews but failed when given only the research paper. Lu et al. [27] improved the base LLM's decision-making process through prompting techniques such as self-reflection [42], few-shot examples [54] and response ensembling [53]. CycleReviewer [55] and DeepReview [61] enhanced LLM-based paper review via fine-tuning on high-quality review dataset.

 https://doi.org/10.52202/085713-0394

In contrast, Less effort is made to explore new paradigms that fundamentally rethink how LLMs can participate in the academic review process.

3 Method

3.1 Framework Overview

We now describe our peer review mechanism (Figure 1) that transforms the traditional paper-level assessment process into a comparative ranking system through pairwise judgments. This approach builds on a key insight that evaluators are generally more reliable when deciding which of two items is better than when assigning each an absolute score [8]. By integrating a large number of pairwise judgments into a global ranking with the Bradley-Terry model, we could potentially mitigate the calibration inconsistencies and personal rating biases that plague conventional peer review systems.

The mechanism operates in three key stages: First, we collect pairwise paper comparisons from multiple LLM agents, where each agent evaluates a randomly sampled pair of papers and produces a binary preference judgment. Second, we apply the Bradley-Terry model to quantify each paper's relative quality through BT coefficients. Finally, we determine the optimal coefficients through maximum likelihood estimation and derive a complete ranking of all papers.

3.2 From Pairwise Comparisons to Rankings

Suppose we have N paper candidates indexed by $i \in [N]$ and M LLM agents participating in the evaluation process. We aim to establish a quality-based ranking through pairwise comparisons. The framework consists of the following steps:

1. **Pairwise Comparison**: Let $\mathcal{S} = \{(i, j) : i \neq j \text{ and } i, j \in [N]\}$ be all possible pairs of papers to compare. For each of M agents, we randomly assign one pair $(i, j) \in \mathcal{S}$ to evaluate. Each agent then analyzes both papers and produces a binary preference judgment $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ indicates that the agent prefers paper i over paper j, and $y_{ij} = 0$ indicates the opposite preference. The full prompt used for pairwise judgment is provided in Supp Figure 7. Denote the collection of all chosen pairs as $\mathcal{A}$, where $|\mathcal{A}| = M$. We collect all comparison results into an observation set $\mathcal{O} = \{(i, j, y_{ij}) : (i, j) \in \mathcal{A}\}$, which contains triplets of the compared papers and their corresponding preference judgments from all agents. Since the size of $|\mathcal{S}|$ grows quadratically with n, we can only afford to assess a small fraction of all possible pairs, and hence it is challenging to recover the ranking under this scenario. Later in Section 4.2, we empirically examined how ranking quality scales with the number of LLM agents. The results reveal a clear scaling law, and we are able to recover a high-quality ranking with less than 2% samples of all possible pairs. [1]
2. **Bradley-Terry Model**: We employ the Bradley-Terry model [6] to recover the paper ranking from pairwise comparison results. The model provides stable statistical estimation from sparse comparisons, which is particularly suitable for our scenario. Specifically, it assumes each paper i has an underlying quality score $\beta_i \in \mathbb{R}$, and the outcome of comparing papers i and j is a Bernoulli random variable, where the probability of paper i beating j is determined by a logistic function of the score difference $\beta_i - \beta_j$, i.e.,

$$p_{ij} := \mathbb{P}(y_{ij} = 1 | (i, j)) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} = \frac{1}{1 + e^{-(\beta_i - \beta_j)}} \tag{1}$$

   To estimate these scores β based on the outcome of observed pairwise comparisons, we utilize the maximum likelihood estimator on the Bradley-Terry model, which maximizes the log-likelihood function below:

$$\mathcal{L}(\beta) = \sum_{(i,j,y) \in \mathcal{O}} \left[y_{ij} \log(p_{ij}) + (1 - y_{ij}) \log(1 - p_{ij}) \right]. \tag{2}$$

3. **Ranking Inference**: Once the scores are estimated, they can be used to rank all candidates in a descending order, with a higher score indicating "stronger" candidates with higher quality.

[1] Theorem 4 in [34] shows that if $m > 12n \log n$ pairs are sampled uniformly at random, then with high probability the estimate $\hat{\theta}$ satisfies $\|\hat{\theta} - \theta^*\| = O\left(n\sqrt{\frac{\log n}{m}}\right)$.

Since all our experiments were conducted using GPT-4o mini [36] for its balance of performance and computational cost, we will refer to our proposed mechanism as the GPT ranking system in the following sections.

4 Experiments

We now empirically validate whether our proposed pairwise-based ranking framework better identifies high-impact papers compared to conventional rating-based approaches.

4.1 Dataset Desciption

We collected papers from major ML conferences publicly available on OpenReview, including *ICLR*, *NeurIPS*, *CoRL*, and *EMNLP*. Papers were categorized into different decision outcomes (e.g., accepted vs. rejected, main vs. findings track, oral vs. poster, etc.), reflecting the quality assessments by the peer review process. The paper PDFs and corresponding decisions were retrieved using the OpenReview API (`https://docs.openreview.net/`). We used ScienceBeam [12] to extract the title, abstract, figure and table captions, and the main text. For each conference, we applied our method to the submitted papers and maintained the original distribution of papers across decision categories. Experiment setup details are available in the Appendix B.

