ALLOCATE MARGINAL REVIEWS TO BORDERLINE PAPERS USING LLM COMPARATIVE RANKING

Elliot L. Epstein[1], **Rajat Dwaraknath**[1], **John Winnicki**[1], **Thanawat Sornwanee**[1]
[1]Stanford University, Stanford, CA 94305, USA
{epsteine, rajatvd, winnicki, tsornwanee}@stanford.edu

ABSTRACT

This paper argues that large ML conferences should allocate marginal review capacity primarily to papers near the acceptance boundary, rather than spreading extra reviews via random or affinity-driven heuristics. We propose using LLM-based comparative ranking (via pairwise comparisons and a Bradley–Terry model) to identify a borderline band *before* human reviewing and to allocate *marginal* reviewer capacity at assignment time. Concretely, given a venue-specific minimum review target (e.g., 3 or 4), we use this signal to decide which papers receive one additional review (e.g., a 4th or 5th), without conditioning on any human reviews and without using LLM outputs for accept/reject. We provide a simple expected-impact calculation in terms of (i) the overlap between the predicted and true borderline sets (ρ) and (ii) the incremental value of an extra review near the boundary (Δ), and we provide retrospective proxies to estimate these quantities.

1 INTRODUCTION

Conferences typically aim to meet a minimum number of reviews per paper. In practice, there is often some surplus reviewer capacity beyond that minimum. Public review corpora report average reviews per paper above three in several venues and years, implying a marginal surplus of review slots in practice (Ebrahimi et al., 2025; Plank and van Dalen, 2019; Su et al., 2025). The natural question is where those marginal reviews should go to improve decisions the most.

The marginal value of an extra review is typically highest near the acceptance boundary, where score variance is high and decisions are most sensitive to reviewer noise. It is lowest for papers that are clearly strong or clearly weak. In practice, surplus capacity is typically absorbed by load balancing and affinity objectives (Kobren et al., 2019; Charlin et al., 2012; Charlin and Zemel, 2013). Randomized assignment has also been proposed to mitigate manipulation in reviewer matching (Jecmen et al., 2020).

Why has this been difficult to implement? Identifying borderline papers typically requires early human reading, which arrives too late to inform reviewer-count decisions at assignment time. By the time human signal accumulates, the review process is already underway and marginal review allocation is harder to adjust.

Recent long-context LLMs (Gemini Team, 2025) make a lightweight pre-review triage pass feasible, enabling reviewer-count decisions before human reviewing begins. We do not need calibrated absolute scores. We only need a rough ranking that separates likely borderline papers from the rest. Pairwise comparisons are a good fit because comparative judgments are less sensitive to calibration drift than absolute scores and can be aggregated into a robust ordering. This motivates using LLMs not to decide outcomes, but to prioritize where human effort is most valuable.

Given this, we argue that **conferences should allocate marginal review capacity to papers near the acceptance boundary, using LLM-based comparative ranking only to target human effort, while keeping accept/reject decisions fully human and keeping the LLM signal hidden from decision-makers.** Figure 1 summarizes the pipeline. This intervention happens before review begins, so it changes only reviewer-count allocation at assignment time (not review content, rebuttals, or decision procedures). It fits within existing workflows where area chairs can adjust assignments after the initial match. It is a minimal change to existing workflows at NeurIPS, ICML, and ICLR, does not

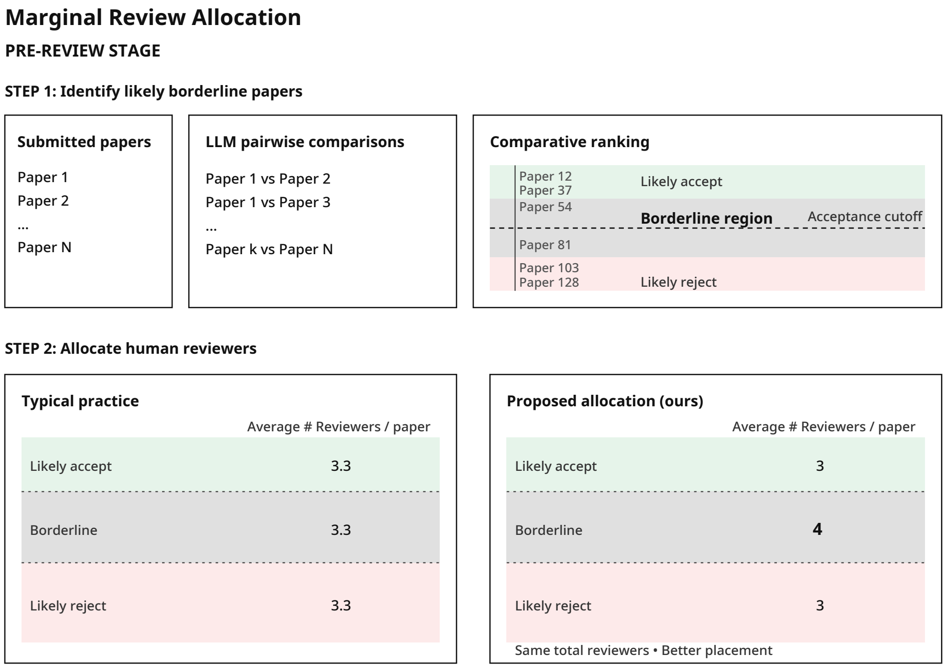


Figure 1: Schematic of the marginal review allocation pipeline. Step 1 uses LLM pairwise comparisons to construct a comparative ranking and identify a borderline band around the acceptance percentile. Step 2 allocates marginal reviews to that band while keeping total reviewer load fixed, with numbers shown for illustration.

introduce a two-stage review that could delay timelines, and remains separate from the final decision process.

