# Pairwise Evaluation at Scale: The Plateau Effect and Efficient Tournament Design

Anonymous ACL submission

## Abstract

Pairwise comparison has become a foundational paradigm for evaluating open-ended systems, underpinning reinforcement learning from feedback, evaluation benchmarks, and widely used LLM leaderboards. Its appeal lies in its robustness over absolute scoring, yet as model capabilities improve and candidate pools grow, evaluation becomes combinatorial: exhaustive comparison scales quadratically, while sparse strategies introduce structural bias. Despite its widespread adoption, the role of comparison structure in ranking quality and stability remains largely under-explored. We recast pairwise evaluation as a budget-constrained problem and propose tournament-based comparison as a principled framework for allocating limited comparisons. Experiments on human-annotated datasets for essay scoring and machine translation reveal a consistent plateau effect: ranking quality saturates after a small fraction of the full comparison budget. Swiss-style tournaments recover most of the accuracy of round-robin at a fraction of the cost, while elimination-based formats exhibit significant instability, demonstrating that tournament structure, not just the ranking model, plays a central role in both accuracy and reproducibility.

## 1 Introduction

The ability to reliably compare and rank system outputs is central to modern machine learning, underpinning applications ranging from reinforcement learning from feedback to evaluation benchmarks and public model leaderboards. In many of these settings, absolute scoring is unreliable: assigning consistent numeric scores to open-ended outputs is difficult for both human annotators and automated judges. As a result, pairwise comparison, i.e., asking which of two outputs is better, has emerged as a dominant paradigm for evaluation, and is now widely used in systems such as Chatbot Arena and AlpacaEval (Devriesere et al., 2025; Dubois et al., 2024; Zheng et al., 2023). Pairwise judgments are typically more consistent and easier to elicit, and can be aggregated into global rankings using models such as Bradley–Terry (Bradley and Terry, 1952).

However, evaluation in practice is fundamentally constrained by cost and scalability. Collecting pairwise comparisons incurs real expense, whether through human annotation or model inference, and this cost scales directly with the number of comparisons performed. As candidate pools grow, exhaustive round-robin evaluation requires $O(N^2)$ comparisons, making it prohibitively expensive at scale (Xu et al., 2025; Liusie et al., 2024). These constraints are particularly acute when evaluation relies on frontier model APIs, where cost scales with the number of calls, making it critical to minimise unnecessary comparisons. More broadly, real-world evaluation pipelines often favor fast and inexpensive proxies over more accurate but costly methods, which explains the continued reliance on automatic metrics such as BLEU during development, as they are faster and cheaper than more reliable but expensive alternatives (Freitag et al., 2022; Papineni et al., 2002).

Under these constraints, pairwise evaluation becomes a design problem: which comparisons should be performed? Simple strategies such as round-robin maximise coverage but are computationally prohibitive, while sparse structures such as single-elimination reduce cost but introduce structural biases, making rankings sensitive to initial pairings and bracket effects (Devriesere et al., 2025). Despite the central role of pairwise comparison in modern evaluation pipelines, the impact of comparison structure on ranking quality and stability remains largely underexplored.

While tournament structures provide a natural mechanism for allocating comparisons under limited budgets, they are primarily studied in domains such as chess and competitive sports, where strong priors over participant skill enable effective seeding and stabilise outcomes (Engist et al., 2021;

Schwenk, 2000; Appleton, 1995). In contrast, large-scale model evaluation typically operates without such priors, requiring tournaments to begin from random initialisation. As a result, the structure of the tournament itself becomes a critical factor: formats that are sensitive to early pairings can produce substantially different rankings across runs, effectively behaving as stochastic processes rather than reliable evaluation procedures. These challenges are further amplified by known pathologies in pairwise judging. As the quality gap between candidates narrows, judges exhibit increased non-transitivity and position bias, making comparisons noisier and less consistent (Xu et al., 2025). This low-signal regime is precisely where ranking is most important, such as comparisons between closely performing models near the top of a leaderboard, yet also where evaluation is most fragile. Together, these factors make comparison structure a critical but underexplored determinant of ranking accuracy and stability.

In this work, we recast pairwise evaluation as a budget-constrained comparison allocation problem and study how the structure of the induced comparison graph affects both ranking accuracy and reproducibility. We take a tournament-based perspective, where different formats correspond to different strategies for allocating limited comparisons, and systematically evaluate how these structures trade off efficiency, discrimination, and stability. We conduct a comprehensive empirical study across multiple evaluation settings, including rubric-based essay scoring and machine translation, using human-annotated datasets to construct full preference graphs as reference rankings. This allows us to measure how accurately each method recovers the underlying ranking under restricted comparison budgets, while also quantifying sensitivity to stochastic factors such as initial pairing and randomisation.

Our results reveal three findings: (i) pairwise evaluation exhibits a plateau effect: most ranking quality is recovered after only a small fraction of the full comparison budget, with diminishing returns thereafter; (ii) tournament structure plays a critical role when signal is weak but non-zero. In this regime, elimination-based formats become highly unstable and fail to recover the underlying ranking, while Swiss-style formats more reliably extract the available signal and maintain reproducibility; (iii) the practical aspect depends not only on accuracy but on execution constraints: Swiss-style tournaments recover most of the benefit of round-robin evaluation at much lower cost, and unlike insertion-based methods such as SWIM (Xu et al., 2025), remain naturally parallelisable. Together, these findings show that comparison structure is a primary driver of both ranking quality and reproducibility, establishing comparison allocation as a key design dimension in large-scale evaluation. Our contributions are as follows:

- **Comparison allocation:** Pairwise evaluation is recast as a budget-constrained allocation problem, where tournament structure determines ranking accuracy and stability.
- **Plateau effect:** Most ranking quality is recovered with a small fraction of comparisons, with diminishing returns thereafter.
- **Practical guidance:** Swiss-style tournaments provide the best trade-off between cost, accuracy, and stability, and remain efficient under real-world computational constraints.

## 2 Related Works

**Pairwise Evaluation and Tournament Design.** The organisation of pairwise evaluation as a tournament is an optimisation problem determined by the competing criteria of efficacy, fairness, and attractiveness (Devriesere et al., 2025). In a candidate pool of size $N$, exhaustive round-robin (where every pair of candidates is compared) evaluation requires $\frac{N(N-1)}{2}$ pairwise comparisons, a quadratic cost that becomes prohibitive at scale. This has motivated research into sparser tournament structures that seek to preserve ranking fidelity at substantially reduced cost. The efficacy of a tournament format is defined as its ability to identify the true latent ranking of candidates using the fewest possible comparisons, a property that is particularly important in settings where match outcomes are noisy and a stronger candidate does not consistently outperform a weaker one (Sziklai et al., 2022). Single-elimination formats (where candidates are eliminated after a single loss) require only $N-1$ matches but produce highly skewed comparison graphs that offer minimal signal for early-eliminated candidates. Recent work has applied single-elimination tournaments to LLM system ranking, demonstrating higher correlation with human Elo ratings[1] than reference-based evaluation while acknowledging that more informative matchmaking could improve upon the format (Son et al., 2025). A complementary line of work addresses budget-constrained ranking through

[1] A rating system that aggregates pairwise wins and losses into a global ranking.

active sampling, where pairs are selected adaptively based on prior comparison outcomes. Chen et al. (2013) propose Crowd-BT, jointly selecting which pair and annotator to query, while Mikhailiuk et al. (2021) extend this to subjective settings via information gain maximisation, reducing comparisons with negligible accuracy loss. However, active sampling introduces a sequential dependency and an indeterminate budget, terminating only upon posterior convergence, which is impractical when comparisons are expensive LLM calls. tournament-style evaluation avoids both issues, fixing all pairings simultaneously with a budget determined in advance.

Adaptive tournament methods have recently emerged as an operationally attractive structure for natural language processing evaluation (Xu et al., 2025). SWIM, a tournament-based ranking method proposed by Xu et al. (2025), is non-eliminating and adaptive: rather than following fixed rounds, candidates are inserted one at a time into a growing pool and compared against opponents of similar estimated quality via Bradley–Terry proximity (a probabilistic model of pairwise preference), concentrating comparisons where informational value is greatest. In contrast, a standard Swiss-style tournament proceeds in discrete rounds where all candidates participate simultaneously, with pairings determined by win-loss records rather than incremental insertion, making it naturally parallelisable within each round. Sziklai et al. (2022) and Sauer et al. (2024) suggest that setting the number of rounds $R$ to approximately $\lceil \log_2 N \rceil$ is sufficient for recovering a reliable global ranking. We analyse SWIM against existing tournament methods and examine how the mechanism performs across varying datasets and scales from modest candidate sets to significantly larger pools, providing a detailed look at its efficiency and stability relative to alternative tournament formats. A related line of work employs tournament-based ranking as a training signal in reinforcement learning pipelines where pointwise rewards are unreliable (Zhang et al., 2026), further motivating the broader shift from absolute scoring towards comparative ranking structures. In the essay scoring setting, Shibata and Miyamura (2025) demonstrate that exhaustive comparison is unnecessary for reliable LLM-based ranking, while the stability of the resulting rankings further depends on the choice of aggregation method (Daynauth et al., 2025; Boubdir et al., 2023), with both identifying failure modes in Elo and Bradley–Terry under LLM-as-a-judge comparisons.

**LLM-as-a-Judge.** These tournament structures operate on pairwise preference judgments, which may be obtained from human annotation or automated evaluation such as LLM-as-a-Judge (Wang et al., 2024). This paradigm is subject to well-documented biases including positional bias, length bias, and self-preference, all of which are most acute when the quality difference between two candidates is marginal. Xu et al. (2025) show that round-robin tournaments combined with Bradley-Terry modelling substantially mitigate the non-transitivity (inconsistent preference cycles) that arises from these biases, improving correlation with human-preference rankings on Chatbot Arena. Tripathi et al. (2025) further demonstrate that pairwise protocols are more susceptible to distracted evaluation than absolute scoring, though Qin et al. (2024) and Chen et al. (2025) show that pairwise comparison retains superior ranking accuracy over pointwise methods across document retrieval and code quality tasks. These findings suggest that the reliability of pairwise evaluation is not a property of the protocol alone but depends critically on how comparisons are structured and aggregated, motivating the systematic study of tournament design as a key variable in LLM evaluation.

Despite these advances, prior work largely treats the comparison process itself as fixed, focusing on aggregation methods or specific tournament designs without systematically analysing how comparison structure affects ranking accuracy and reproducibility under realistic budget constraints. In particular, the effects of stochastic factors such as initial pairing and the behaviour of different structures in low-signal regimes remain under-explored. This work addresses these gaps by treating tournament structure as a first-class design variable, evaluating how different structures trade off efficiency, accuracy, and stability across varying pool sizes and budgets, and quantifying their sensitivity to random initialisation.