4.2 Agent Scaling Boosts the Performance of GPT Ranking System

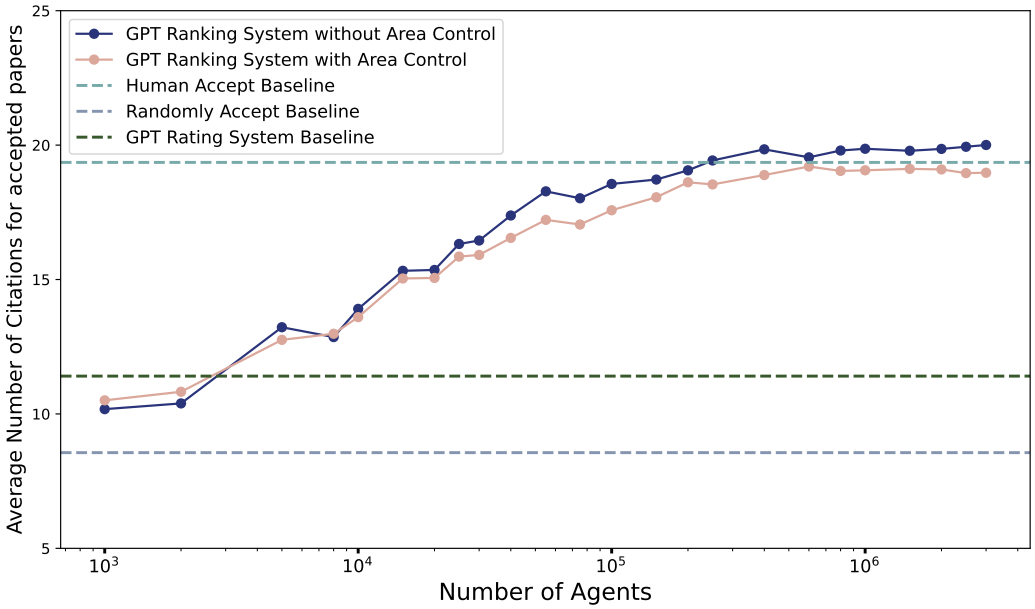

Figure 2: **Scaling of average citation counts for accepted papers by GPT ranking system with increasing pairwise comparisons**. We introduce two variants of our GPT ranking system: a basic version without area control, and one with area control, which ensures that the distribution of accepted papers across primary research areas matches that of human peer review. These two variants are compared against three baselines: (1) human acceptance decisions, (2) random acceptance based on the conference acceptance rate, and (3) GPT rating system (acceptance depends on the rating synthesized by a meta-reviewer from three independent GPT reviewers). As the number of comparisons increases, both variants steadily improve and approach human-level performance. At maximum scale, the GPT ranking system without area control achieves 20.00 average citations, while the version with area control reaches 18.97 average citations. These compare favorably to human acceptance decisions (19.36 average citations), and significantly outperform both random acceptance (8.56 average citations) and the GPT rating system baseline (11.41 average citations).

https://doi.org/10.52202/085713-0394

We investigate the scaling performance of our GPT ranking system with the number of agents, where each agent conducts a single pairwise comparison between randomly sampled papers from *ICLR 2024*. Figure 2 shows the overall academic impact (measured by their average citation counts) of the accepted papers as the number of agents increases. In particular, the system exhibits a steady increasing trend in average citations as the number of agents scales from around 10^3 to 10^5. At insufficient sample sizes (e.g., 10^3 agents), the system performance approaches that of random acceptance. Since each paper receives too few comparisons, it fails to establish reliable relative rankings across the entire pool of 7158 papers. As the scale increases beyond 10^5, the growth rate gradually slows down, with the average citation count eventually plateauing at approximately 20.

Given that the GPT ranking system demonstrates different acceptance patterns across research areas compared to humans (Section 4.5), we include two variants in our evaluation: one with area control, which maintains the same distribution of accepted papers across primary research areas as the human peer review system, and one without such controls. Nevertheless, they all yield comparable results to human-accepted papers at the convergence point. In the subsequent analysis, we focus on the GPT ranking system without area control as it better reflects the true properties of our system without deliberate controls.

We also compare our GPT ranking system with another baseline that follows a similar setup as prior works [27, 18]. This baseline, called the GPT rating system, involves three GPT reviewers independently generating reviews for each paper. Another agent then acts as a meta-reviewer to synthesize these individual reviews and provide a final rating (on a scale of 1 to 10). Acceptance decisions are made based on the rating provided by the meta-reviewer. However, it revealed a critical limitation: GPT-generated ratings exhibited less diversity, with most ratings concentrated around 6 and 7. Consequently, the system's ability to differentiate papers of varying quality was significantly constrained, leading to only minor improvements over the random baseline.

4.3 Discriminative Capability of GPT ranking system in Research Evaluation

With sufficient scale, our GPT ranking system approaches human-level performance in terms of average citation counts for accepted papers. However, beyond raw citation averages, a key question remains: does the GPT ranking system effectively distinguish between highly influential and less impactful papers in a manner comparable to human reviewers? To address this question, we analyzed the system's ability to discriminate papers across the entire citation distribution spectrum.

Across multiple top AI conferences (*ICLR 2024*, *EMNLP 2023*, *ICLR 2023*, and *CoRL 2023*) and under various decision conditions (accepted vs. rejected, main vs. findings track, oral vs. poster, etc.), papers ranked highly by our system consistently received more citations than those ranked lower (Figure 3, Supp Figures 8, 9). This pattern closely mirrors the citation advantages observed in human-selected papers across the same decision categories. The statistical significance of these differences (indicated by asterisks) further validates that our GPT ranking system can serve as reliable proxies for human peer review when identifying work likely to generate greater impact in the research community. The consistency of this pattern across different conferences and years suggests that as the system employs sufficient agents to conduct comprehensive pairwise comparisons, it could also capture the subtle quality signals that correlate with future scholarly influence.

4.4 Consistency Analysis between GPT Ranking System and Human Peer Review

We further examined the decision consistency between the GPT ranking system and conventional human peer review in *ICLR 2024* and *ICLR 2023*. As shown in Table 1, each paper was independently categorized by human reviewers and by the GPT ranking system into one of four decision categories — Oral, Spotlight, Poster, or Reject. Papers withdrawn after review releases were considered as rejected for this analysis.

Notably, 41.0% of the papers accepted by human reviewers in *ICLR 2024* were also accepted by the GPT ranking system (42.2% for papers accepted in *ICLR 2023*; Supp Table 4). The result aligns with findings from the *NeurIPS 2021* consistency experiment [3], where 48.0% of papers accepted by the first committee were also accepted by the second committee. This level of agreement between AI and human reviewers is roughly on par with the consistency observed between independent human review committees, suggesting that the GPT ranking system could offer valuable insights for identifying

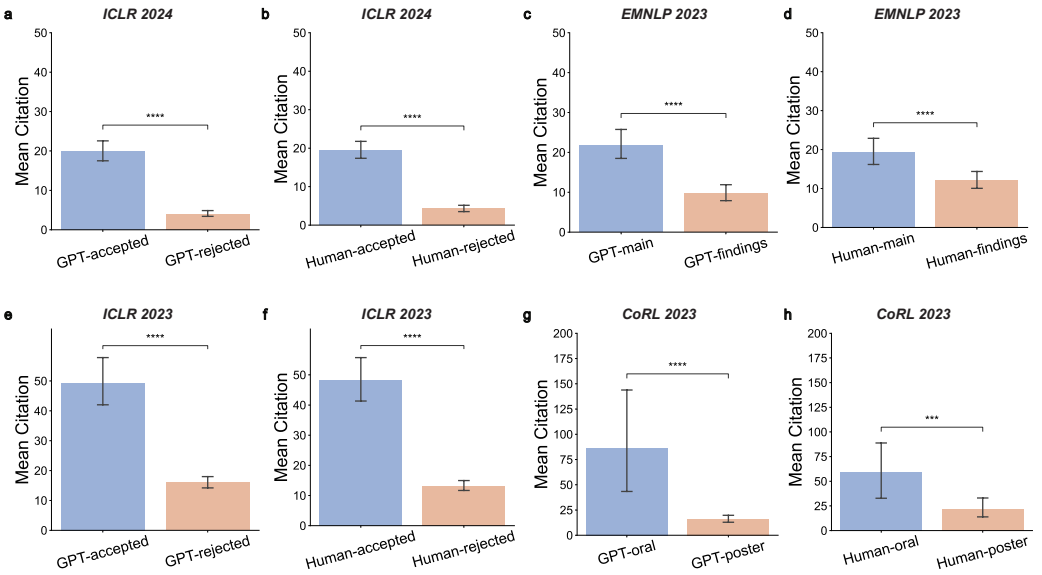

Figure 3: **Comparison of mean citation counts across multiple AI conferences under two-fold decisions (e.g., accepted vs. rejected, main vs. findings track, oral vs. poster)**. The results consistently show that higher-tier papers (both GPT-selected and human-selected) receive substantially higher citation counts than lower-tier papers. This implies that our GPT ranking system can effectively distinguish more influential papers, performing comparably to human peer review system in identifying works likely to generate greater impact in the academic community. Error bars represent 95% confidence intervals. *P $<$ 0.05, **P $<$ 0.01, ***P $<$ 0.001, and ****P $<$ 0.0001.

