Kurate Paper Rankings

VALIDATION REPORT

24.06.2026

swiss economics

DOCUMENT INFORMATION

Kurate Paper Rankings, 24.06.2026

Authors: Nicolas Oderbolz, Matthias Hafner

Contact: Matthias Hafner, +41 79 830 14 32, matthias.hafner@swiss-economics.ch

Executive Summary

This report validates the AI-generated paper rankings produced by Kurate (kurate.org) by comparing them against expert human assessments from the ICLR 2026 conference. Three Kurate ranking methods are evaluated: Win Rate and TrueSkill rankings based on pairwise paper comparisons, and an AI-produced Single Score ranking. These were compared against different ground-truth rankings derived from human reviewer scores and committee decisions, using Spearman and Kendall correlation coefficients as the main performance metrics. The analysis finds that all three Kurate ranking methods produce rankings meaningfully correlated with observed expert judgement, with Spearman coefficients ranging from 0.50 to 0.71. Single Score consistently outperforms the two methods based on pairwise paper comparison and outperforms the level of agreement observed between single human reviewers and program committee decisions in the ICLR 2026 dataset. These findings are robust across research subfields and are replicated in an out-of-sample comparison using ICLR 2024/25 data. The results support the use of Kurate as a scalable and cost-effective tool for automated paper screening.

Contents

1 Introduction

BACKGROUND

Kurate (kurate.org) is an automated scientific paper ranking system that uses large language models (LLMs) to evaluate the potential impact of academic preprints. It is designed as a scalable, low-cost alternative to traditional peer review. By automating the initial screening and ranking of papers, Kurate aims to help researchers, institutions, and funding bodies identify high-potential work more efficiently than conventional review processes allow. The platform currently covers a range of scientific fields including Machine Learning, Artificial Intelligence, Robotics, Game Theory, Economics, and Physics, drawing on preprints from arXiv. Rankings are derived from LLM-based assessments of relative paper quality, aggregated into a final ordering using established ranking algorithms.

The present analysis evaluates the validity of Kurate's AI-generated rankings by comparing them against human ground-truth rankings from the ICLR 2026 conference, where papers were assessed by expert reviewers on OpenReview. The goal is to assess how closely Kurate's rankings correspond to the consensus judgements of these domain experts.

RESEARCH QUESTION

The goal of this report is to validate the methodology Kurate uses to rank research papers by comparing Kurate rankings against rankings constructed from observable expert reviews. Kurate produces rankings via three distinct methods: Single Score, which directly elicits a quality judgement from the LLM for each paper, and different tournament-based methods, which derive rankings by aggregating the outcomes of pairwise paper comparisons.

Specifically, we ask the following research questions:

- Do Kurate rankings (both single-item and tournament-based) align with domain expert assessments of paper quality? Where is alignment with domain experts high, where is there divergence?
- Which specific Kurate ranking methods perform best relative to the expert benchmark? Is it possible to make methodological recommendations?

APPROACH

To evaluate these research questions, we proceed in four steps:

1. We apply Kurate's ranking methodology to a random sample of ICLR 2026 submissions, producing three ranking variants: Single Score, Win Rate, and TrueSkill.
2. We create different benchmark rankings that are based on individual human reviews and committee decisions of the papers in the ICLR dataset (“ground-truth” rankings).

3. We compare the different Kurate rankings against the constructed benchmarks, measuring the correspondence between Kurate and ground-truth rankings using visual analysis and correlation measures. As an additional benchmark, we evaluate how well single expert reviews in the ICLR 2026 sample correlate with the constructed ground-truth rankings.
4. We verify our results are robust against sample selection effects. We run the analysis on different subsamples (topics in ICLR), and we compare results from selected topics between 2025 and 2026 conferences.

Based on this analysis, we present an overall assessment of how effective the Kurate ranking methods are and whether they can outperform human domain experts.

STRUCTURE

The remainder of this paper is structured as follows:

- **Chapter 2** describes the data and the rankings, including sample construction, Kurate and ground-truth ranking methods, performance metrics, and the human benchmark design.
- **Chapter 3** presents the main results, reporting correlation coefficients and decile-level acceptance distributions for all three Kurate ranking methods.
- **Chapter 4** reports robustness checks across research subfields, an out-of-sample replication using ICLR 2024/25 data, and a convergence analysis for the pairwise ranking methods.
- **Chapter 5** discusses the findings, draws methodological recommendations, and outlines limitations and directions for future work.

2 Methods

This section describes the data, sample construction, and analytical approach underlying the validation. We first outline the primary dataset and the ground-truth rankings derived from it, before describing the Kurate ranking methods applied to the same sample. We then define the correlation metrics used to measure agreement between rankings and introduce the human benchmark against which Kurate's performance is evaluated.

2.1 Data

The present validation approach draws on a sample of papers submitted to the 2026 International Conference on Learning Representations (ICLR 2026).[1] Expert reviews and acceptance decisions for this conference are publicly available on OpenReview.[2] The submission window ran from 1 September to 25 September 2025.[3] Importantly, this means that reviewer evaluations were published to OpenReview after the knowledge cutoff of the models employed by Kurate (see Table 1), ensuring that the paper assessments underlying the Kurate rankings are not contaminated by the data used to construct the "ground-truth rankings.