Our contributions are threefold. First, we advance a position: LLMs should be used to allocate scarce human reviewer effort to identify where marginal reviews matter most, rather than to automate accept/reject decisions. Second, we formalize marginal review allocation as a policy choice by defining a borderline band from surplus capacity and deriving expected net improved paper decisions in terms of ρ (borderline overlap fraction) and Δ (marginal benefit of an extra review), and we outline a minimal-change pipeline that uses LLM pairwise comparisons with Bradley–Terry ranking to construct the band early while keeping final decisions fully human. Third, we provide initial evidence from 1,000 papers at ICLR 2025 and probe robustness to band fraction and centering.

The paper is organized as follows. Section 2 describes the LLM comparative ranking pipeline and how we fit the Bradley–Terry model. Section 3 situates our proposal in the literature on reviewer assignment and LLM-based reviewing. Section 4 formalizes expected net improved paper decisions, and Section 5 details how we estimate ρ and Δ. Section 6 presents empirical validation and ablations, followed by a discussion in Section 7 and alternative views in Section 8. We close with a conclusion in Section 9.

2 LLM Comparative Ranking

We describe one concrete pipeline to make the proposed allocation policy operational; the position itself does not depend on this specific modeling choice. Conceptually, we only require a coarse ordering that separates likely borderline papers from clearly strong or weak ones. We estimate paper quality with pairwise LLM comparisons and a Bradley-Terry model (Bradley and Terry, 1952; Zhang et al., 2025). For paper i with latent score θ_i, the win probability against paper j is

$$P(i \succ j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)}.$$

Given observed outcomes from many pairwise matches, we fit $\{\theta_i\}$ by maximum likelihood and rank papers by θ_i. Equivalently, for each observed match (i, j) with outcome $y_{ij} \in \{0, 1\}$ indicating whether i wins, the log-likelihood is

$$\sum_{(i,j)} \left[y_{ij} \log P(i \succ j) + (1 - y_{ij}) \log\bigl(1 - P(i \succ j)\bigr) \right],$$

which we maximize over $\{\theta_i\}$ subject to an identifiability constraint (e.g., $\sum_i \theta_i = 0$). The resulting scores are unique up to a constant shift, and their ordering is the ranking.

Match system. We run multiple rounds of random pairings across the submission set, recording a binary winner for each match. These outcomes define the Bradley–Terry likelihood; the fitted scores yield a total ordering used to define the borderline band.

Prompt and parsing. Each comparison uses a structured prompt with two papers and asks the LLM to choose one. The prompt includes title, abstract, figure/table captions (when available), and main text, and requests a JSON output with a single chosen paper. We parse the JSON field ("paper_1" or "paper_2") and ignore all other content.

Why pairwise comparisons. Pairwise comparisons are attractive in this setting because comparative judgments are less sensitive to scale drift and calibration than pointwise scores, and they can be aggregated into a total order with standard models such as Bradley-Terry. This makes the approach robust to model upgrades and prompt changes, provided the pairwise preferences are reasonably consistent.

PDF extraction. We extract text from PDFs with ScienceBeam, following Zhang et al. (2025). We truncate inputs to the first $P_{\max} = 10$ pages before parsing to reduce length bias and cost.

3 Related Work

Reviewer assignment and conference operations

Large conferences rely on automated reviewer assignment based on reviewer bids, text similarity, subject areas, and conflict constraints, typically solved as an optimization problem under load and coverage constraints (Charlin et al., 2012; Charlin and Zemel, 2013; Liu et al., 2014; Leyton-Brown et al., 2022; 2024). Fairness- and accuracy-aware objectives have been formalized for assignment, including max–min fairness and statistical accuracy guarantees (Stelmakh et al., 2019). Randomized assignment has been proposed to mitigate manipulation, preserve anonymity, and enable counterfactual evaluation of matching policies (Jecmen et al., 2020; Xu et al., 2023; Saveski et al., 2023). Large Conference Matching (LCM) integrates data aggregation, constrained optimization, and a two-phase reviewing process (often described as a 2+2 scheme) that shifts reviewer resources toward papers near the decision boundary after initial human reviews are in (Leyton-Brown et al., 2022; 2024). Our proposal is complementary but differs in timing: we allocate marginal reviews *before* any human has read the papers, using LLM comparative ranking to pre-identify the likely borderline band once a minimum coverage target is met. More broadly, prior allocation methods are typically framed around meeting a minimum reviewer count and optimizing assignment quality, but they are less explicit about how to deploy *marginal* reviewers when total capacity exceeds the minimum. For example, classic assignment systems such as the Toronto Paper Matching System formalize coverage and load constraints but do not, to our knowledge, specify a dedicated policy for surplus capacity beyond those constraints (Charlin and Zemel, 2013). In practice, surplus capacity is typically absorbed by load-balancing and affinity-based assignment objectives (Charlin and Zemel, 2013; Liu et al., 2014), whereas our focus is to explicitly target that surplus to the likely borderline region.