Table 1: **Summary of recommendations for ICLR 2024 papers by two review systems: the human peer review (Human) and the GPT ranking system (GPT).** Each row represents humans' decision, while each column shows how the GPT ranking system categorized the same papers.

Human\GPT	Oral	Spotlight	Poster	Reject
Oral	6	11	28	41
Spotlight	10	32	130	191
Poster	29	116	556	1,085
Reject	41	204	1,072	3,606

potentially impactful papers that might otherwise be rejected due to the inherent randomness in the traditional peer review process.

4.5 Different Acceptance Patterns across Research Areas

We found notable disparities between the GPT ranking system and human peer review in their acceptance rates of several research areas (Figure 4, Supp Figure 10): GPT ranking system shows significantly higher acceptance rates for applied research areas such as robotics/autonomy applications (0.56 vs. 0.32) and societal considerations (0.51 vs. 0.30). In contrast, theoretical areas that received relatively high acceptance rates from human reviewers, such as learning theory (0.50 vs. 0.12) and optimization (0.31 vs. 0.12), had much lower acceptance rates from the GPT ranking system.

There are several ways to interpret these findings. First, LLMs are trained on massive web-crawled corpora rich in practical, application-oriented content. Consequently, these models may exhibit an inherent preference for studies that demonstrate immediate real-world applicability. Alternatively, large language models often struggle with complex mathematical reasoning [33], which could hinder their understanding of papers that require deep theoretical and mathematical rigor beyond memorized

https://doi.org/10.52202/085713-0394

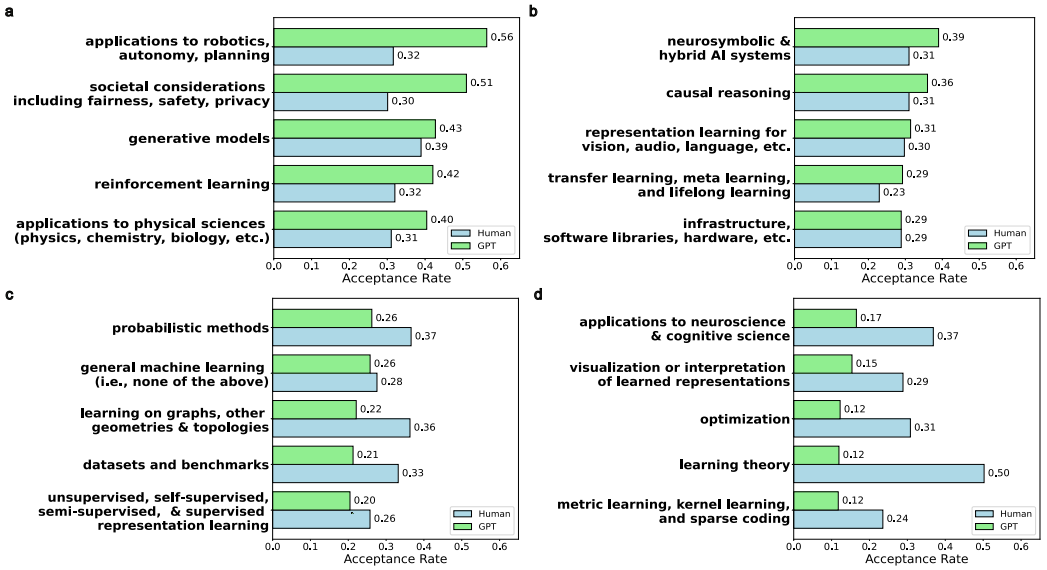

Figure 4: **Comparative acceptance rates of *ICLR' 24* papers by humans and GPT ranking system across research areas**. We sort areas by the GPT ranking system's acceptance rate from highest to lowest. The GPT ranking system exhibits noticeably higher acceptance rates in application-oriented fields compared to human reviewers, showing the most striking disparities in robotics (0.56 vs. 0.32) and societal considerations (0.51 vs. 0.30). In contrast, for more theoretical or methodologically focused areas, it assigns significantly lower acceptance rates than human reviewers. Learning theory demonstrates the largest gap, with acceptance rate at 0.12 versus humans at 0.50. The categorization is based on the 20 primary areas by *ICLR' 24*.

patterns. This limitation can result in a preference for research with more concrete implementations than abstract theoretical work. Future interdisciplinary research could explore these hypotheses.

4.6 Potential Bias against Novel Research Topics

We explored how our GPT ranking system influences the novelty of topics in selected papers. To measure the novelty of the topics, we embed each paper's abstract with OpenAI's text-embedding-3-small model, generating a vector representation for each abstract. We then compute the distance between each paper's vector and the closest neighbor within the abstracts from the same conference. A smaller distance indicates greater similarity between abstracts, thus lower novelty of the topic.

Through examination of papers categorized into top decision tiers by human reviewers and our GPT ranking system, our analysis revealed a consistent pattern (Figure 5): papers receiving higher rankings from the GPT system exhibited substantially smaller average distances to their nearest neighbors compared to papers selected by human reviewers. This might indicate that GPT exhibit a bias toward papers that cover topics similar to those in the existing literature, potentially undervaluing work that delves into novel research areas.

4.7 Impact on Academic Inequality

We also observed that our GPT ranking system tend to favor papers from established research institutions, potentially amplifying existing imbalances in the academic ecosystem. To quantify this imbalance, we track the first authors' affiliation distribution across papers presented in higher decision tiers. Using the Gini coefficient, a statistical measure commonly applied to income inequality, we quantified the degree of publication concentration across institutional affiliations.

We found a concerning pattern of institutional bias: the GPT ranking system consistently exhibited higher institutional inequality compared to humans, with significantly higher Gini coefficients across all conferences studied. The most significant disparity was observed in *ICLR 2023* and *ICLR 2024*.