## 3 Data

We evaluate on two human-annotated datasets that provide reliable pairwise preference signals across distinct generation tasks. For essay scoring, we use the DREsS dataset (Yoo et al., 2025), comprising EFL essays annotated across content, organisation, and language (`DREsS New` and `DREsS Std`). For machine translation, we use WMT 2022 to 2024

(Kocmi et al., 2024) via the mt-metrics-eval toolkit,[2] evaluating language pairs from en-zh and en-cs independently. We focus on two language pairs to isolate the effect of tournament structure from confounds introduced by cross-lingual variation, selecting pairs that differ in typological distance from English to ensure our findings generalise beyond a single language family. We additionally include a case study on borderline peer review decisions, where score differences are minimal and ranking is particularly challenging (see Appendix H).

**Tournament Structures.** We frame the evaluation as a series of matchmaking tournaments where a large language model acts as the judge, deploying one of the tournament structures to generate the pairwise comparison graph for a pool of size N (increments of 10 up to 100). For machine translation, each pool consists of N candidate system translations of the same source sentence, drawn from the full set of systems available for that language pair; for essay scoring, each pool consists of N essays drawn from the same essay group. To mitigate prompt position bias, every match presents each candidate ordering an equal number of times, with early stopping once a majority is guaranteed. Initial pairings are strictly randomly assigned, stress-testing every format against worst-case initialisation conditions. We sample k=10 distinct data subsets from each corpus, stratified to ensure an evenly distributed range of ground-truth scores. For each subset, we independently repeat the entire tournament process 10 times with different random seeds. Dataset statistics are shown in Appendix B.

## 4 Model Selection

We select an evaluator by screening candidate LLMs for pairwise accuracy above chance ($>50\%$), as tournament-based ranking is unreliable if the judge cannot consistently distinguish between candidates. We balance computational efficiency with reasoning capability across: Qwen3 (8B, 14B) (Team, 2025), Llama 3.1 8B-Instruct (Grattafiori et al., 2024), Gemma-3-12B-it (Kamath et al., 2025), InternLM3-8B (Cai et al., 2024), Ministral 8B-Instruct (Liu et al., 2026), Phi-4 and -mini (Abouelenin et al., 2025), OLMo 3 7B-Instruct (Olmo et al., 2025), and GPT-OSS 20B (Agarwal et al., 2025). Results are reported in Table 5. Phi-4 achieves the highest average pairwise accuracy, while GPT-OSS 20B offers a stronger balance between accuracy and inference speed. We therefore select GPT-OSS 20B as the primary evaluator and replicate core experiments (MT and Essay) with Phi-4 to verify that our findings are not model-specific (see Appendix I). Additionally, we evaluate GPT-5.4-mini[3] as a closed proprietary model to assess budget constraints and generalisability.

## 5 Tournament Format

**Round-Robin (All-Pairs):** An exhaustive, all-pairs evaluation that constructs a complete, dense comparison graph. This format provides the maximum statistical power for global rank but requires quadratic match complexity.

**Single Elimination:** A binary bracket structure, in which candidates are eliminated after a single loss. While minimising the total match count, it remains highly sensitive to early-round stochasticity and "unlucky" pairings (if two high quality candidates meet at the start then one will be eliminated early on).

**Double Elimination:** A dual-bracket structure (Winners and Losers) providing increased robustness. A candidate is only eliminated after two losses, ensuring that a single sub-optimal pair does not terminate a high-quality sample.

**Swiss-System:** A non-eliminating format that pairs candidates in successive rounds against opponents with similar win–loss records, approximating comparisons between similarly ranked items. It maximises the informational entropy of each match while bounding total rounds logarithmically.

**Hybrid Swiss-Elimination:** We evaluate hybrid variants that utilise a Swiss seeding phase to inform a subsequent elimination, specifically Swiss into Single Elimination **(Swiss→SE)** and Swiss into Double Elimination **(Swiss→DE)**. For these variants, the seeding phase uses one fewer round than the standard Swiss configuration, $R_{\text{seed}} = R - 1$, where $R = \lceil \log_2(N) \rceil$. For all Swiss-based formats, we evaluate three round configurations: $R - 1$, $R$, and $R + 1$ to characterise the sensitivity and diminishing returns of additional rounds.

**SWIM:** An adaptive insertion-based format proposed by Xu et al. (2025) that builds the ranking incrementally. Candidates are inserted one at a time into a growing pool; upon insertion, each candidate is compared against $\lceil \log_2(s) \rceil$ opponents selected from the current pool of size $s$, beginning with a

[2] https://github.com/google-research/mt-metrics-eval

[3] gpt-5.4-mini-2026-03-17

random opponent and then targeting candidates closest in estimated quality under a BT model.

All tournament formats produce pairwise match outcomes that are aggregated using the Bradley–Terry model (Bradley and Terry, 1952), which estimates a latent quality score for each candidate via maximum likelihood estimation over the entire match history. The probability of one candidate defeating another is modelled as a logistic function of the difference between their respective latent quality scores. While we adopt BT as our primary aggregation method, we additionally evaluate alternative aggregators including TrueSkill (Minka et al., 2018), Crowd-BT (Chen et al., 2013), and ASAP (Mikhailiuk et al., 2021) to assess whether our findings are sensitive to the choice of aggregation model. A summary of match complexities for each format is provided in Appendix D.

**Evaluation Metrics.** We evaluate ranking quality by comparing predicted rankings against reference rankings derived from human annotations using two standard rank correlation coefficients. **Kendall's** $\tau$ measures the proportion of concordant pairs, capturing pairwise ranking consistency. **Spearman's** $\rho$ evaluates the monotonic relationship between predicted and ground-truth rankings. Beyond absolute accuracy, we quantify reproducibility across random initialisations via **Normalised Rank Deviation**, a scale-invariant stability metric. For a candidate pool of size $N$, raw rank variance measures the average absolute ranking shift an item experiences across multiple runs, but its theoretical maximum grows with $N$, making direct comparisons across pool sizes misleading. We therefore normalise by pool size: $\text{NormRankDev} = \frac{\sqrt{\text{RankVar}}}{N}$. A value of 0 indicates perfect structural stability. For reference, uniformly random rankings yield an expected standard deviation of $N/\sqrt{12}$, corresponding to $\text{NormRankDev} \approx 0.289$, with values approaching this threshold indicating near-random rankings.

# 6 Results

## 6.1 Comparison to Prior Work

We begin by establishing a direct comparison to prior work. To ensure comparability with Xu et al. (2025), we replicate their experimental setup on the LC AlpacaEval dataset,[4] using the same pool of 20 models. This provides a controlled setting in which SWIM was originally designed and evaluated. As shown in Appendix G, SWIM matches round-robin performance at a fraction of the comparison budget, consistent with the findings of Xu et al. (2025). At this scale, SWIM also outperforms Swiss and its hybrid variants on both Spearman and Kendall correlation with the Chatbot Arena oracle, whilst maintaining lower NormRankDev. This confirms our implementation is faithful and provides a strong baseline for the subsequent tournament comparisons. Importantly, $N{=}20$ represents a pool size for SWIM to perform the best, as it was specifically designed and evaluated at this scale.

## 6.2 Plateau Effect and Scaling Behaviour

As shown in Figure 1 and 2, SWIM and Swiss-style formats consistently recover the majority of round-robin ranking quality at a fraction of the comparison budget across both tasks and all pool sizes. A clear plateau effect emerges: ranking quality saturates rapidly, with Swiss formats reaching near round-robin performance well before exhausting the full comparison budget. Round sensitivity analysis confirms that the theoretical optimum $R = \lceil \log_2(N) \rceil$ provides a sound operating point, with gains beyond this being negligible (see Appendix L).

Single Elimination performs worst across all settings, exhibiting both the lowest ranking accuracy and the highest NormRankDev, confirming that sparse unstructured pairwise evaluation introduces substantial bracket luck (where outcomes depend heavily on early pairings). Double Elimination improves over Single Elimination but remains well below Swiss-style formats at comparable budgets at higher values of $N$.

The relative performance of SWIM and Swiss-style formats depends on scale. SWIM performs best at small pool sizes (around $N{=}20$), consistent with its original design and evaluation setting (see Appendix G). However, as $N$ grows, its advantage diminishes, and Swiss-style formats become more effective, both in accuracy and stability. This reflects a fundamental difference: SWIM is sequential and optimised for small pools, whereas Swiss-style formats scale more effectively with larger candidate sets. The stability advantage of Swiss-style formats stems from a structural property: all candidates remain active throughout the tournament, allowing early mismatches to be corrected as more signal accumulates across rounds. Elimination-based formats make irreversible decisions after one or two losses, meaning an unlucky early pairing permanently removes a candidate before sufficient

[4] https://github.com/tatsu-lab/alpaca_eval

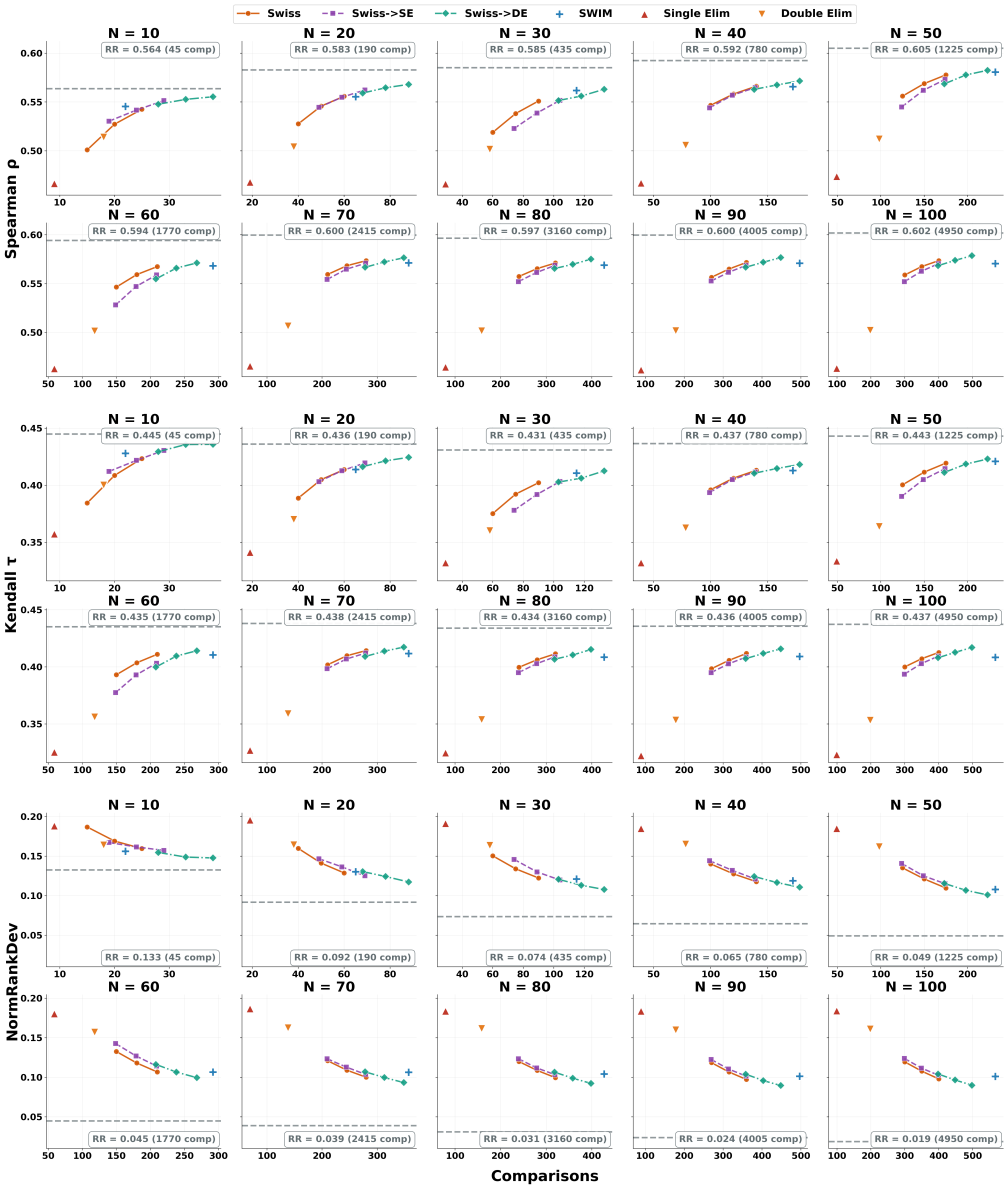



Figure 1: Ranking recovery on WMT across pool sizes and comparison budgets, using GPT-OSS 20B. Each Swiss-family line traces increasing round counts (R-1, R, R+1). Dashed line shows full round-robin performance.

evidence has been gathered. This asymmetry is particularly damaging in low-signal regimes, where individual comparison outcomes are noisy and a single incorrect elimination cannot be recovered.