Public review corpora provide evidence that mean review counts often exceed three, implying a marginal surplus of review slots beyond a three-review baseline. RottenReviews reports average reviews per paper of 4.47 (NeurIPS 2024) and 3.86 (ICLR 2024) (Ebrahimi et al., 2025), and CiteTracked reports NeurIPS averages above three across 2013–2018, with some years substantially higher (Plank and van Dalen, 2019). The ICML 2023 ranking experiment reports 3.08 reviews per submission pre-rebuttal and 3.29 post-rebuttal for its ranked-submission subset (Su et al., 2025). These figures vary by venue, year, and subset, but they indicate that marginal review capacity exists in practice, which creates an allocation question about where to place those extra reviews.

Decision variability and the marginal value of extra reviews

Peer-review outcomes exhibit substantial variability and decision noise near the acceptance boundary, as shown by analyses of NeurIPS 2016 (NIPS 2016) and related Bayesian estimates of arbitrariness (Shah et al., 2018; Francois, 2015). Mechanism-design and calibration perspectives further highlight how noisy ratings and strategic behavior can persist under structured review policies (Lu and Kong, 2023; Srinivasan and Morgenstern, 2023). Recent work leverages author-provided rankings to calibrate scores and improve decision reliability, including the Isotonic Mechanism and the ICML 2023 ranking experiment (Su, 2021; Su et al., 2025). These findings motivate explicitly modeling the marginal benefit of additional reviewers rather than treating review capacity as fixed.

LLMs for reviewing and automated quality estimation

Most LLM-focused work treats LLMs as substitutes or assistants for human reviewers, evaluating their ability to generate reviews, predict scores with calibrated uncertainty estimates Epstein et al. (2026), or critique papers (Liu and Shah, 2023; Robertson, 2023; Idahl and Ahmadi, 2025; Zhu et al., 2025; Zhou et al., 2024; Liang et al., 2023; Zhang and Abernethy, 2025). Recent conference pilots and reports, including a AAAI program, emphasize AI assistance that provides factual review content without scores, leaving accept or reject decisions entirely to humans (AAAI, 2025). Our proposal is complementary and could be deployed alongside such systems, since we use LLMs only to target marginal reviewer allocation rather than to generate reviews. Risk analyses emphasize susceptibility to manipulation, bias, and unreliable judgments under long-context or incomplete inputs (Ye et al., 2024; Akella et al., 2025). Surveys synthesize the growing literature on automated scholarly paper review and its limitations (Zhuang et al., 2025). Parallel lines of work explore pairwise or debiased quality estimation, including LLM-based pairwise comparisons and scalable pairwise training with pointwise inference (Zhang et al., 2025; Zhao et al., 2025). NAIPv2 in particular trains on pairwise preferences but performs pointwise inference at deployment, which could reduce latency and cost compared to full pairwise aggregation, and we do not yet compare against that alternative.

Gap: allocating human review resources with LLM signal

Despite extensive work on assignment and on LLMs as reviewers, there is limited focus on using LLM-derived comparative signals to steer the allocation of human review effort. Existing pipelines such as LCM shift resources toward borderline papers but do not leverage LLM comparative ranking to identify those papers (Leyton-Brown et al., 2022; 2024). Our work targets this gap by using LLM pairwise ranking to estimate the borderline band and by quantifying the marginal value of reallocating human reviews to that band.

4 Expected Net Improved Paper Decisions

Assume N submissions and a mean surplus of s reviews per paper beyond the minimum $r_{\min}$, so the venue has sN "+1 review" slots available. If each paper can receive at most one additional review, then at most sN papers can be upgraded from $r_{\min}$ to $r_{\min} + 1$; we choose these papers as a band of width $w = s$ centered at the expected acceptance percentile.

Let ρ denote the overlap (precision) of the LLM-defined band: the fraction of papers in the LLM-selected band that fall in the true borderline region under the venue's eventual outcomes. Then the expected number of extra reviews that land on truly borderline papers is ρsN. Under a random baseline that selects sN papers uniformly, the expected overlap with a true borderline set of size sN is s^2N papers.

Let δ_B denote the marginal flip rate for borderline papers and let $\delta_{\neg B}$ denote the marginal flip rate for non-borderline papers. Define $\Delta := \delta_B - \delta_{\neg B}$ as the incremental decision-reliability gain from allocating an extra review to a borderline paper rather than a non-borderline paper (measured via a flip-sensitivity proxy in §5, and ideally via randomized estimation in a pilot). The expected number of net improved paper decisions is therefore $(\rho s - s^2)N\Delta$. This expression is a first-order accounting: it translates overlap quality (ρ) and marginal review value (Δ) into expected improvements under a fixed extra-review budget. It abstracts away heterogeneity across papers, reviewer calibration differences, and topic-dependent variance; those effects would naturally lead to a distribution of gains rather than a single scalar. We therefore interpret the formula as a transparent scale estimate rather than a precise causal prediction.

Using ICLR 2025 retrospective estimates (see §5) with $\rho = 0.41$ and $\Delta = 0.024$, and taking $N = 30{,}000$ and $s = 0.3$, the expected net improved paper decisions are about 24 corrected decisions.

5 Estimating Parameters

We study a random sample of 1,000 ICLR 2025 submissions. We run 40 rounds of random pairwise comparisons and fit a Bradley–Terry model. This yields 40 rounds × (1000/2) matches per round, for a total of 20,000 pairwise battles. For efficiency, we truncate each paper to at most the first 10 pages before LLM extraction. In this 1,000-paper setting, the full LLM-based ranking cost about $120. We will release the paper list, extracted text, and LLM pairwise responses used in these experiments.