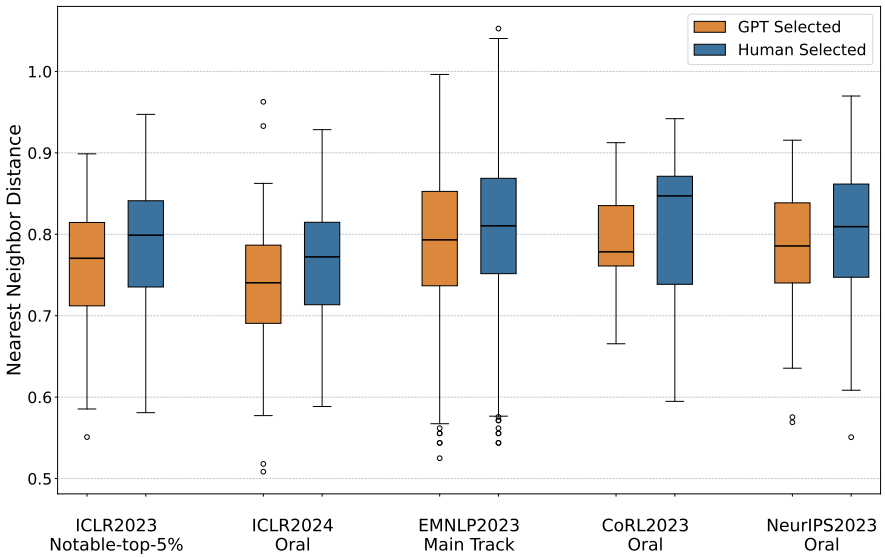

Figure 5: **Nearest neighbor distances of papers selected for top-tier acceptance by GPT ranking system versus human reviewers**. Higher distances indicate greater novelty. Human-selected papers (blue) consistently show higher nearest neighbor distances than GPT-selected papers (orange) across all conferences. The differences are statistically significant ($p < 0.05$) for ICLR 2023 Notable-top-5%, ICLR 2024 Oral, and EMNLP 2023 Main Track. For CoRL 2023 Oral and NeurIPS 2023 Oral, we did not observe statistically significant differences, which may be due to smaller sample sizes.

In *ICLR 2023*, 43.8% of GPT ranking system's top-tier papers came from 10 institutions compared to only 27.0% in human decisions, while in *ICLR 2024*, the gap persisted with 37.2% of GPT ranking system's top selections coming from 10 institutions versus 26.7% in human evaluations.

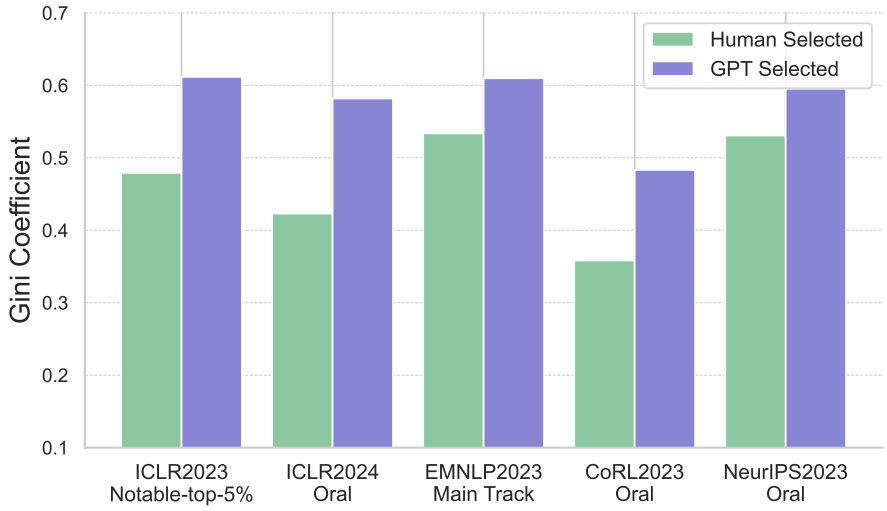

Figure 6: **Comparison of Gini coefficients for papers selected by humans versus GPT ranking system across conferences**. Larger Gini coefficients represent greater inequality or imbalance in the distribution. The consistently higher Gini coefficients for GPT-selected papers (purple bars) compared to human-selected papers (green bars) indicate that GPT ranking system exhibit greater institutional concentration, potentially favoring papers from established research institutions.

https://doi.org/10.52202/085713-0394

5 Conclusion

Existing LLM-assisted peer review efforts largely replicate traditional workflows, framing evaluation as absolute scoring followed by aggregation [18, 27, 49, 55, 61]. However, we argue that the scalability of LLMs open new opportunities to redesign scholarly evaluation structures rather than merely automate human processes. In this work, we introduce and explore a novel mechanism using LLM agents to evaluate academic papers through pairwise comparisons. Rather than assigning isolated scores, the system constructs a global ranking by aggregating local relative judgments between submissions. Through empirical experiments, we find that with sufficient scale, our system effectively identifies high-impact papers across multiple conferences, significantly exceeding traditional rating-based methods. The overlap between papers accepted by our system and those accepted by human reviewers aligns with human-human agreement levels observed in previous studies, suggesting potential value as a complementary tool in the review process.

At the same time, our analysis reveals important challenges. The system exhibits area-specific preferences that diverge from human reviewers, shows a measurable decline in topic novelty among selected papers, and concentrates acceptances among a smaller set of prestigious institutions than human-selected papers. These patterns suggest that while scalable LLM-driven evaluation systems hold promise, careful design will be critical to ensure they promote diversity, equity, and innovation rather than reinforcing existing hierarchies.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] B. Alberts, B. Hanson, and K. L. Kelner. Reviewing peer review, 2008.

[3] A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.

[4] L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 64(1):45–80, 2008.

[5] L. Bornmann and R. Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222, 2015.

[6] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[7] M. Breuning, J. Backstrom, J. Brannon, B. I. Gross, and M. Widmeier. Reviewer fatigue? why scholars decline to review their peers' work. *PS: Political Science & Politics*, 48(4):595–600, 2015.

[8] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *European Conference on Information Retrieval*, pages 16–27. Springer, 2008.

[9] M. D'Arcy, T. Hope, L. Birnbaum, and D. Downey. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.

[10] M. Demeter, A. Jele, and Z. B. Major. The international development of open access publishing: A comparative empirical analysis over seven world regions and nine academic disciplines. *Publishing Research Quarterly*, 37(3):364–383, 2021.

[11] M. S. Di Bitetti and J. A. Ferreras. Publish (in english) or perish: The effect on citation rate of using languages other than english in scientific publications. *Ambio*, 46:121–127, 2017.

[12] D. Ecer and G. Maciocci. Sciencebeam - using computer vision to extract pdf data. Elife Blog Post, https://elifesciences.org/labs/5b56aff6/sciencebeam-using-computer-vision-to-extract-pdf-data, 2017. Accessed: 26 March 2025.