At small pool sizes, Swiss and Swiss→DE briefly match or outperform round-robin. Notably, as round-robin requires only 45 comparisons, meaning Swiss formats are operating at a comparison count already close to the exhaustive budget, further reducing the practical gap between them. As $N$ grows, round-robin's density advantage reasserts itself, though Swiss formats continue to recover most of its performance at substantially reduced cost.

Finally, SWIM and Swiss-style formats maintain low rank deviation across random initialisations, converging toward round-robin's perfect stability as additional rounds are added. In particular, Swiss-style formats exhibit more consistent scaling behaviour at larger $N$, where stability plateaus more reliably.These trends are consistent across both the MT and Essay tasks and replicated under Phi-4 as an alternative judge (see Appendix I).

To further probe robustness in challenging conditions, we evaluate tournament behaviour in a low-signal setting based on borderline peer review decisions, where score differences are minimal and initialisation can introduce bias. Results show that elimination-based formats are highly sensitive to seeding, while Swiss-style formats remain robust (see Appendix H), consistent with the stability

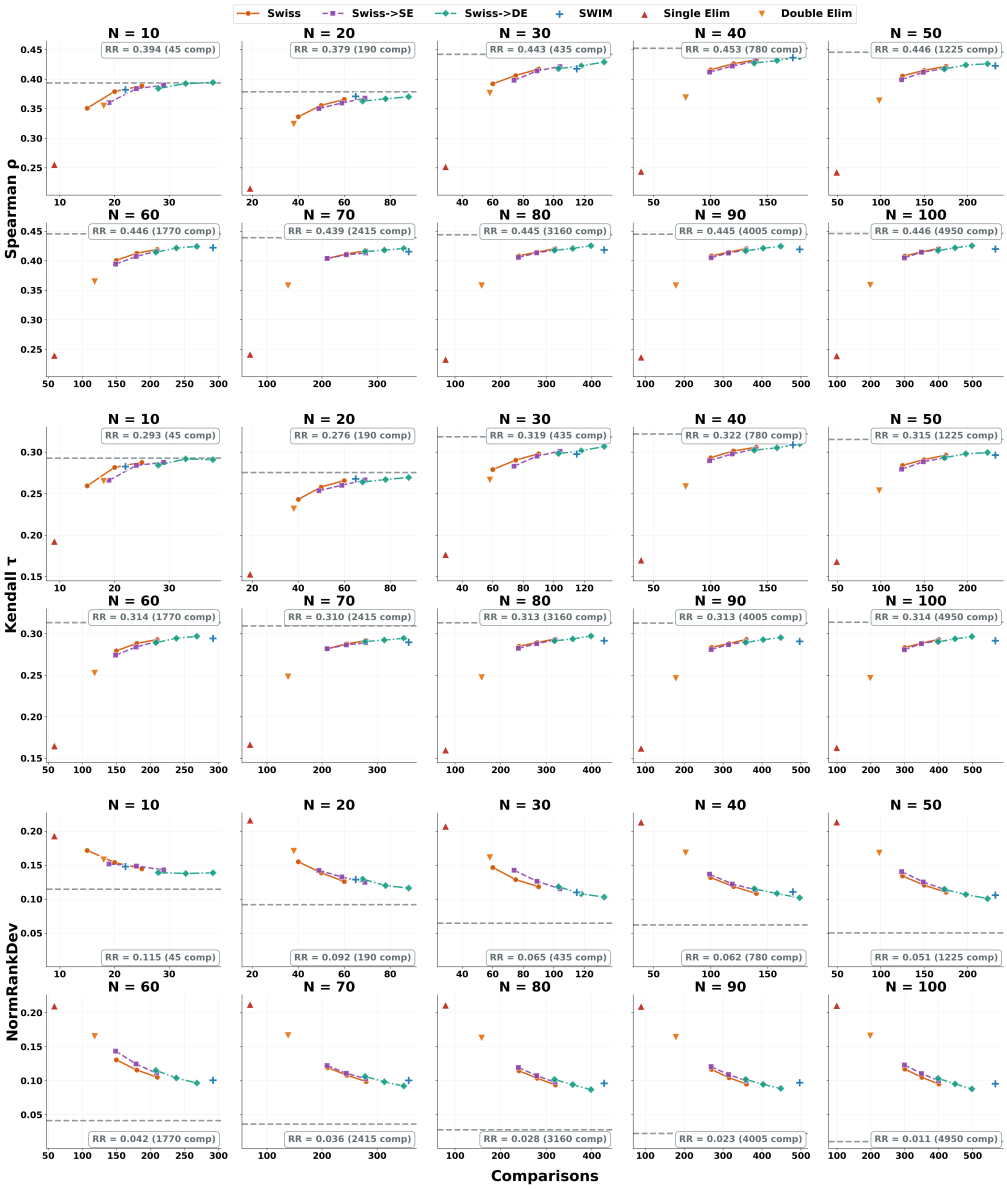



Figure 2: Ranking recovery on DResS across pool sizes and comparison budgets, using GPT-OSS 20B. Each Swiss variant line traces increasing round counts (R-1, R, R+1). Dashed line shows full round-robin performance.

trends observed above for higher values of $N$.

**Aggregation Methods.** As shown in Appendix F, we compare aggregation methods against Swiss-paired variants at matched budget. TrueSkill consistently fails across both tasks and pool sizes, producing near-zero or negative correlations, while Crowd-BT and ASAP lag behind Swiss and SWIM at matched budget. Hybrid methods (Swiss+TrueSkill, Swiss+Crowd-BT, Swiss+ASAP) show no consistent trend, with performance varying across datasets and pool sizes. These results confirm that tournament structure is the primary driver of ranking quality, and are not an artefact of the BT aggregation choice.

### 6.3 Cost–Accuracy Trade-offs

Building on these results, we examine whether the gains of structured tournament design translate into practical advantages under computational constraints.Evaluation cost and latency scale with the number of model calls, particularly in settings such as reinforcement learning from feedback and large-scale benchmarking. We therefore compare Swiss, SWIM, and Single Elimination against a pointwise baseline.Tournament formats differ not only in accuracy but in execution. Swiss pairings are parallel within each round, requiring only $\lceil \log_2(N) \rceil$ sequential steps. In contrast, SWIM is inherently sequential, as each comparison

depends on updated Bradley–Terry estimates (Xu et al., 2025), and elimination formats offer limited parallelism due to their sparse structure.

We evaluate GPT-5.4-mini on DREsS using the zero-shot ICL scoring prompt from Yoo et al. (2025) as our pointwise baseline alongside tournament formats at $N$=100. As shown in Table 1, Swiss surpasses the pointwise baseline at approximately 200 comparisons, demonstrating that structured pairwise evaluation can achieve higher accuracy at comparable cost. Single Elimination never recovers pointwise-level accuracy, confirming that sparse unstructured pairwise evaluation is strictly worse than absolute scoring at equivalent cost which is consistent with the findings of Tripathi et al. (2025). SWIM requires substantially more comparisons and cannot be batched, making it less practical in constrained settings. Furthermore SWIM looks at each candidate in sequential order hence capping the budget reduces the number of candidates with an actual ranking.

**Performance vs. Comparison Difficulty.** We analyse results as a function of score gap. We bin pairs by gold score gap to compute pairwise accuracy, and classify triplets into LL, LM, ML, and MM difficulty scenarios (where L and M denote low- and medium/high-difficulty comparisons based on score gap) following Xu et al. (2025). As shown in Table 2, all methods perform near chance at small gaps [0,1), with pointwise marginally leading, indicating that near-identical candidates are intrinsically difficult to distinguish. As gaps widen, performance improves across all methods, with Swiss consistently outperforming alternatives and achieving the largest gains, demonstrating that structured comparison allocation more effectively exploits available signal. Consistent trends are observed in the MT and Essay experiments (see Appendix J), where elimination-based formats often underperform pointwise scoring at small gaps, while Swiss-style formats yield clear advantages once sufficient signal is present.

Taken together, these findings demonstrate that the advantages of structured tournament design are consistent across evaluators (both open source proprietary models) and tasks, confirming that the observed trends are driven by tournament structure rather than judge choice.

## 7 Conclusion

Pairwise evaluation is often treated as a reliable default, but our results show that its effectiveness depends critically on how comparisons are allocated. We demonstrate that ranking quality is not determined solely by the number of comparisons or the choice of aggregation model, but by the structure of the comparison graph itself. Across tasks and scales, we find that most ranking signal can be recovered with a small fraction of comparisons, and that structured tournament designs, particularly Swiss-style formats, consistently achieve strong performance while remaining efficient under practical constraints. In contrast, unstructured or sequential schemes either fail to utilise available signal or incur prohibitive execution costs. These findings suggest a shift in perspective: evaluation should be viewed not as scoring or ranking in isolation, but as a resource allocation problem over comparisons. Designing efficient and reliable evaluation protocols therefore requires careful consideration of both statistical and computational structure.

| | SE | | SWIM | | Swiss | |
|---|---|---|---|---|---|---|
| **#Comp** | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| pointwise | 0.484 | 0.344 | 0.484 | 0.344 | 0.484 | 0.344 |
| 50 | 0.259 | 0.191 | 0.072 | 0.053 | 0.316 | 0.224 |
| 100 | 0.297 | 0.210 | 0.122 | 0.086 | 0.427 | 0.299 |
| 150 | – | – | 0.193 | 0.136 | 0.471 | 0.324 |
| 200 | – | – | 0.209 | 0.143 | 0.508 | 0.349 |
| 250 | – | – | 0.357 | 0.241 | 0.542 | 0.372 |
| 300 | – | – | 0.358 | 0.242 | 0.549 | 0.379 |
| 350 | – | – | 0.331 | 0.223 | 0.570 | 0.393 |
| 400 | – | – | 0.373 | 0.252 | 0.580 | 0.398 |
| 450 | – | – | 0.414 | 0.281 | 0.597 | 0.412 |
| 500 | – | – | 0.444 | 0.307 | 0.594 | 0.412 |
| 550 | – | – | 0.461 | 0.316 | 0.591 | 0.409 |
| 567 | – | – | 0.468 | 0.321 | – | – |