Table 1: Knowledge cutoffs of models employed by Kurate

Model	Knowledge Cutoff	Source
GPT 5.4	August 31 2025	https://developers.openai.com/api/docs/models/gpt-5.2
Claude Opus 4.6	End of August 2025	https://support.claude.com/en/articles/8114494-how-up-to-date-is-claude-s-training-data
Gemini 3 Pro	January 2025	https://www.temso.ai/blog/ai-knowledge-cutoff-dates-every-major-llm-updated-for-2026

The ICLR 2026 review process entails multiple anonymous expert reviewers assigning each paper submission a numerical score on a discrete scale from 2 to 10 with a step size of two. Papers are subsequently accepted by a conference committee decision for Oral or Poster presentation, or rejected (see Figure 1). Together, these scores and acceptance decisions allow the construction of paper rankings that approximate the relative quality of each submission.

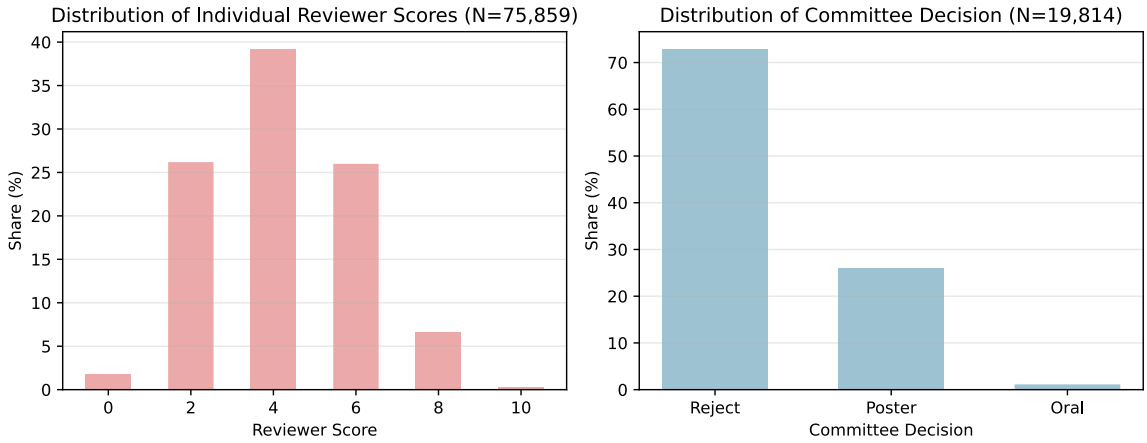
---

[1] Source: Berens Lab, University of Tübingen

[2] OpenReview is an open platform for scientific peer review that hosts paper submissions, reviews, and author responses for academic conferences and journals. It provides open access to submitted papers, reviewer comments, and author rebuttals.

[3] See ICLR 2026 Conference | OpenReview

Figure 1: Distribution of reviewer scores and committee decisions ICLR 2026



Additionally, around half of the submissions include specific topic labels, enabling analysis of ranking quality across different subtopics (see Table 2).[4]

Table 2: Frequency of topic labels ICLR 2026

Label	N	Share (%)
LLMs	2'380	22.7%
diffusion models	814	7.8%
RL	701	6.7%
language models	684	6.5%
optimization	357	3.4%
graphs	339	3.2%
vision-language models	300	2.9%
transformers	291	2.8%
safety	252	2.4%
adversarial	250	2.4%
unlabeled	*9,269*	*46.8%*
Total	**19'754**	**100.0%**

2.2 Sample Selection

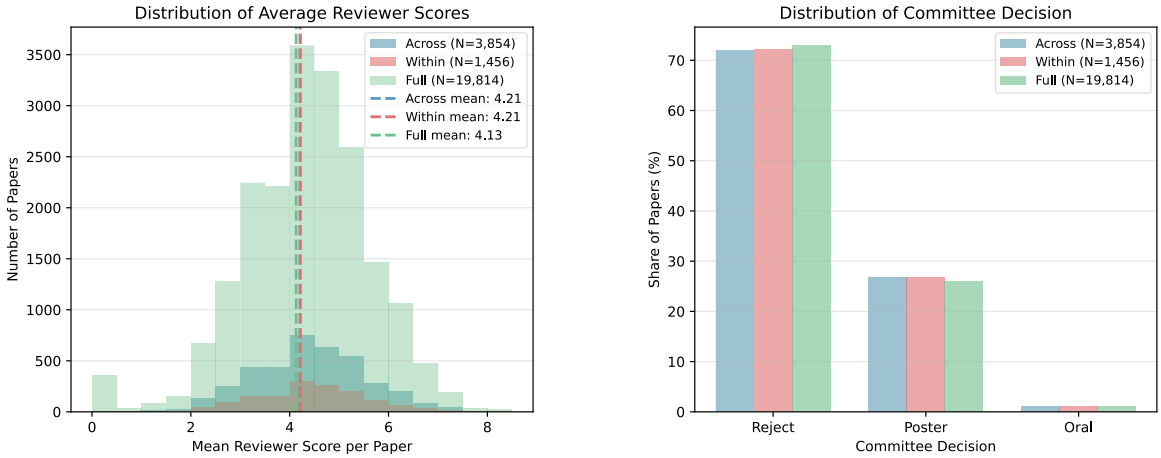
The Kurate ranking relies on computationally costly AI prompts. To keep the analysis tractable, we employ random sampling stratified by topic label, selecting 20% of papers from each topic label, including the "unlabeled" category. This yields a dataset of 3'854 papers, which we term the **"Across"** sample. We further construct a restricted dataset