5.1 Estimating the borderline overlap fraction ρ

We estimate ρ retrospectively on ICLR 2025 submissions with public outcomes. We report the API model identifier logged during the runs (GPT-5-mini). We run pairwise LLM comparisons (40 rounds over 1000 papers using GPT-5-mini) and fit a Bradley–Terry model to obtain an LLM ranking. We define a proxy "human ordering" by decision tier (Reject < Accept < Spotlight < Oral) and mean reviewer score within tier, acknowledging that this operationalizes the conference process rather than ground-truth quality. This operationalizes the borderline set using observed outcomes and scores, so it reflects the review process rather than a latent ground-truth threshold. We then define the borderline band as a quantile window centered at the borderline center (acceptance percentile). Let c denote the borderline center (acceptance percentile) and let w denote the borderline band fraction. The band spans the quantile interval $[c - w/2,\ c + w/2]$. In our experiments we use $c = 1 - 0.25$ and $w = 0.3$. We compute ρ as the borderline overlap fraction of the LLM borderline set with respect to the human borderline set (overlap divided by the size of the LLM set). Because we enforce $|B_{\text{LLM}}| = |B_{\text{human}}|$, this overlap is equal to both precision and recall; we report it as ρ for simplicity.

$$\rho = \frac{|B_{\text{LLM}} \cap B_{\text{human}}|}{|B_{\text{LLM}}|}.$$

In practice, ACs and SACs may adjust assignments after the initial match; those adjustments can be applied on top of a targeted allocation policy. We therefore compare the LLM-defined band to a random baseline (selecting sN papers uniformly), which provides a neutral reference point for marginal review placement in the absence of any targeted policy.

5.2 Estimating the marginal review effect Δ

We estimate Δ on the same 1000-paper ICLR 2025 sample, restricted to papers with ≥ 4 reviews. We use the same human ranking and borderline band as above. For a paper with scores $s_1, \ldots, s_k$, we compute the mean μ and variance σ^2 of the k scores. We fit a logistic model $p = \text{sigmoid}(\beta_0 + \beta_1\mu + \beta_2\sigma^2)$ on observed decisions, where $p = P(\text{accept} \mid \mu, \sigma^2)$. This calibrated flip model is an operational proxy and does not identify a causal effect of adding a review; a randomized extra-review design could estimate Δ causally. For each review i, we form the leave-one-out statistics $(\mu_{-i}, \sigma^2_{-i})$ and compute p_{-i}. A *flip* occurs if the accept indicator changes after removing a review: $\mathbf{1}\{p \geq 0.5\} \neq \mathbf{1}\{p_{-i} \geq 0.5\}$. We aggregate flips over all leave-one-out trials within the borderline set and within the non-borderline set, and define Δ as the difference in flip rates. Let δ_B denote the

ρ	Δ	Expected net improved paper decisions
0.41	0.024	24

Table 1: ICLR 2025 retrospective estimates from 1000 papers (40 rounds, GPT-5-mini); borderline band fraction 30% around the acceptance percentile center; Δ computed on papers with ≥ 4 reviews using the calibrated flip model. Expected net improved paper decisions are computed as $(\rho s - s^2)N\Delta$ with $N = 30{,}000$ and $s = 0.3$.

flip rate within the borderline set and let $\delta_{\neg B}$ denote the flip rate outside the borderline set.

$$\begin{aligned} \Delta &= \delta_B - \delta_{\neg B}, \\ \delta_B &= \frac{\#\text{ flips in } B}{\#\text{ trials in } B}, \\ \delta_{\neg B} &= \frac{\#\text{ flips in } \neg B}{\#\text{ trials in } \neg B}. \end{aligned}$$

We report a Wald 95% confidence interval for Δ using a difference-in-proportions standard error,

$$\begin{aligned} \mathrm{SE}(\Delta) &= \sqrt{\frac{\delta_B(1-\delta_B)}{n_B} + \frac{\delta_{\neg B}(1-\delta_{\neg B})}{n_{\neg B}}}, \\ \mathrm{CI}_{0.95}(\Delta) &= \Delta \pm 1.96\,\mathrm{SE}(\Delta), \end{aligned}$$

where n_B and $n_{\neg B}$ are the numbers of leave-one-out trials in the borderline and non-borderline sets. This yields an average Δ and likely masks heterogeneity across papers with different score variance and reviewer calibration.

Robustness and falsifiability. These estimates are retrospective proxies, so we report ablations and confidence intervals to assess stability. If ρ were close to the random baseline or if Δ were indistinguishable from zero on the 1,000-paper sample, the policy would have little expected benefit and would not justify deployment. Conversely, persistent separation between the LLM ranking and the random baseline, together with a positive Δ, supports the position that marginal reviews should be targeted rather than spread uniformly.

6 Empirical Validation

The practical usefulness of our position hinges on empirical support: how many paper decisions are improved in expectation, and how accurately LLMs identify borderline papers. We therefore treat this section as empirical validation aimed at understanding how impactful the position is in terms of expected net improved paper decisions, rather than as standalone algorithmic results. Table 1 summarizes the retrospective estimates from ICLR 2025 (1000-paper sample, 40 rounds with GPT-5-mini, 30% borderline band fraction). The expected net improved paper decisions are computed as $(\rho s - s^2)N\Delta$ with $N = 30{,}000$ and $s = 0.3$.