Table 1: Correlation at each budget using GPT-5.4-mini. Single Elimination (SE) terminates at 99 comparisons as its bracket structure permits no further rounds. Shaded cells mark the optimal comparisons for each format. Budgets increase in steps of 50 to ensure every candidate is paired at least once per increment.

| Gap | #Pairs | SE | PW | SWIM | Swiss |
|---|---|---|---|---|---|
| [0, 1) | 1653 | 0.513 | **0.532** | 0.493 | 0.499 |
| [1, 2) | 2181 | 0.499 | 0.529 | 0.563 | **0.565** |
| [2, 3) | 1968 | 0.555 | 0.612 | 0.601 | **0.633** |
| [3, 4) | 1776 | 0.592 | 0.661 | 0.635 | **0.683** |
| [4, 5) | 1584 | 0.619 | 0.709 | 0.668 | **0.739** |
| [5, 6) | 1392 | 0.683 | 0.745 | 0.696 | **0.766** |
| [6, 7) | 1200 | 0.701 | 0.774 | 0.717 | **0.801** |
| [7, 8) | 1008 | 0.720 | 0.804 | 0.794 | **0.865** |
| [8, 9) | 816 | 0.724 | 0.800 | 0.824 | **0.884** |
| [9, 10) | 624 | 0.736 | 0.848 | 0.859 | **0.934** |
| [10, 11) | 432 | 0.594 | 0.817 | 0.854 | **0.949** |
| [11, 12) | 216 | 0.679 | 0.801 | 0.893 | **0.989** |

Table 2: Pairwise ranking accuracy by score gap using GPT-5.4-mini. SE is single elimination and PW is pointwise. Swiss dominates at larger gaps while all methods converge near chance for closely-scored pairs. With non-transitivity analysis shown in Appendix K.

## Limitations

Our experiments are limited to three task domains (essay scoring, machine translation and borderline peer review decisions), and findings may not generalise to other open-ended generation tasks such as dialogue or summarisation. While we selected datasets with large candidate pools and dense human annotation to enable robust subsampling, our conclusions remain empirically grounded and a theoretical account of when and why ranking quality saturates remains an open question.

While we mitigate prompt-level biases such as positional and order bias through repeated candidate orderings, we do not account for biases inherent in the human annotations used as ground truth, including annotator disagreement, cultural bias, and domain-specific scoring inconsistencies. However, as all methods are evaluated under the same annotation conditions, any such biases affect all tournament formats equally, and our relative findings remain consistent.

We selected two open and one closed model for our experiments due to computational and access constraints. However, we believe our findings generalise across ranking models: the tournament structure operates independently of the underlying LLM, which affects only the noise level of individual comparisons. Models with higher pairwise accuracy will naturally yield more stable rankings, while those performing near or below chance are not useful for tournament-based evaluation regardless of structure.

## Ethics Statement

This work involves the analysis of publicly available ICLR 2025 and 2026 data via OpenReview, which releases accepted and rejected papers alongside their reviews under open access. No private or confidential reviewer information was accessed beyond what is publicly available through the official OpenReview platform. The application of tournament-based evaluation is presented strictly as an empirical study of ranking methodology and should not be interpreted as a recommendation to replace or automate human peer review decisions. Deploying automated ranking systems in high-stakes decision-making carries significant risks, including the potential to amplify biases present in the underlying review text or judge model, and further validation across venues, disciplines, and judge models is necessary before any practical deployment could be considered.

## References


Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

David R Appleton. 1995. May the best man win? *Journal of the Royal Statistical Society Series D: The Statistician*, 44(4):529–538.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Li-Guo Chen, Zheng Xiao, Yi-Jiang Xu, Rui-Chuan An, Xin Wang, Yang-Ning Li, Ying-Hui Li, Yi-Dong Wang, Zheng-Ran Zeng, Qing Gao, and 1 others. 2025. Coderankeval: Benchmarking and analyzing llm performance for code ranking. *Journal of Computer Science and Technology*, 40(5):1220–1233.

Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202.

Roland Daynauth, Christopher Clarke, Krisztian Flautner, Lingjia Tang, and Jason Mars. 2025. Ranking unraveled: Recipes for LLM rankings in head-to-head AI combat. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26078–26091, Vienna, Austria. Association for Computational Linguistics.

Karel Devriesere, László Csató, and Dries Goossens. 2025. Tournament design: A review from an operational research perspective. *European Journal of Operational Research*, 324(1):1–21.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Oliver Engist, Erik Merkus, and Felix Schafmeister. 2021. The effect of seeding on tournament outcomes: Evidence from a regression-discontinuity design. *Journal of Sports Economics*, 22(1):115–136.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sangkeun Jung, Goun Pyeon, Inbum Heo, and Hyungjin Ahn. 2025. What drives paper acceptance? a process-centric analysis of modern peer review. *arXiv preprint arXiv:2509.25701*.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jing Yang, and Nedjma Ousidhoum. 2025. Insights from the iclr peer review and rebuttal process. *arXiv preprint arXiv:2511.15462*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk. 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2559–2566. IEEE.

Tom Minka, Ryan Cleven, and Yordan Zaykov. 2018. Trueskill 2: An improved bayesian skill rating system. *Technical Report*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sukannya Purkayastha, Nils Dycke, Anne Lauscher, and Iryna Gurevych. 2025. Decision-making with deliberation: Meta-reviewing as a document-grounded dialogue. *arXiv preprint arXiv:2508.05283*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. 2025. Reviewertoo: Should ai join the program committee? a look at the future of peer review. *arXiv preprint arXiv:2510.08867*.

Pascal Sauer, Ágnes Cseh, and Pascal Lenzner. 2024. Improving ranking quality and fairness in swiss-system chess tournaments. *Journal of Quantitative Analysis in Sports*, 20(2):127–146.

Allen J Schwenk. 2000. What is the correct way to seed a knockout tournament? *The American Mathematical Monthly*, 107(2):140–150.

Takumi Shibata and Yuichi Miyamura. 2025. LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29988–30001, Suzhou, China. Association for Computational Linguistics.

Seonil Son, Ju-Min Oh, Heegon Jin, Cheolhun Jang, Jeongbeom Jeong, and Kuntae Kim. 2025. Arena-lite: Efficient and reliable large language model evaluation via tournament-based direct comparisons. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7068–7086.

Balázs R Sziklai, Péter Biró, and László Csató. 2022. The efficacy of tournament designs. *Computers & Operations Research*, 144:105821.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. Pairwise or pointwise? evaluating feedback protocols for bias in LLM-based evaluation. In *Second Conference on Language Modeling*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating non-transitivity in LLM-as-a-judge. In *Forty-second International Conference on Machine Learning*.

Jing Yang. 2025. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Jing Yang, Qiyao Wei, and Jiaxin Pei. 2025. Paper copilot: Tracking the evolution of peer review in ai conferences. *arXiv preprint arXiv:2510.13201*.

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. DREsS: Dataset for rubric-based essay scoring on EFL writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13439–13454, Vienna, Austria. Association for Computational Linguistics.

Qiang Zhang, Boli Chen, Fanrui Zhang, Ruixue Ding, Shihang Wang, Qiuchen Wang, Yinfeng Huang, Haonan Zhang, Rongxiang Zhu, Pengyong Wang, and 1 others. 2026. Arenarl: Scaling rl for open-ended agents via tournament-based relative ranking. *arXiv preprint arXiv:2601.06487*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hangxiao Zhu, Yian Yin, and Yu Zhang. 2025. Survivors, complainers, and borderliners: Upward bias in online discussions of academic conference reviews. *arXiv preprint arXiv:2509.16831*.

# A Appendix

Evaluation was conducted on two Nvidia RTX 4090 GPUs. All models evaluated used the default parameters from vLLMs. The closed source GPT-5.4-mini models was called using the openai api[5].

# B Dataset Statistics

| Dataset | $N$ | Mean | Std | Min | Max |
|---|---|---|---|---|---|
| en-zh | 38262 | 79.79 | 15.75 | 0.00 | 100.00 |
| en-cs | 37100 | 82.06 | 20.90 | 0.00 | 100.00 |
| DREsS_New | 2112 | 10.34 | 2.26 | 3.00 | 15.00 |
| DREsS_Std | 6508 | 8.92 | 2.55 | 0.56 | 15.00 |

Table 3: Dataset statistics.

# C Prompts

**MT Pairwise Evaluation Prompt**

```
You are an expert machine translation evaluator.
Compare two {tgt_lang} translations.

<pair_a>
  <source lang="{src_lang}">
    {item_a.src}
  </source>
  <translation>{item_a.mt}</translation>
</pair_a>

<pair_b>
  <source lang="{src_lang}">
    {item_b.src}
  </source>
  <translation>{item_b.mt}</translation>
</pair_b>

Evaluate on adequacy, fluency, terminology.
Respond with ONLY:
<choice>A</choice> or <choice>B</choice>
```

**Essay Pairwise Evaluation Prompt**

```
You are an expert essay evaluator for English
as a Foreign Language (EFL) writing.
Compare two student essays written in response
to the prompt.
<prompt_a>{item_a.prompt_text}</prompt_a>
<essay_a>{item_a.essay}</essay_a>

<prompt_b>{item_b.prompt_text}</prompt_b>
<essay_b>{item_b.essay}</essay_b>

Evaluate on content, organization, and
language quality.
Respond with ONLY:
<choice>A</choice> or <choice>B</choice>
```

[5] https://openai.com/api/

**ICLR Pairwise Evaluation Prompt**

```
You are an expert academic reviewer for a top
machine learning conference.
Compare two submitted papers based on their
reviews and meta-review.
<paper_a>
  paper_a metadata, reviews, and meta-review
</paper_a>

<paper_b>
  paper_b metadata, reviews, and meta-review
</paper_b>

Evaluate on technical soundness, novelty,
clarity, reviewer consensus,
and severity of weaknesses.
Respond with ONLY:
<choice>A</choice> or <choice>B</choice>
```

## D Tournament match complexity

As shown in Table 4, round-robin grows quadratically and becomes prohibitive at scale, while single elimination is the most frugal at $N-1$ matches but offers minimal ranking signal. Swiss operates in $\mathcal{O}(N \log N)$ matches, providing a middle ground between coverage and cost, with the theoretical optimum at $R = \lceil \log_2(N) \rceil$ rounds. Hybrid formats (Swiss→SE, Swiss→DE) incur the Swiss seeding cost plus an elimination phase, with double elimination adding roughly twice the overhead of single elimination. SWIM also scales as $\mathcal{O}(N \log N)$ but its sequential insertion distributes the budget unevenly: earlier-inserted candidates receive fewer comparisons than later ones, which can disadvantage strong candidates inserted early.

| Format | Match Count |
|---|---|
| Round Robin | $\frac{N(N-1)}{2}$ |
| Single Elim | $N-1$ |
| Double Elim | $2N-2$ or $2N-1$ |
| Swiss | $\frac{N \cdot \lceil \log_2(N) \rceil}{2}$ |
| Swiss→SE | $\frac{N \cdot R_{\text{seed}}}{2} + (N-1)$ |
| Swiss→DE | $\frac{N \cdot R_{\text{seed}}}{2} + \{2N-2, 2N-1\}$ |
| SWIM | $\sum_{i=1}^{N} \lceil \log_2(i) \rceil$ |

Table 4: Match complexity for each tournament format as a function of pool size $N$, where $R_{\text{seed}} = R-1$.