[13] C. W. Fox, A. Y. Albert, and T. H. Vines. Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review*, 2:1–6, 2017.

[14] C. W. Fox, J. Meyer, and E. Aimé. Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Functional Ecology*, 37(5):1144–1157, 2023.

[15] M. L. Gordon, M. S. Lam, J. S. Park, K. Patel, J. Hancock, T. Hashimoto, and M. S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[16] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] M. Hosseini and S. P. Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research integrity and peer review*, 8(1):4, 2023.

 https://doi.org/10.52202/085713-0394

[18] Y. Jin, Q. Zhao, Y. Wang, H. Chen, K. Zhu, Y. Xiao, and J. Wang. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*, 2024.

[19] M. Lamont. *How professors think: Inside the curious world of academic judgment*. Harvard University Press, 2009.

[20] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin. Bias in peer review. *Journal of the American Society for information Science and Technology*, 64(1):2–17, 2013.

[21] W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, L. Chen, H. Ye, S. Liu, Z. Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *International Conference on Machine Learning*, pages 29575–29620. PMLR, 2024.

[22] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Y. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, D. A. McFarland, and J. Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024. doi: 10.1056/AIoa2400196. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400196.

[23] J. Lin, J. Song, Z. Zhou, Y. Chen, and X. Shi. Moprd: A multidisciplinary open peer review dataset. *Neural Comput. Appl.*, 35(34):24191–24206, Sept. 2023. ISSN 0941-0643. doi: 10.1007/s00521-023-08891-5. URL https://doi.org/10.1007/s00521-023-08891-5.

[24] J. Lin, J. Song, Z. Zhou, Y. Chen, and X. Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101830. URL https://www.sciencedirect.com/science/article/pii/S156625352300146X.

[25] D. Lindsey. Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid. *Scientometrics*, 15:189–203, 1989.

[26] R. Liu and N. B. Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing, 2023. URL https://arxiv.org/abs/2306.00622.

[27] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[28] Y. Lu and Y. Kong. Calibrating "cheap signals" in peer review without a prior. *Advances in Neural Information Processing Systems*, 36:20640–20661, 2023.

[29] S. Ma, Q. Chen, X. Wang, C. Zheng, Z. Peng, M. Yin, and X. Ma. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. *arXiv preprint arXiv:2403.16812*, 2024.

[30] A. McCook. Is peer review broken? submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. what's wrong with peer review? *The scientist*, 20(2):26–35, 2006.

[31] B. Mehmani and A. Ghildiyal. Rethinking Reviewer Fatigue. *EON*, nov 15 2024. https://eon.pubpub.org/pub/3ana9ey0.

[32] B. Merriman. Peer review as an evolving response to organizational constraint: Evidence from sociology journals, 1952–2018. *The American Sociologist*, 52:341–366, 2021.

[33] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL https://arxiv.org/abs/2410.05229.

[34] S. Negahban, S. Oh, and D. Shah. Rank centrality: Ranking from pairwise comparisons. *Oper. Res.*, 65(1):266–287, Feb. 2017. ISSN 0030-364X. doi: 10.1287/opre.2016.1534. URL https://doi.org/10.1287/opre.2016.1534.

[35] L. Niu, N. Xue, and C. Pöpper. Unveiling the sentinels: Assessing ai performance in cybersecurity peer review. *arXiv preprint arXiv:2309.05457*, 2023.

[36] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-04-02.

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[38] Z. Robertson. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*, 2023.

[39] D. Scherbakov, N. Hubig, V. Jansari, A. Bakumenko, and L. A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. *arXiv preprint arXiv:2409.04600*, 2024.

[40] N. B. Shah. Challenges, experiments, and computational solutions in peer review. *Commun. ACM*, 65(6):76–87, May 2022. ISSN 0001-0782. doi: 10.1145/3528086. URL https://doi.org/10.1145/3528086.

[41] Y. Shao, V. Samuel, Y. Jiang, J. Yang, and D. Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.

[42] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652, 2023.

[43] C. Si, D. Yang, and T. Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

[44] F. Squazzoni, G. Bravo, M. Farjam, A. Marusic, B. Mehmani, M. Willis, A. Birukou, P. Dondio, and F. Grimaldo. Peer review and gender bias: A study on 145 scholarly journals. *Science Advances*, 7(2):eabd0299, 2021. doi: 10.1126/sciadv.abd0299. URL https://www.science.org/doi/abs/10.1126/sciadv.abd0299.

[45] S. Srinivasan and J. Morgenstern. Auctions and peer prediction for academic peer review, 2023. URL https://arxiv.org/abs/2109.00923.

[46] I. Stelmakh, N. B. Shah, A. Singh, and H. Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17, 2021.

[47] I. Stelmakh, C. Rastogi, R. Liu, S. Chawla, F. Echenique, and N. B. Shah. Cite-seeing and reviewing: A study on citation bias in peer review. *PLOS ONE*, 18(7):e0283980, July 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0283980. URL http://dx.doi.org/10.1371/journal.pone.0283980.

[48] W. J. Su. You are the best reviewer of your own papers: an owner-assisted scoring mechanism. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[49] C. Tan, D. Lyu, S. Li, Z. Gao, J. Wei, S. Ma, Z. Liu, and S. Z. Li. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.

[50] J. P. Verharen. Chatgpt identifies gender disparities in scientific peer review. *Elife*, 12:RP90230, 2023.

[51] J. Wang. Unpacking the matthew effect in citations. *Journal of Informetrics*, 8(2):329–339, 2014.

[52] J. Wang and N. B. Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 864–872, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

 https://doi.org/10.52202/085713-0394

[53] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[54] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[55] Y. Weng, M. Zhu, G. Bao, H. Zhang, J. Wang, Y. Zhang, and L. Yang. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.

[56] J. Wu, H. Xu, Y. Guo, and W. Su. An isotonic mechanism for overlapping ownership, 2025. URL https://arxiv.org/abs/2306.11154.

[57] Y. Yan, W. J. Su, and J. Fan. Isotonic mechanism for exponential family estimation in machine learning peer review. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf025, 2025.

[58] R. Ye, X. Pang, J. Chai, J. Chen, Z. Yin, Z. Xiang, X. Dong, J. Shao, and S. Chen. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review, 2024. URL https://arxiv.org/abs/2412.01708.

[59] J. Zhang, H. Zhang, Z. Deng, and D. Roth. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv preprint arXiv:2211.06398*, 2022.

[60] R. Zhou, L. Chen, and K. Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.816/.

[61] M. Zhu, Y. Weng, L. Yang, and Y. Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.

[62] Z. Zhuang, J. Chen, H. Xu, Y. Jiang, and J. Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103332. URL https://www.sciencedirect.com/science/article/pii/S1566253525004051.

[63] H. Zuckerman and R. K. Merton. Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, pages 66–100, 1971.