6.1 Ablations

Using the Wald interval from Section 5, we obtain $\Delta = 0.024$ with 95% CI $[0.003, 0.045]$. We ablate the borderline band fraction and centering using cached LLM rankings and the calibrated Δ estimator. For the centering ablation, we report 95% confidence intervals for ρ under a binomial overlap model with $m = |B_h|$ and $K = |B_h \cap B_{\mathrm{LLM}}|$. We use the Wald interval

$$\hat{\rho} = \frac{K}{m}, \qquad \mathrm{CI}_{0.95}(\rho) = \hat{\rho} \pm 1.96\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{m}}. \tag{1}$$

For the band-fraction ablation, we set the borderline band fraction to match the marginal reviewer fraction s and recompute ρ or Δ under the new band definition. We report all ablations on the 1000-paper sample with 40 rounds. We summarize the band-fraction and centering sensitivity of expected net improved paper decisions and ρ in Figures 2a–3a, including the random-baseline reference.

Rounds	Papers	n_+	n_-	AUC
40	1000	366	634	0.708
30	1000	366	634	0.705
20	1000	366	634	0.701
10	1000	366	634	0.686
5	1000	366	634	0.652

Table 2: Mann–Whitney AUC for the LLM ranking against binary accept labels. The table reports n_+ accepted papers and n_- rejected papers in each sample.

Figure 3b reports the corresponding sensitivity of Δ under the calibrated flip rule. In Figure 2a, expected net improved paper decisions increase with the marginal reviewer fraction; at $s = 0.5$ this yields roughly 50 improved decisions in our setting. Figure 2b shows that across a wide range of marginal reviewer fractions, the borderline overlap fraction remains statistically above the random baseline, indicating that the LLM ranking is informative for identifying borderline papers. Figure 3a shows higher overlap as the borderline center moves to higher acceptance percentiles, consistent with the LLM being better at identifying top papers than separating near-cutoff papers. Figure 3b indicates that the marginal value of an extra review is positive across a wide range of marginal reviewer fractions.

Summary. Across ablations, the LLM ranking consistently exceeds the random baseline in overlap, yielding positive expected net improved paper decisions under the allocation rule. This suggests that the policy is not overly sensitive to a single hyperparameter choice, which is important for operational deployment.

6.2 Ranking AUC

Mann–Whitney AUC. We also report a global ranking metric based on the Mann–Whitney AUC between LLM scores and binary accept labels. Let θ_i be the LLM score for paper i, and let $y_i \in \{0, 1\}$ indicate accept vs. reject. Let n_+ be the number of accepted papers and n_- be the number of rejected papers in the sample. The AUC is

$$\mathrm{AUC} = \Pr(s_{i^+} > s_{i^-}) + \tfrac{1}{2} \Pr(s_{i^+} = s_{i^-}),$$

which is the probability that a randomly chosen accepted paper ranks above a randomly chosen rejected paper, with ties split evenly. We estimate it from the finite sample by

$$\widehat{\mathrm{AUC}} = \frac{1}{n_+ n_-} \sum_{i:y_i=1} \sum_{j:y_j=0} \Big[\mathbb{I}(\theta_i > \theta_j) + \tfrac{1}{2}\mathbb{I}(\theta_i = \theta_j) \Big].$$

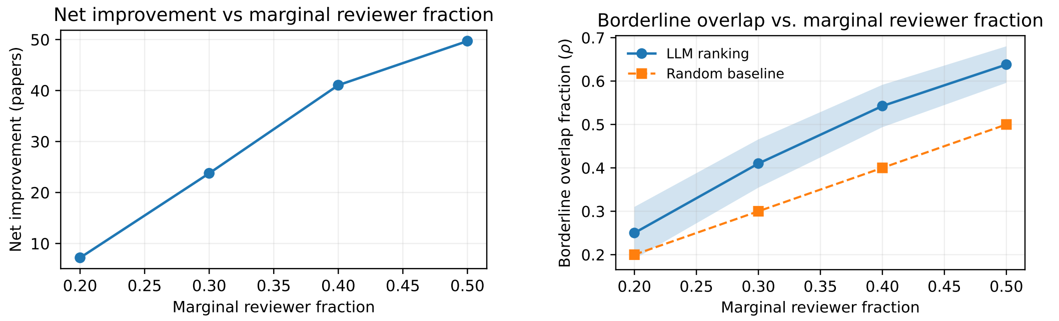
Table 2 reports the Mann–Whitney AUC for the 1,000-paper run (random baseline AUC $= 0.5$) and shows diminishing returns as the number of rounds increases.

We present ablations on model capability as well as full-paper vs abstract comparisons in the Appendix.

7 Discussion

Gaming. Authors might try to steer their paper toward the borderline band to attract an extra review: If the paper is originally good, extra review reduces noise, thereby increasing acceptance chance. An originally not as good paper can also reduce its quality to steer away from getting an extra reviewer. However, we suspect that this gaming is impossible in practice since it requires knowledge about others' qualities.[1] LLM ranking is comparative and noisy, further reducing the incentive of downgrading paper (See figure 4). We also recommend allocating only a fraction of extra reviews via the LLM signal, with the remainder assigned uniformly at random as a decoy, so that receiving a fourth review is not a clear signal of borderline status.