# E Preliminary pairwise results

| Model | Avg. Acc. | Essay Evaluation Acc. | Essay Evaluation 95% CI | MT Evaluation Acc. | MT Evaluation 95% CI | Avg. Wall Time (s) |
|---|---|---|---|---|---|---|
| **Phi-4** | **0.646** | 0.680 | [0.622, 0.738] | **0.612** | [0.552, 0.672] | 31.35 |
| **GPT-OSS-20B** | 0.639 | **0.705** | [0.647, 0.764] | 0.572 | [0.506, 0.638] | 18.13 |
| Gemma-3-12B-it | 0.618 | 0.644 | [0.585, 0.703] | 0.592 | [0.531, 0.653] | 29.07 |
| Qwen3-14B | 0.590 | 0.612 | [0.552, 0.672] | 0.568 | [0.507, 0.629] | 30.03 |
| InternLM3-8B | 0.576 | 0.636 | [0.576, 0.696] | 0.516 | [0.454, 0.578] | 20.57 |
| Qwen3-8B | 0.576 | 0.600 | [0.539, 0.661] | 0.552 | [0.490, 0.614] | 16.71 |
| Phi-4-mini | 0.536 | 0.528 | [0.466, 0.590] | 0.544 | [0.482, 0.606] | **8.55** |
| OLMo-3-7B | 0.532 | 0.516 | [0.454, 0.578] | 0.548 | [0.486, 0.610] | 16.94 |
| Ministral-8B | 0.532 | 0.584 | [0.523, 0.645] | 0.480 | [0.418, 0.542] | 16.79 |
| Llama-3.1-8B | 0.522 | 0.488 | [0.426, 0.550] | 0.556 | [0.494, 0.618] | 16.45 |

Table 5: Model performance sorted by Average Accuracy.

# F Aggregation Methods

| method | $N=20$ | | $N=50$ | | $N=100$ | |
|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| MT | | | | | | |
| `swiss` | 0.341 | 0.239 | 0.507 | 0.366 | 0.570 | 0.411 |
| `swim` | 0.288 | 0.198 | 0.539 | 0.395 | 0.575 | 0.419 |
| `trueskill` | -0.287 | -0.196 | -0.196 | -0.153 | 0.176 | 0.117 |
| `crowd_bt` | 0.247 | 0.156 | 0.450 | 0.315 | 0.519 | 0.368 |
| `asap` | 0.281 | 0.177 | 0.547 | 0.396 | 0.562 | 0.408 |
| `swiss_trueskill` | 0.314 | 0.219 | 0.515 | 0.369 | 0.550 | 0.398 |
| `swiss_crowd_bt` | 0.298 | 0.200 | 0.534 | 0.387 | 0.562 | 0.409 |
| `swiss_asap` | 0.327 | 0.214 | 0.552 | 0.399 | 0.548 | 0.394 |
| Essay | | | | | | |
| `swiss` | 0.334 | 0.233 | 0.287 | 0.200 | 0.467 | 0.328 |
| `swim` | 0.324 | 0.214 | 0.321 | 0.236 | 0.452 | 0.316 |
| `trueskill` | -0.226 | -0.163 | -0.279 | -0.179 | -0.307 | -0.197 |
| `crowd_bt` | 0.379 | 0.268 | 0.288 | 0.203 | 0.407 | 0.282 |
| `asap` | 0.277 | 0.177 | 0.338 | 0.233 | 0.423 | 0.299 |
| `swiss_trueskill` | 0.336 | 0.242 | 0.348 | 0.246 | 0.418 | 0.291 |
| `swiss_crowd_bt` | 0.322 | 0.221 | 0.325 | 0.225 | 0.448 | 0.315 |
| `swiss_asap` | 0.356 | 0.240 | 0.329 | 0.233 | 0.457 | 0.322 |

Table 6: Spearman and Kendall correlations across pool sizes $N$.

## G Alpaca results

| Method | #Match | Budget% | Spearman | Kendall | NRD |
|---|---|---|---|---|---|
| Round Robin | 190 | 100.0 | 0.962 | 0.863 | 0.000 |
| Random | – | – | 0.000 | 0.000 | 0.289 |
| Single Elim | 19 | 10.0 | 0.725 | 0.552 | 0.179 |
| Double Elim | 38 | 20.0 | 0.879 | 0.724 | 0.117 |
| SWIM | 65 | 34.2 | **0.962** | **0.859** | **0.034** |
| Swiss $R$=4 | 40 | 21.1 | 0.931 | 0.792 | 0.093 |
| Swiss $R$=5 | 50 | 26.3 | 0.930 | 0.806 | 0.078 |
| Swiss $R$=6 | 60 | 31.6 | 0.943 | 0.823 | 0.046 |
| Swiss→SE $R$=3 | 49 | 25.8 | 0.929 | 0.798 | 0.084 |
| Swiss→SE $R$=4 | 59 | 31.1 | 0.955 | 0.842 | 0.056 |
| Swiss→SE $R$=5 | 69 | 36.3 | 0.958 | 0.853 | 0.045 |
| Swiss→DE $R$=3 | 68 | 35.8 | 0.944 | 0.825 | 0.060 |
| Swiss→DE $R$=4 | 78 | 41.1 | 0.961 | 0.855 | 0.041 |
| Swiss→DE $R$=5 | 88 | 46.3 | 0.955 | 0.838 | 0.042 |

Table 7: LC Alpaca eval: Tournament format comparison for ranking 20 LLMs. Budget% is the fraction of full round-robin comparisons used. Spearman and Kendall measure rank correlation with the Chatbot Arena oracle. NRD is normalised rank deviation (lower is more stable).

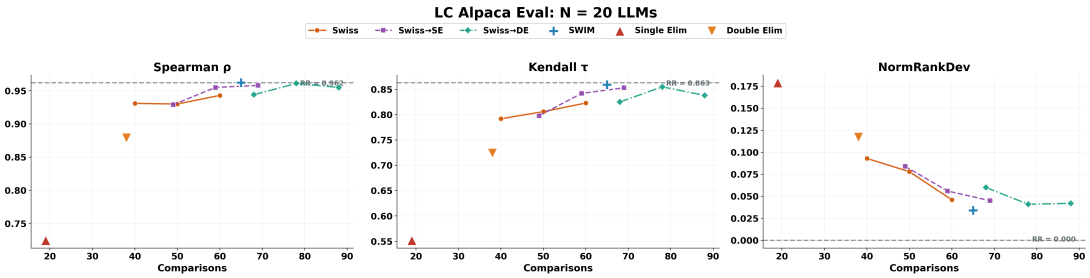



Figure 3: Visualisation of the results from Table 7

## H Borderline Decisions in Peer Review

### H.1 Setup

Borderline peer review decisions are an inherently high-stakes ranking problem where absolute scoring is known to fail. The difficulty of synthesising noisy reviewer signals into reliable accept/reject decisions motivates moving beyond static pointwise scores toward comparative structures (Sahu et al., 2025; Purkayastha et al., 2025). Score distributions exhibit systematic upward bias Zhu et al. (2025), and borderline outcomes are driven by reviewer influence rather than stable quality signals, reinforcing the case for tournament-based evaluation where absolute grading collapses (Kargaran et al., 2025; Jung et al., 2025).

We restrict our analysis to ICLR,[6] the only major ML venue that systematically releases both accepted and rejected submissions (Yang et al., 2025; Yang, 2025). NeurIPS,[7] ICML,[8] and COLM[9] either withhold rejections or rely on opt-in disclosure, introducing survivorship bias. We include both ICLR 2025 and 2026: following a security incident, ICLR 2026 scores were reverted to their pre-rebuttal state,[10] making them a particularly noisy seeding signal, with 2025 serving as a cleaner baseline. Borderline regions are defined as $[4.0, 6.0]$ for ICLR 2026 and $[5.0, 7.0]$ for ICLR 2025 (see Figure 4b and Figure 4a).

Borderline papers are grouped into semantically coherent clusters via TF-IDF keyword scoring:

$$\text{score}(k) = f_k \cdot \log\left(\frac{N}{c_k}\right),$$

where $f_k$ is keyword frequency, $c_k$ is the number of papers containing it, and $N$ is the total borderline pool. Jaccard-based deduplication removes keywords whose paper sets overlap by more than 0.5 with any higher-ranked keyword (see Figure 5 for examples). Groups of up to 100 papers[11] are evaluated using 5 independent tournament replicates under both random and score-based seeding, with ground-truth decisions withheld from the prompt.

[6] https://docs.openreview.net/reference/api-v2

[7] https://neurips.cc/Conferences/2025/ReviewerGuidelines

[8] https://icml.cc/Conferences/2025/PeerReviewFAQ

[9] https://colmweb.org/cfp.html

[10] https://blog.iclr.cc/2025/12/03/iclr-2026-response-to-security-incident/

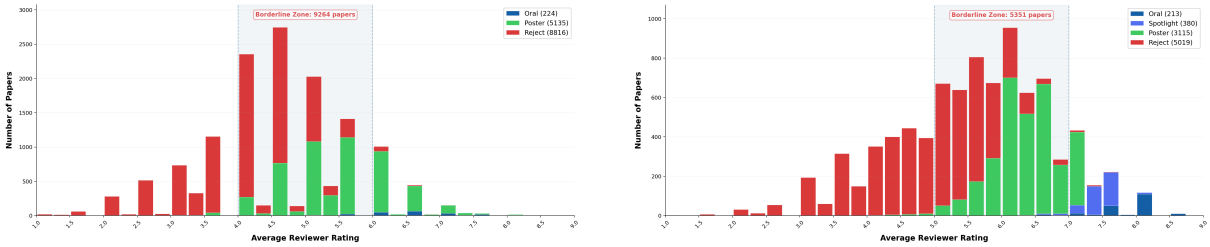
[11] This design requires approximately $17.5\times$ the LLM calls of a single Swiss tournament at fixed $N$.

### H.2 Results

For the ICLR borderline setting, where paper decisions are withheld from the judge, we assess discriminative power independently of fixed thresholds. **AUROC** measures the probability that a randomly sampled accepted paper ranks above a rejected one, with $0.5$ indicating random performance and $1.0$ a perfect ranking.

The results reveal a clear asymmetry in how seeding affects different tournament formats (As shown in Table 8). Score-based seeding actively degrades elimination formats, with single and double elimination both showing measurable AUROC and Kendall declines when seeded, as pre-seeding biases the LLM judge toward existing human consensus and prevents correction of initial misalignments in the borderline pool. Elimination formats consequently tend to preserve a potentially flawed initial ordering rather than adjudicating merit. Swiss-style formats and SWIM are invariant to seeding strategy, with AUROC and Kendall remaining stable or marginally improving regardless of initialisation, demonstrating a self-correcting property that makes them robust precisely where absolute scoring fails. Crucially, results for both 2025 and 2026 suggest that the tournament process is invariant to the noiser seeding conditions shown in ICLR 2026.