A Discussion

A.1 Limitations

In this study, we rely on citation counts as a proxy for academic impact. Although citation metrics are widely used and provide a quantifiable measure of influence, they are affected by numerous factors beyond quality, such as the visibility of research communities, prevailing research trends, and established reputations [25, 4, 51, 11]. More critically, papers selected by human reviewers inherently receive more visibility through conference presentations and proceedings. The fact that our GPT ranking system identified papers with comparable citation impact despite this disadvantage demonstrates a certain level of effectiveness in our approach. Nevertheless, this circularity problem makes it difficult to establish a truly independent measure of quality. To further verify the robustness of our proposed method, we conducted experiments using alternative measures. Specifically, we measured the Spearman correlation between our system's output and human review scores. In both ICLR 2023 and ICLR 2024, the BT scores produced by our mechanism show moderate correlation with the average human ratings—27% and 24%, respectively. This indicates that our system aligns with signals in the current review process, while not fully replicating its outcomes. We believe this complementary nature is essential for identifying potential improvements to existing peer review systems.

Furthermore, while we exclusively used GPT-4o mini for our large-scale experiments due to its balance of performance and computational cost, we acknowledge that there are other diverse LLMs capable of academic assessment, and they might yield substantially different results. A more comprehensive evaluation across multiple models would provide better insight into the generalizability of our method and findings. In Appendix C.4, we tested two additional models beyond GPT-4o mini: Gemini 2.0 Flash and Claude-3-Haiku-20240307. We scaled the number of comparisons to over 10^6 and found that, in both cases, our pairwise comparison framework performs significantly better than the GPT rating system baseline. Specifically, papers selected by Gemini 2.0 Flash receive an average of 18.3 citations (vs. 11.4), while those chosen by Claude-3-Haiku average 16.8 (vs. 11.4). These results suggest that the advantage of our approach is robust across different LLMs.

Finally, while our current work constructs a single global ranking, future extensions could explore personalized or multi-objective evaluation systems that explicitly account for epistemic diversity and evolving community goals.

A.2 Expanding Peer Review through Pairwise Evaluation

Beyond the immediate results, our framework and explorations open broader directions for redesigning scholarly review systems.

First, because pairwise comparisons produce local relative judgments that can be incrementally aggregated, our approach naturally lends itself to continuous evaluation. Rather than operating within a fixed submission and decision timeline, conferences could maintain an ongoing review pipeline, where new papers are progressively compared against the existing submission pool. Such a dynamic process could allow for rolling acceptances, faster feedback loops, and better accommodation of late-breaking research.

Second, pairwise evaluation offers distinct advantages for emerging or interdisciplinary fields where traditional scoring rubrics are poorly defined or difficult to establish. Absolute scoring requires consensus on evaluation criteria and careful calibration across reviewers, which can be challenging in fast-moving or nascent research areas. Comparative judgments, by focusing on relative rather than absolute assessments, can surface high-potential work even when shared evaluation norms are still evolving, making the system more adaptable to domains where innovation resists rigid checklist-based quantification.

Third, the flexibility of the aggregation process suggests opportunities for personalized and diversified peer review. Different weighting schemes could prioritize novelty, interdisciplinarity, methodological rigor, or other dimensions based on the goals of specific conferences, tracks, or even reviewer communities. In principle, such mechanisms could allow peer review to better reflect the heterogeneous values of different research communities, moving beyond the one-size-fits-all model [15] currently dominant in scientific publishing.

 https://doi.org/10.52202/085713-0394

Finally, an important direction for future exploration lies in integrating human expertise and LLM evaluations [41, 29]. Rather than viewing LLM-based and human-based assessments as competing alternatives, hybrid models could combine the strengths of both: leveraging LLMs for large-scale, consistent pairwise comparisons while relying on human reviewers to provide deeper qualitative insights, assess boundary cases, and adjudicate particularly novel or interdisciplinary submissions. Designing effective protocols for human-AI collaboration in peer review could further enhance both the scalability and the fairness of the evaluation process.

Together, these directions highlight how reframing peer review around relative comparisons, enriched by human-AI collaboration, could support a more scalable, inclusive, and adaptive scholarly communication ecosystem.

B Experimental Details

B.1 Dataset Details

Here we include additional details on the datasets used for our experiments.

Table 2: **Academic paper data from major ML conferences.**

Conference	Decision Types	# of Papers
ICLR 2023	• Notable-top-5%	89
	• Notable-top-25%	281
	• Poster	1,193
	• Reject (Withdrawn submissions included)	3,303
ICLR 2024	• Accept (Oral)	86
	• Accept (Spotlight)	363
	• Accept (Poster)	1,786
	• Reject (Withdrawn submissions included)	4,923
NeurIPS 2023	• Accept (Oral)	67
	• Accept (Spotlight)	374
	• Accept (Poster)	2,748
EMNLP 2023	• Accept-Main	975
	• Accept-Findings	993
CoRL 2023	• Accept (Oral)	32
	• Accept (Poster)	166

B.2 Implementation Details

We use the official OpenAI's Batch API for GPT-4o mini and set temperature as 0 during pairwise comparison. The number of API calls and estimated cost for each simulated conference is as follows:

Table 3: **API usage and estimated costs for pairwise comparison across each conference.** *EMNLP 2023* and *CoRL 2023* used all available pairs, while others were randomly downsampled to 3M pairs.

Conference	# of API Calls	Estimated Cost (USD)
ICLR 2023	3,000,000	$\sim 1,350$
ICLR 2024	3,000,000	$\sim 1,350$
NeurIPS 2023	3,000,000	$\sim 1,350$
EMNLP 2023	3,871,056	$\sim 1,700$
CoRL 2023	39,006	~ 17

https://doi.org/10.52202/085713-0394

B.3 LLM prompts used in the study

```
Please act as an impartial judge and evaluate the quality of the
    following two papers. As the area chair for a top ML conference,
    you can only select one paper. Start with a brief meta-review/
    reasoning of the pros and cons for each paper (two sentences), and
     then provide your choice in a binary format. Start with a brief
    meta-review/reasoning of the pros and cons for each paper,
    focusing on novelty, significance, clarity, methodology, and
    practical implications. Be very critical and do not be biased by
    what the author claimed. Finally, provide your choice in a binary
    format.


Please provide your analysis in JSON format.

### Paper 1:
Submission Title: {title}

‘‘‘
Abstract: {abstract}

Figures Captions:
{figure_and_table_captions}

Main:
{main_content}
‘‘‘


### Paper 2:
Submission Title: {title}

‘‘‘
Abstract: {abstract}

Figures Captions:
{figure_and_table_captions}

Main:
{main_content}
‘‘‘

Your JSON output should look like this:

{{
    "paper_1_review": "Your meta-review and reasoning for paper 1",
    "paper_2_review": "Your meta-review and reasoning for paper 2",
    "chosen_paper": "paper_1 or paper_2"
}}
"""
```

Figure 7: Example prompt for pairwise comparison.