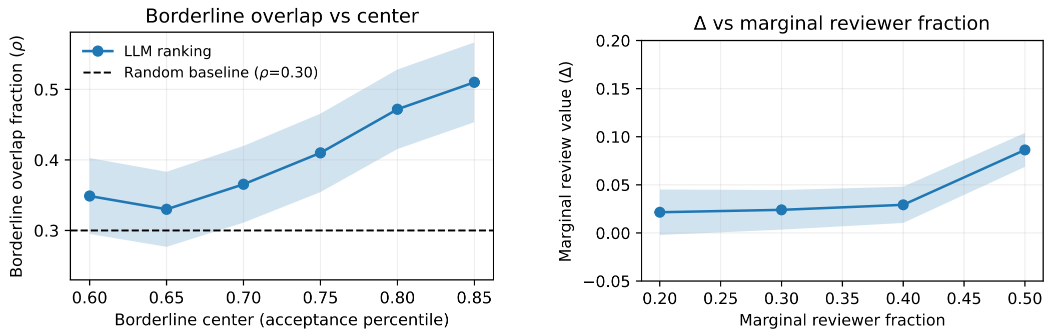
[1] This is usually feasible only with continuous observation Edelman et al. (2007) or continuum of agents with a known distribution Sornwanee (2025). Both conditions are not met in a practical peer review setting.



(a) Expected net improved paper decisions sensitivity to the marginal reviewer fraction. We vary the marginal reviewer fraction and set the borderline band fraction to match it, then recompute ρ and the expected net improved paper decisions under the review reallocation. Expected net improved paper decisions are shown relative to random allocation, using $(\rho s - s^2)N\Delta$ with s fixed.

(b) Borderline overlap fraction as a function of the marginal reviewer fraction. The dashed series shows the random-baseline ρ implied by each band fraction, and shaded bands show the 95% Wald confidence interval for ρ.

Figure 2: Sensitivity to marginal reviewer fraction.



(a) Centering ablation for the borderline center (acceptance percentile). We shift the band center and recompute ρ; the horizontal line shows the random-baseline ρ implied by the 30% band fraction. Shaded bands show the 95% Wald confidence interval for ρ under the binomial overlap model.

(b) Marginal review value sensitivity to the marginal reviewer fraction. We vary the marginal reviewer fraction (and set the borderline band fraction to match it) and recompute Δ from leave-one-review counterfactuals under the calibrated flip rule.

Figure 3: Sensitivity to centering and marginal review value.

8 Alternative Views

Decision accuracy is not the right objective. Some may argue that the objective should be to maximize the chance that the very best papers are accepted. That would favor allocating extra reviews to the top-ranked papers rather than to the borderline band. Others may argue the opposite, that weaker papers deserve more reviews to receive helpful feedback. We see these as different objectives at different stages. Identifying the very best papers can be prioritized later, for example when area chairs and senior area chairs deliberate on awards, talks, or top-presentation slots, which typically occur after accept or reject outcomes are fixed and need not respect tight resubmission timelines. By contrast, accept or reject decisions must be timely, and they are most sensitive near the boundary, so reallocating marginal reviews there yields the highest expected improvement in correct decisions. On the low end, additional reviews may be less informative for outcomes and can be redundant

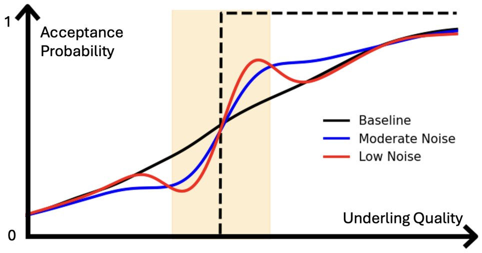


Figure 4: Toy Example of Acceptance Probability as a function of Underlying Quality: The black dotted line represents the switch from 0 probability to 1 probability of acceptance when the quality exceeds a certain threshold (target quantile). This is the first best behavior that could only be achieved when the underlying quality is known to us. Reviewer feedback is a noisy signal of the quality, leading to the black curve. Although having quality exceeding the threshold may no longer guarantee acceptance, higher quality leads to higher acceptance probability. The red curve represents the acceptance probability under our scheme when LLM has extremely high fidelity: if an author has an option to reduce one's quality, they could intentionally reduce their quality to increase the acceptance probability. However, we can see that, when an LLM has moderate noise, the acceptance probability follows the blue curve, which is monotone, rendering gaming impossible.

when multiple reviewers already agree; structured feedback and mentoring can be provided without changing the review-count allocation. Our position is narrower: given a fixed extra-review budget for acceptance decisions, we target the region where marginal reviews are most likely to change accept or reject outcomes.

Avoid LLMs in reviewing. Concerns about LLM review reliability are grounded in recent evaluations and surveys (Akella et al., 2025; Zhou et al., 2024; Zhuang et al., 2025). Another view is that any LLM involvement introduces unacceptable bias. Recent work reports systematic differences in LLM rankings across paper categories, which could skew where extra reviews are placed. This concern is strongest when LLMs are used to make accept or reject decisions, because any bias can directly change outcomes and creates incentives for authors to optimize for the prompt rather than scientific impact. Our design uses LLMs only to allocate review *counts*, keeps the signal hidden from decision-makers, and preserves fully human outcomes. As a result, any bias mainly affects where additional scrutiny is applied, not the decision rule itself, and the incentive to game the LLM is sharply reduced. We also recommend allocating only a fraction of extra reviews via the LLM signal, with the remainder assigned uniformly at random as a decoy, so that four reviews do not reliably indicate a paper is borderline. These mitigations do not eliminate bias risk, but they reduce the chance that LLM outputs systematically tilt final decisions.