The per-group ICLR breakdown in Figure 6 and Figure 7 further shows that the same plateau effect observed in MT and Essay experiments replicates in the peer review setting: AUROC and Kendall saturate rapidly as comparisons increase, with SWIM and Swiss-style formats consistently approaching their performance ceiling well before exhausting the comparison budget, and Single Elimination remaining highly unstable across group sizes.




(a) ICLR 2026 Ratings vs. Decisions. The 4.0–6.0 range is the borderline region.

(b) ICLR 2025 Ratings vs. Decisions. The 5.0–7.0 range is the borderline region.

Figure 4: Comparative analysis of rating distributions and decision outcomes for ICLR 2026 and 2025. Oral and Spotlight papers are shown in Blue, Poster outcomes in Green, and Reject outcomes in Red.

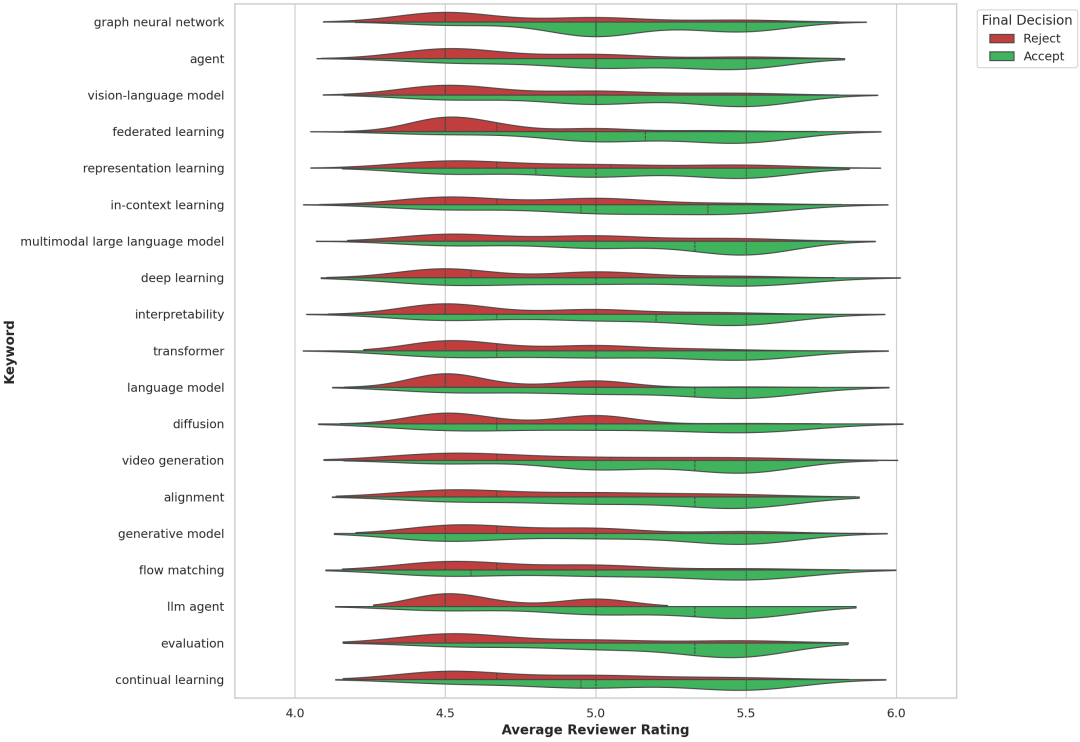



Figure 5: Examples of Accept vs. Reject Distributions for Keywords in ICLR 2026.

| Dataset | Method | AUROC | | | Kendall | | |
|---|---|---|---|---|---|---|---|
| | | R | S | Δ | R | S | Δ |
| ICLR 2025 | Single Elim | 0.803 | 0.781 | −2.74% | 0.426 | 0.396 | −7.04% |
| | Double Elim | 0.925 | 0.895 | −3.24% | 0.599 | 0.557 | −7.01% |
| | SWIM | 0.960 | 0.960 | +0.00% | 0.648 | 0.648 | +0.00% |
| | Swiss | 0.944 | 0.951 | +0.74% | 0.625 | 0.635 | +1.60% |
| | Swiss→SE | 0.954 | 0.959 | +0.52% | 0.640 | 0.646 | +0.94% |
| | Swiss→DE | **0.963** | **0.968** | +0.52% | **0.652** | **0.659** | +1.07% |
| ICLR 2026 | Single Elim | 0.803 | 0.783 | −2.49% | 0.414 | 0.387 | −6.52% |
| | Double Elim | 0.909 | 0.895 | −1.54% | 0.561 | 0.541 | −3.57% |
| | SWIM | 0.941 | 0.941 | +0.00% | 0.605 | 0.605 | +0.00% |
| | Swiss | 0.925 | 0.930 | +0.54% | 0.583 | 0.589 | +1.03% |
| | Swiss→SE | 0.932 | 0.935 | +0.32% | 0.593 | 0.597 | +0.67% |
| | Swiss→DE | **0.943** | **0.945** | +0.21% | **0.608** | **0.610** | +0.33% |

Table 8: Decision quality across formats under random (R) and score-based (S) seeding for ICLR 2025 and 2026 borderline submissions. Δ shows percentage change from random to score-based seeding, using GPT-OSS 20B.

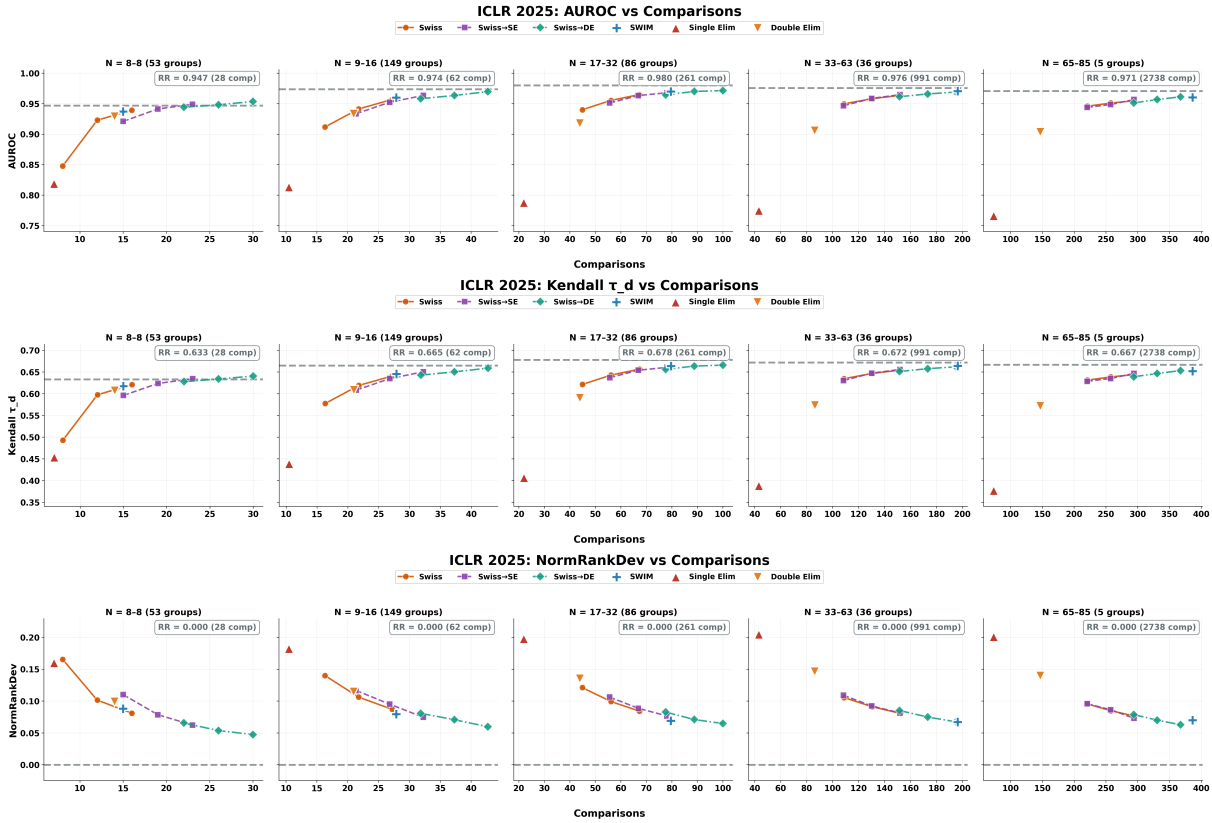



Figure 6: Comparative analysis of rating distributions and decision outcomes for ICLR 2025, using GPT-OSS 20B.

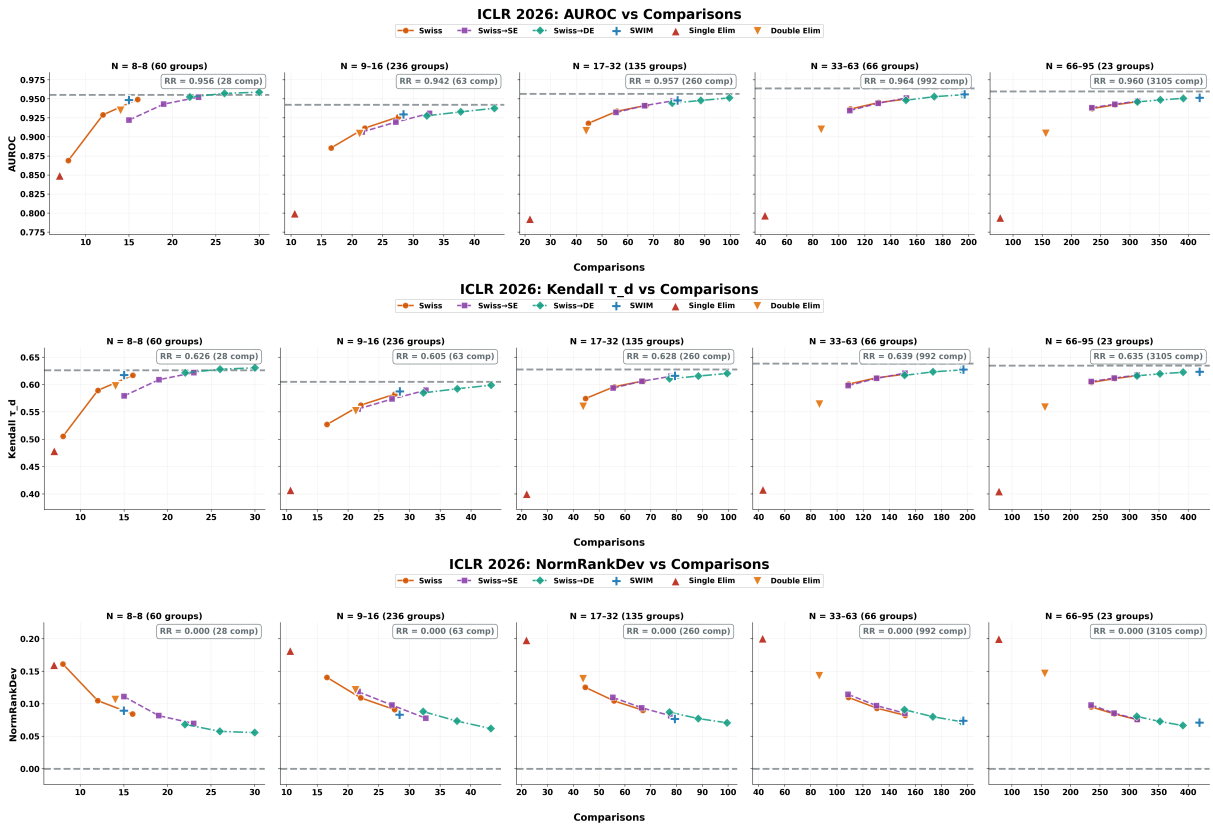



Figure 7: Comparative analysis of rating distributions and decision outcomes for ICLR 2026, using GPT-OSS 20B.