C Additional Results

C.1 Discriminative Capability of GPT ranking system in Research Evaluation

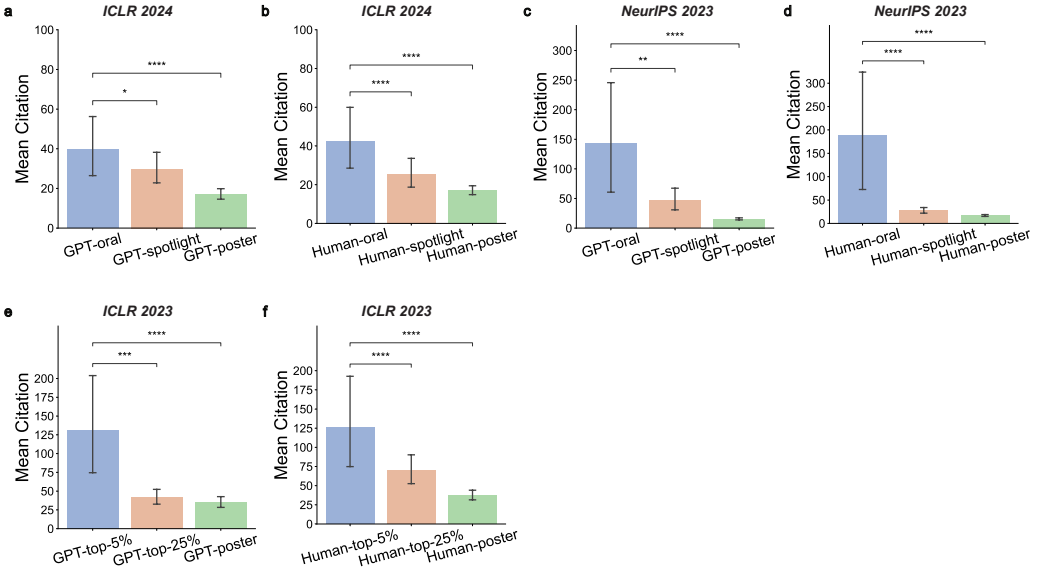

Figure 8: **Comparison of mean citation counts across multiple AI conferences under fine-grained decisions in accepted papers.** The results consistently show that higher-tier papers (both GPT-selected and human-selected) receive overall higher influence than lower-tier papers. Error bars represent 95% confidence intervals. *P $<$ 0.05, **P $<$ 0.01, ***P $<$ 0.001, and ****P $<$ 0.0001.

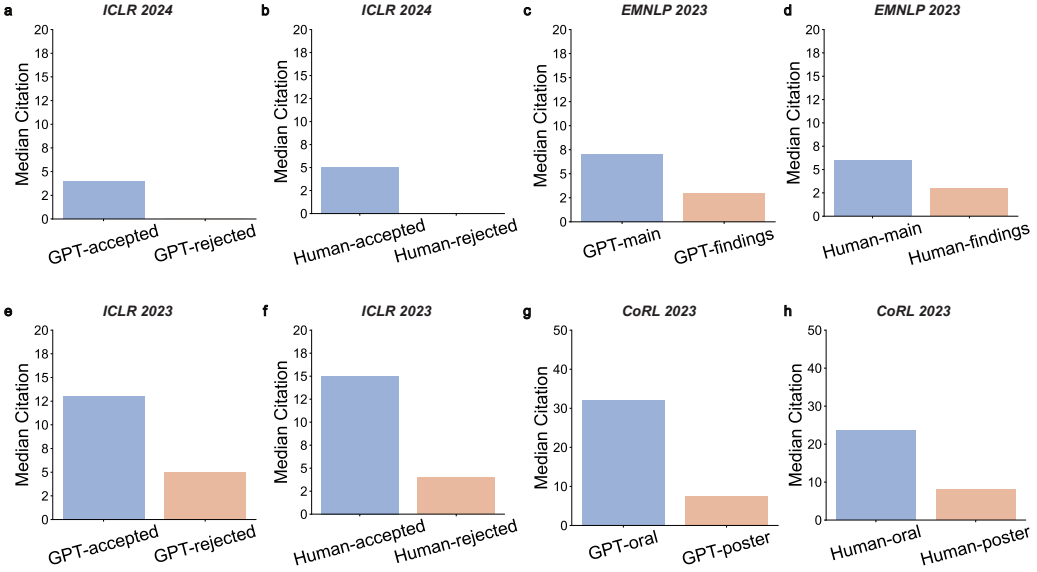

Figure 9: **Comparison of median citation counts across multiple AI conferences under two-fold decisions (e.g., accepted vs. rejected, main vs. findings track, oral vs. poster).** The results consistently show that higher-tier papers (both GPT-selected and human-selected) receive overall higher influence than lower-tier papers.

https://doi.org/10.52202/085713-0394

C.2 Consistency Analysis between GPT Ranking System and Human Peer Review

Table 4: **Summary of recommendations for ICLR 2023 papers by two review systems: the human peer review (Human) and the GPT ranking system (GPT).** Each row represents humans' decision, while each column shows how the GPT ranking system categorized the same papers.

Human\GPT	Notable top 5%	Notable top 25%	Poster	Reject
Notable top 5%	10	9	26	44
Notable top 25%	11	27	95	148
Poster	29	93	360	711
Reject	39	152	712	2,400