Time and cost. One could argue that improving a modest number of decisions is not worth the cost of running an LLM ranking. In our 1,000-paper run, the full ranking cost about $120. Scaling linearly to a NeurIPS-scale full run ($N = 30{,}000$) yields about $3,600 before batch discounts (about $1,800 with a 50% discount), holding prompt lengths and pricing fixed. Dividing by our expected net improved paper decisions (Section 6) yields an implied cost per improved decision of about $75 (using 1,800 and 24). Given the total time invested by expert reviewers and the reputational cost of errors at scale, this cost may be reasonable. Operationally, the ranking can run in parallel with reviewer bidding, so it need not add elapsed time to the review cycle.

9 Conclusion

We argue that reallocating spare reviewer capacity toward a statistically defined borderline band can improve decision quality without changing reviewer load. Our empirical estimates of the borderline overlap fraction ρ and the marginal review value Δ, together with ablations, indicate that the gains are robust to reasonable variations in band fraction and centering. More broadly, this position advocates using LLMs to guide *allocation* rather than *judgment*: the goal is to focus scarce human attention where it is most likely to change outcomes while keeping final decisions entirely human. Operationally, this is a minimal change to current reviewing workflows; with explicit guardrails, it carries low risk of bias amplification and low incentives for gaming.

References

AAAI. Association for the advancement of artificial intelligence launches ai-powered peer review assessment system. Press release, 2025. URL https://aaai.org/wp-content/uploads/2025/05/AAAI-LLM-Press-Release.pdf. Accessed 2026-01-28.

Akhil Pandey Akella, Harish Varma Siravuri, and Shaurya Rohatgi. Pre-review to peer review: Pitfalls of automating reviews using large language models, 2025. URL https://arxiv.org/abs/2512.22145.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Laurent Charlin and Richard S. Zemel. The toronto paper matching system: An automated paper-reviewer assignment system, 2013. URL https://api.semanticscholar.org/CorpusID:680003.

Laurent Charlin, Richard S. Zemel, and Craig Boutilier. A framework for optimizing paper matching, 2012. URL https://arxiv.org/abs/1202.3706.

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

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259, 2007.

Elliot L Epstein, John Winnicki, Thanawat Sornwanee, and Rajat Vadiraj Dwaraknath. LLMs are overconfident: Evaluating confidence interval calibration with fermieval. In *AAAI 2026 Workshop on Assessing and Improving Reliability of Foundation Models in the Real World*, 2026. URL https://openreview.net/forum?id=yUyFITL0wv.

Olivier Francois. Arbitrariness of peer review: A bayesian analysis of the nips experiment, 2015. URL https://arxiv.org/abs/1507.06411.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews, 2025. URL https://arxiv.org/abs/2412.11948.

Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12533–12545. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/93fb39474c51b8a82a68413e2a5ae17a-Paper.pdf.

Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints, 2019. URL https://arxiv.org/abs/1905.11924.

Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. Matching papers and reviewers at large conferences, 2022. URL https://arxiv.org/abs/2202.12273.

Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. Matching papers and reviewers at large conferences. *Artificial Intelligence*, 331: 104119, 2024. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2024.104119. URL https://www.sciencedirect.com/science/article/pii/S0004370224000559.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023. URL https://arxiv.org/abs/2310.01783.

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

Xiang Liu, Torsten Suel, and Nasir Memon. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 25–32, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645749. URL https://doi.org/10.1145/2645710.2645749.

Yuxuan Lu and Yuqing Kong. Calibrating "cheap signals" in peer review without a prior. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xr3KAzboHY.

Barbara Plank and Reinard van Dalen. Citetracked: A longitudinal dataset of peer reviews and citations. In *BIRNDL@SIGIR*, 2019. URL https://api.semanticscholar.org/CorpusID:198489688.

Zachary Robertson. Gpt4 is slightly helpful for peer-review assistance: A pilot study, 2023. URL https://arxiv.org/abs/2307.05492.

Martin Saveski, Steven Jecmen, Nihar B Shah, and Johan Ugander. Counterfactual evaluation of peer-review assignment policies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rhIfzCZoXG.

Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike von Luxburg. Design and analysis of the nips 2016 review process, 2018. URL https://arxiv.org/abs/1708.09794.

Thanawat Sornwanee. 1-dimensional normal competitive market equilibrium. 2025. URL https://arxiv.org/abs/2505.08425.

Siddarth Srinivasan and Jamie Morgenstern. Auctions and peer prediction for academic peer review, 2023. URL https://arxiv.org/abs/2109.00923.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. Peerreview4all: Fair and accurate reviewer assignment in peer review, 2019. URL https://arxiv.org/abs/1806.06237.

Buxin Su, Jiayao Zhang, Natalie Collina, Yuling Yan, Didong Li, Kyunghyun Cho, Jianqing Fan, Aaron Roth, and Weijie Su. The icml 2023 ranking experiment: Examining author self-assessment in ml/ai peer review. *Journal of the American Statistical Association*, 0(0):1–12, 2025. doi: 10.1080/01621459.2025.2510006. URL https://doi.org/10.1080/01621459.2025.2510006.