## I Model agnostic Validation

To verify the robustness and generalisability of the tournament system on Essay and MT, we reran the experiments on Phi-4. As shown in Figure 8 and Figure 9, the results show the same trends as in GPT-OSS.

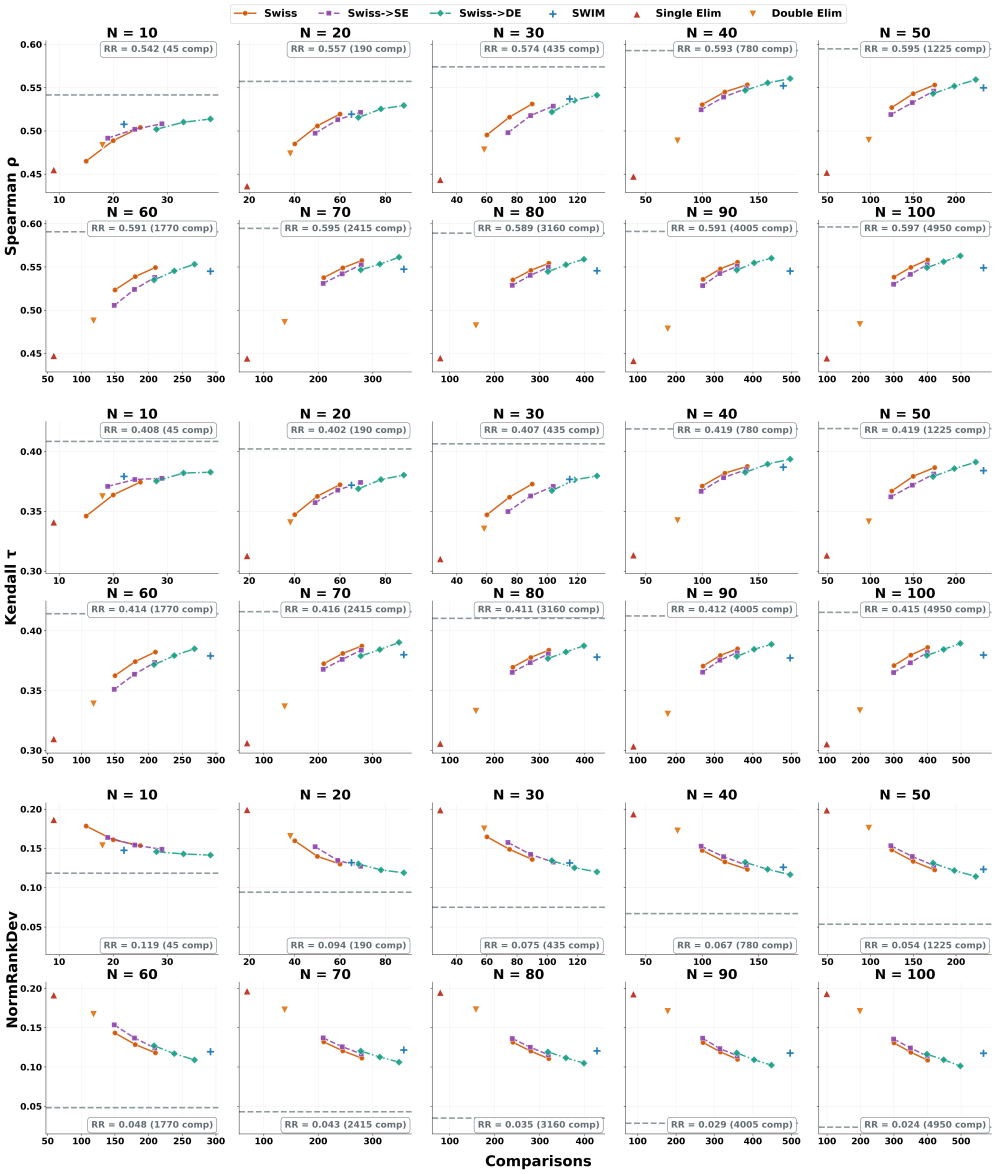



Figure 8: Ranking recovery on WMT across increasing pool sizes and comparison budgets, using Phi-4. Each Swiss-family line traces increasing round counts (R-1, R, R+1). Dashed line shows full round-robin performance.

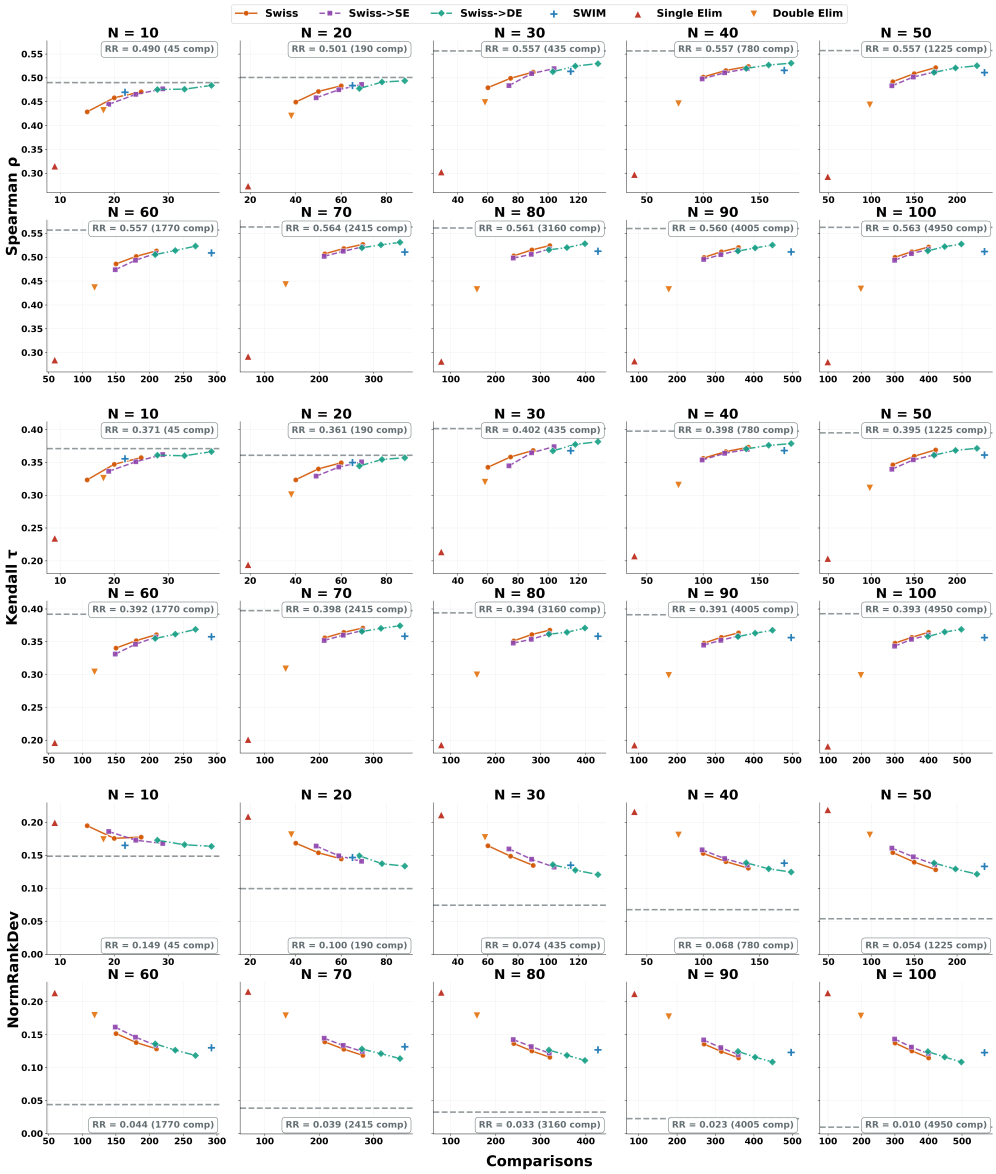



Figure 9: Ranking recovery on DResS across increasing pool sizes and comparison budgets, using Phi-4. Each Swiss-family line traces increasing round counts (R-1, R, R+1). Dashed line shows full round-robin performance.

## J Pointwise

As shown in Figure 10 and Figure 11, while pointwise scoring achieves median ranking accuracy approaching the tournament formats, it exhibits substantially higher variance across evaluation runs, suggesting that its performance is dataset-sensitive and unreliable. Swiss formats maintain consistently tight performance across conditions. However this backs up the points made in Tripathi et al. (2025), as they demonstrate that naive pairwise comparison is more vulnerable to distracted evaluation than pointwise scoring. Our results complement this finding: elimination-based formats, which represent the most unstructured form of pairwise evaluation, exhibit the highest instability and lowest ranking accuracy. However, Swiss-style tournaments substantially mitigate these weaknesses, recovering round-robin accuracy at a fraction of the cost and outperforming pointwise scoring on ranking fidelity and stability.

## K Non-transitivity

Following Xu et al. (2025), we measure non-transitivity in our LLM judge using PNT (Percentage of Non-Transitive cases) and SNTD (Soft Non-Transitivity Deviation) across four triplet scenarios: **LL** (both quality gaps large), **LM/ML** (one large, one marginal), and **MM** (both marginal). As shown in Figure 14, non-transitivity is highest in MM and lowest in LL across both tasks, replicating the core finding of Xu et al. (2025) that judges struggle most when candidates are similar in quality. This directly manifests in scenario accuracy: all formats achieve their highest all-3-correct rates in LL and converge to a shared lower bound in MM, where no tournament structure can fully compensate for judge uncertainty. Table 9 further corroborates this pattern using GPT-5.4-Mini, Swiss achieves the highest all-3-correct rate in the easiest (LL) and hardest (MM) scenarios, while Single Elimination consistently underperforms across all scenarios. on the other hand, pointwise performance is comparatively stronger in frontier models for non-transitivity, while for smaller open models they struggle.