C.3 Different Acceptance Patterns across Research Areas

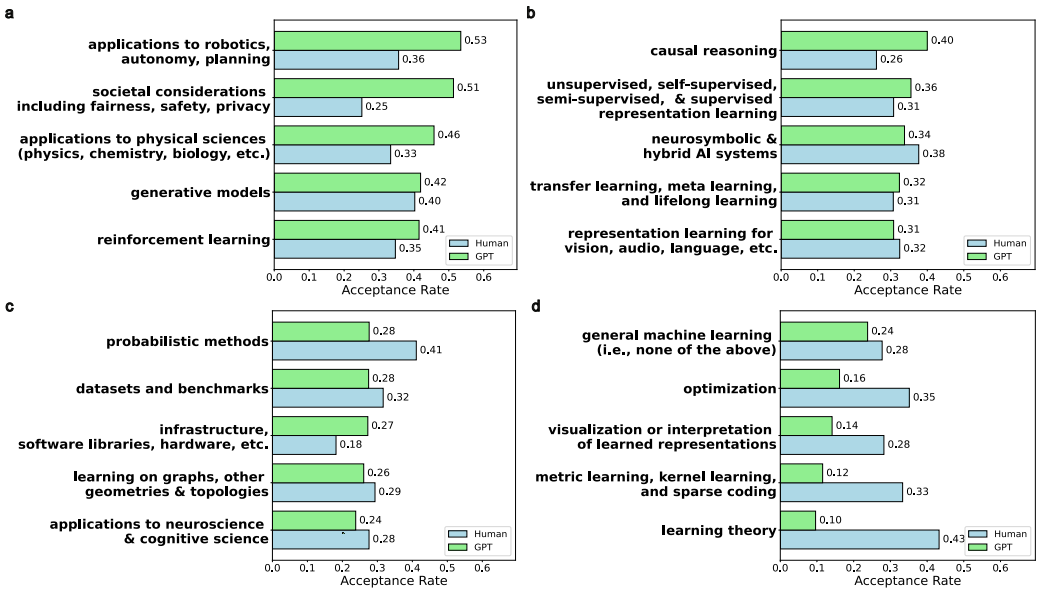

Figure 10: **Comparative acceptance rates of *ICLR' 23* papers by humans and GPT ranking Systems across research areas**. We sort areas by the GPT ranking system's acceptance rate from highest to lowest. The GPT ranking system exhibits noticeably higher acceptance rates in application-oriented fields compared to human reviewers, showing the most striking disparities in robotics (0.53 vs. 0.36) and societal considerations (0.51 vs. 0.25). In contrast, for more theoretical or methodologically focused areas, it assigns significantly lower acceptance rates than human reviewers. Learning theory demonstrates the largest gap, with acceptance rate at 0.10 versus humans at 0.43.

https://doi.org/10.52202/085713-0394

C.4 Using other LLMs as agents

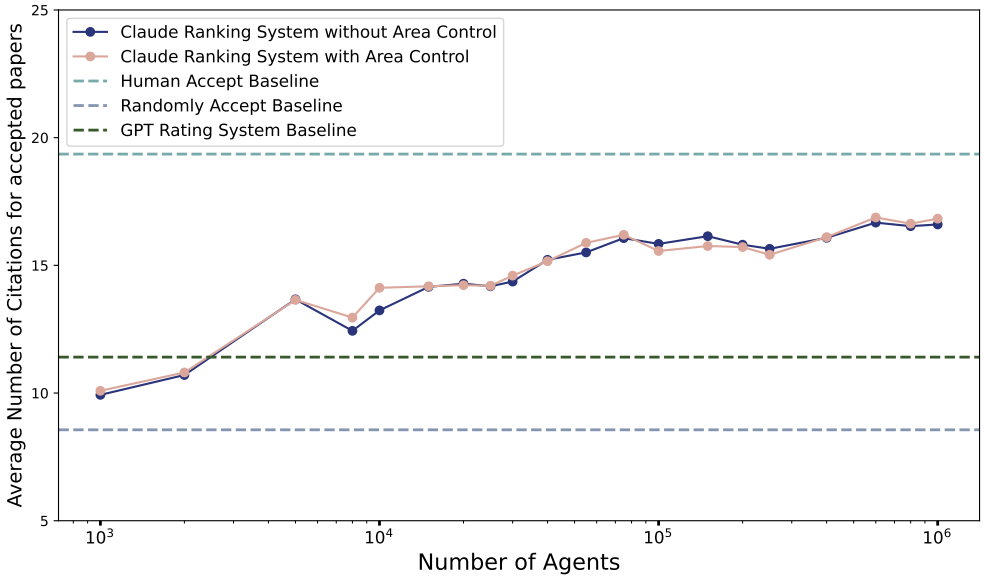

Figure 11: **Scaling of average citation counts for accepted papers by Claude ranking system with increasing pairwise comparisons**. We observe a similar temporal scaling pattern as in the GPT ranking system, with citation counts increasing as the number of pairwise comparisons grows.

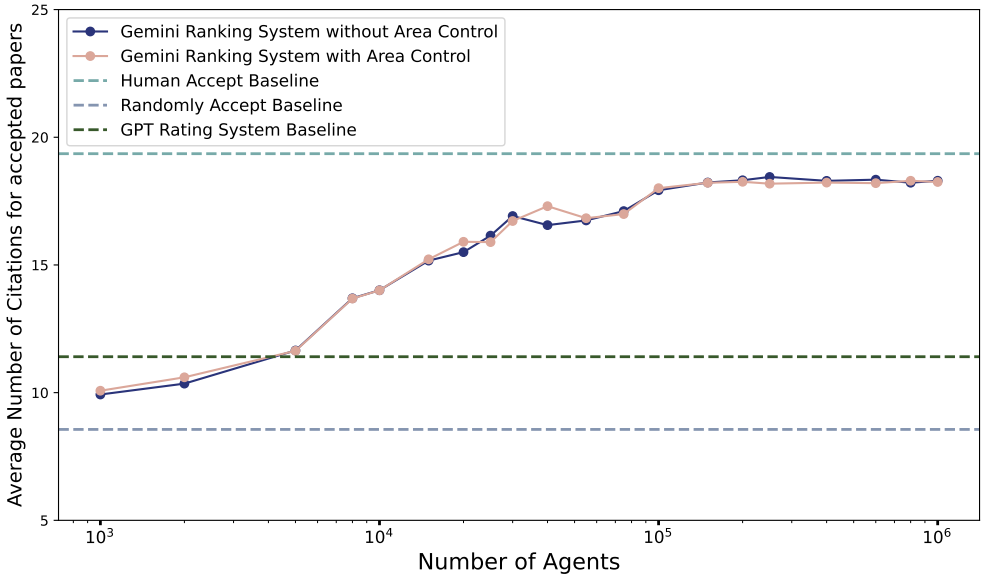

Figure 12: **Scaling of average citation counts for accepted papers by Gemini ranking system with increasing pairwise comparisons**. We observe a similar temporal scaling pattern as in the GPT ranking system, with citation counts increasing as the number of pairwise comparisons grows.

https://doi.org/10.52202/085713-0394

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the primary contributions of the paper, including the introduction of LLM-based pairwise ranking system, the significant performance improvements and emergent biases in the selection process through experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Appendix A.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There are no theoretical results in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explained comprehensive details on dataset (Section 4.1, Appendix B.1), experimental setups (Appendix B), and methods used (Section 3), ensuring that the results can be reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/085713-0394

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All numerical experiments performed are simple enough to reproduce without access to code. Code and data are available upon request.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are explained in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars and includes information on the statistical significance of the experimental results, ensuring the reliability of the findings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments mainly involved inference using API calls to GPT-4o mini. The type of compute workers, memory and time of execution are managed by the API provider. Details about the number of API calls and configurations are provided in Appendix B.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential societal impacts in Section 5, Appendix A.2.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/085713-0394

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such models or datasets are involved.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the creators of existing assets used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/085713-0394

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our LLM-based pairwise ranking system employs LLMs as evaluators to produce binary preference judgments between paper candidates. Please refer to Section 3.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.