Weijie J Su. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=xmx5rE9QP7R.

Yixuan Even Xu, Steven Jecmen, Zimeng Song, and Fei Fang. A one-size-fits-all approach to improving randomness in paper assignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=D94QKZA7UP.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng 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.

Tianmai M. Zhang and Neil F. Abernethy. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation, 2025. URL https://arxiv.org/abs/2505.23824.

Yaohui Zhang, Haijing ZHANG, Wenlong Ji, Tianyu Hua, Nick Haber, Hancheng Cao, and Weixin Liang. From replication to redesign: Exploring pairwise comparisons for LLM-based peer review. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=z5KTxW5sJd`.

Penghai Zhao, Jinyu Tian, Qinghua Xing, Xin Zhang, Zheng Li, Jianjun Qian, Ming-Ming Cheng, and Xiang Li. Naipv2: Debiased pairwise learning for efficient paper quality estimation, 2025. URL `https://arxiv.org/abs/2509.25179`.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen 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/`.

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

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. Large language models for automated scholarly paper review: A survey. *Inf. Fusion*, 124(C), December 2025. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.103332. URL `https://doi.org/10.1016/j.inffus.2025.103332`.

A Additional Empirical Validation

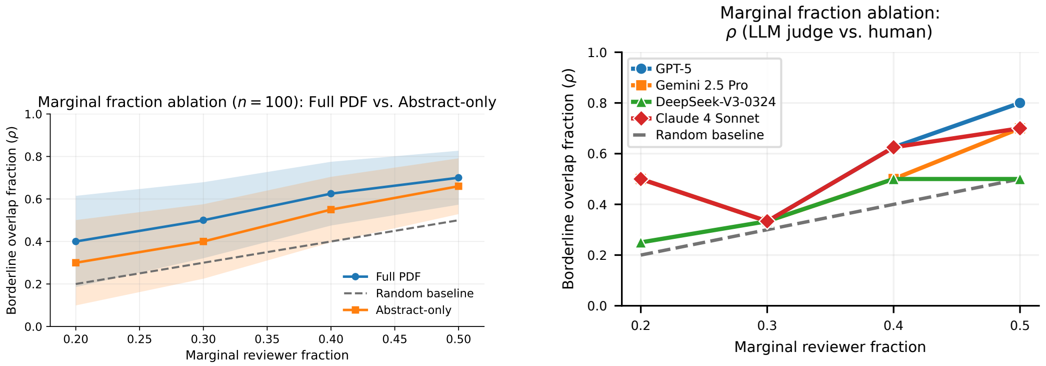
A.1 Full paper vs. abstract-only inputs.

We ablate how much paper content the LLM judge sees by running the marginal-fraction ablation with either the full PDF text or only the abstract across 100 papers with 5 rounds. Figure 5a shows that both settings improve with larger marginal fractions and both outperform the random baseline, but access to the full paper yields consistently higher agreement with humans. Concretely, ρ increases from ≈ 0.40 to ≈ 0.70 (full PDF) as the marginal reviewer fraction rises from 0.2 to 0.5, while abstract-only rises from ≈ 0.30 to ≈ 0.66, leaving a persistent (though narrowing) absolute gap of roughly 0.04–0.10.

This ablation highlights a cost–accuracy tradeoff. Abstracts contain a large share of the signal needed for pairwise comparisons, but the full text provides additional evidence that measurably improves fidelity, especially at smaller marginal fractions. From a practical standpoint, this matters because (as discussed in our Alternative Views cost analysis) total dollar cost and latency are dominated by input tokens, and full-PDF comparisons are materially more expensive than abstract-only prompts. A natural compromise is a staged pipeline: run abstract-only comparisons broadly, and reserve full-PDF comparisons for papers near the estimated borderline where the extra signal is most valuable.

A.2 Model capability ablation.

We ablate the choice of LLM judge and measure agreement with human judgments using the borderline overlap fraction ρ (Figure 5b) in a small 20-paper, 30-round run. Across the marginal fractions shown, the higher-capability judges tend to achieve higher agreement, with differences most apparent at larger marginal reviewer fractions. At the largest fraction we evaluate (0.5), GPT-5 reaches $\rho \approx 0.8$, Gemini 2.5 Pro and Claude 4 Sonnet are around $\rho \approx 0.7$, while DeepSeek-V3-0324 is closer to the random baseline trend (around $\rho \approx 0.5$). At smaller marginal fractions the curves are closer together and not strictly monotonic, which is plausible given the small scale of this test.



(a) Marginal fraction ablation ($n = 100$): full PDF vs. abstract-only. We report agreement ρ between the LLM judge and human judgments as a function of the marginal reviewer fraction when the judge compares either the full paper text or only the abstract; the dashed line shows the random baseline (shaded bands indicate uncertainty).

(b) Model capability ablation. Borderline overlap fraction ρ (LLM judge vs. human) as a function of marginal reviewer fraction. In this small run, higher-capability judges tend to achieve higher agreement, while weaker judges track closer to the random baseline (dashed).

Figure 5: Ablations on information access and judge capability.

Overall, this ablation is intended as a lightweight sanity check (not a definitive ranking), but it supports the practical takeaway that the choice of judge can meaningfully affect how closely the evaluation tracks human decisions, especially when the marginal fraction is larger.