## L Swiss Round Analysis

As shown in Figure 15, adding rounds beyond the theoretical optimum $R = \lceil \log_2(N) \rceil$ yields only marginal gains across both tasks and all three Swiss-style formats. Moving from $R{-}1$ to $R$ provides a modest improvement in ranking accuracy and stability, with further gains at $R{+}1$ being less significant. This suggests that the theoretical optimum $R$ provides a sound operating point, and additional rounds offer diminishing returns relative to their added computational cost.

| Method | LL | LM | ML | MM |
|---|---|---|---|---|
| Pointwise | 0.431 | 0.306 | 0.360 | 0.194 |
| Single Elim | 0.290 | 0.275 | 0.281 | 0.136 |
| SWIM | 0.376 | **0.339** | 0.328 | 0.220 |
| Swiss | **0.461** | 0.333 | **0.362** | **0.222** |

Table 9: All-3-correct rate by scenario for GPT-5.4-Mini results from subsection 6.3. LL = both gaps large (easy), MM = both gaps marginal (hard).

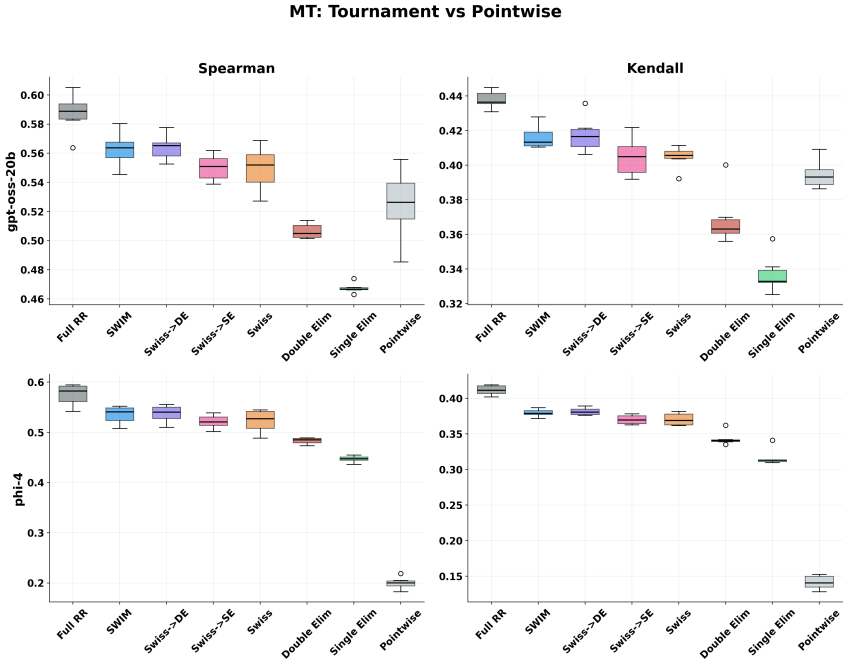



Figure 10: Comparison of tournament formats against pointwise scoring on MT, using GPT-OSS-20B and Phi-4 as judges.

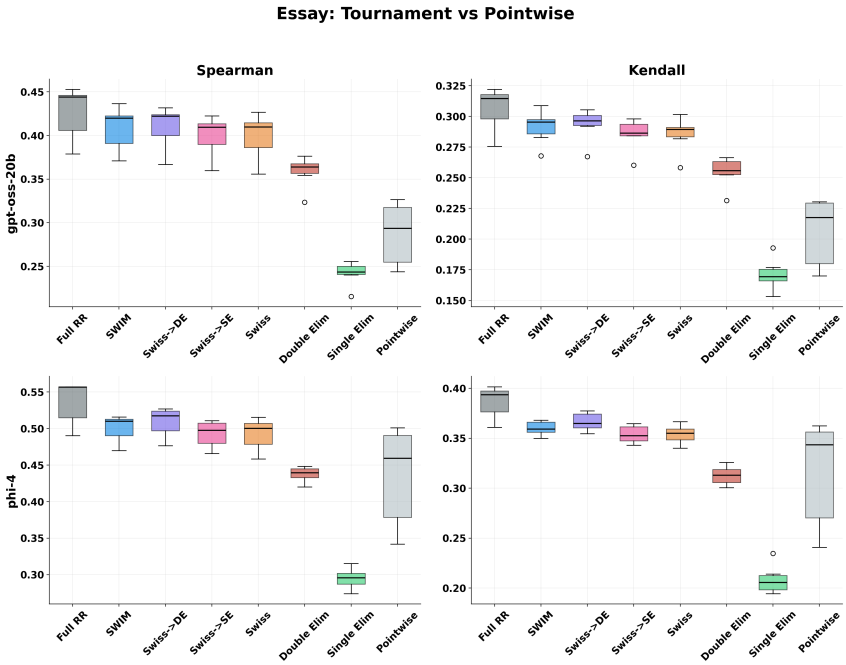



Figure 11: Comparison of tournament formats against pointwise scoring on Essay, using GPT-OSS-20B and Phi-4 as judges.

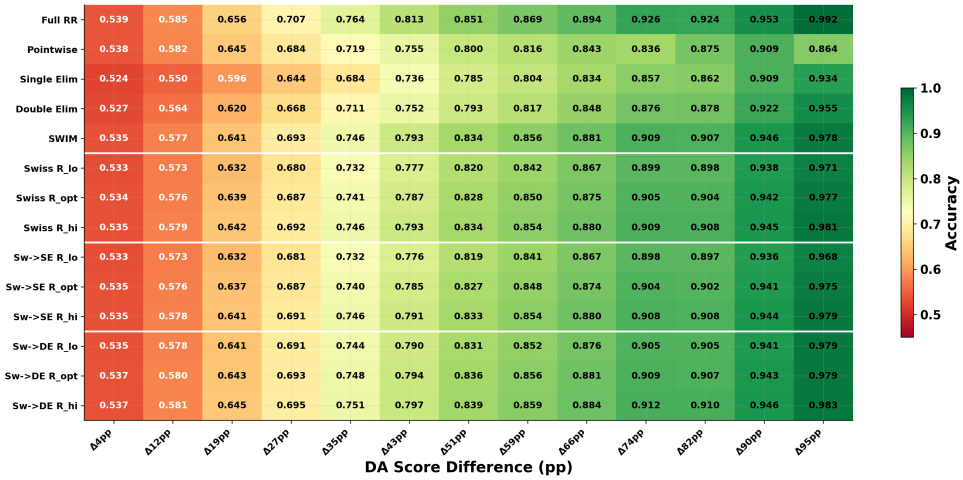



Figure 12: Score gap heatmap for pairwise comparisons for different tournament types and pointwise, using GPT-OSS-20B.

Figure 13: Score gap heatmap for pairwise comparisons for different tournament types and pointwise, using GPT-OSS-20B.

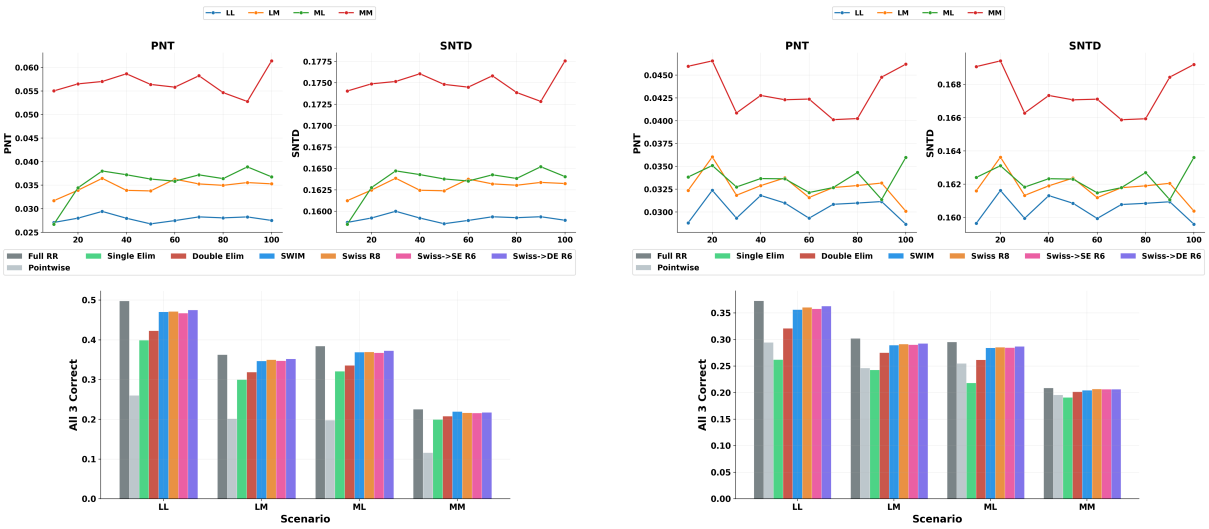



Figure 14: **Non-transitivity across MT (left) and Essay (right).** *Top:* PNT and SNTD by scenario; MM exhibits highest non-transitivity, LL lowest. *Bottom:* All-3-correct accuracy per scenario and format.

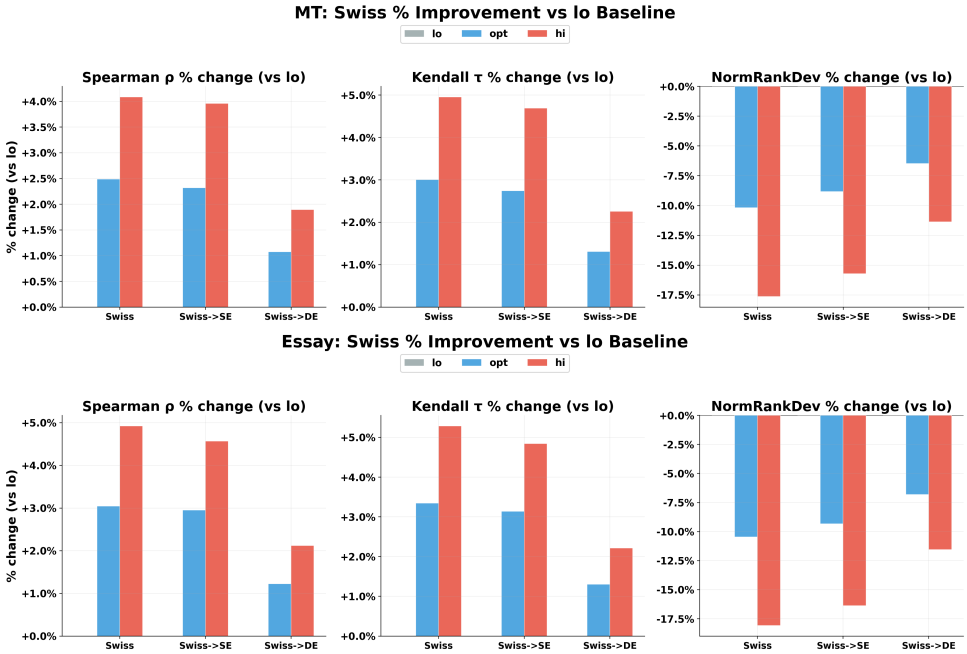



Figure 15: **Swiss round sensitivity for MT (left) and Essay (right).** Percentage improvement in Spearman, Kendall, and NormRankDev for $R{-}1$ (lo), $R$ (opt), and $R{+}1$ (hi).