[4] To label the dataset, Berens Lab relied on the author-provided keywords and used them to assign papers to non-overlapping classes. For more detail, see https://github.com/berenslab/iclr-dataset.

limited to topic labels with at least 30 sampled papers, yielding 1'456 papers, which we term the **"Within"** sample. Figure 2 shows the respective distributions of reviewer scores and committee decisions, which are consistent across samples.

For each paper in the sample, we subsequently select 30 opponent papers to form the basis for applying the Kurate pairwise comparison methods. In the "Across" sample, opponents are drawn randomly from the full set of sampled papers, meaning pairwise comparisons can occur between papers with different topic labels. In the "Within" sample, opponents are restricted to papers sharing the same topic label, so that all comparisons occur within a single research area. This approach allows us to evaluate whether the performance of Kurate's pairwise rankings is robust to comparisons across different topics.

Figure 2: Distribution of reviewer ratings and committee decisions



2.3 Kurate Rankings

We apply Kurate's ranking methodology to both samples, producing three ranking variants.

The **Single Score** ranking asks the AI models employed by Kurate to assign each paper an individual numerical score from 0 to 10 in increments of 0.1, reflecting its assessed quality and predicted scientific impact.[5] Papers are then ranked by their assigned score.

The **pairwise comparison rankings** are constructed by asking the AI models to select the stronger paper from each of the 30 paper pairs defined above. This produces a win matrix recording the outcome of each comparison, which is then aggregated into a unified ranking using one of two algorithms:

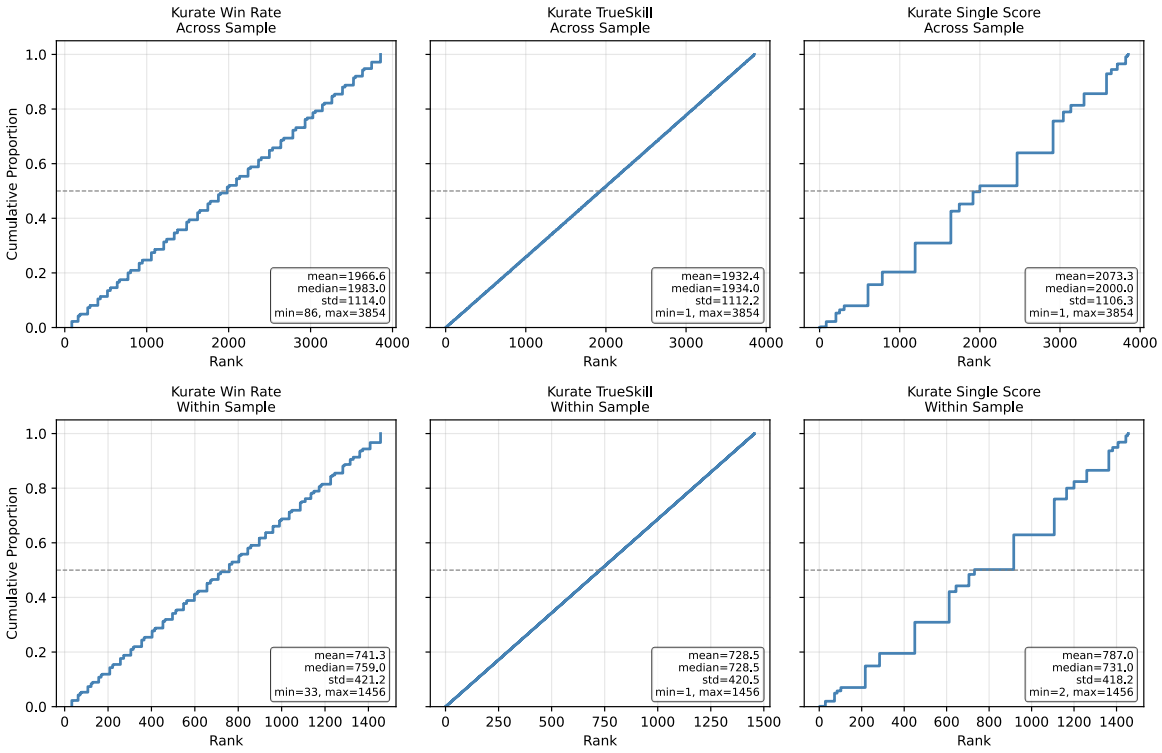
- **Win Rate** ranks papers by the share of pairwise comparisons they win out of their 30 matches.

[5] The prompts used to generate these scores are available at kurate.org/prompts.

- **TrueSkill** is a Bayesian rating algorithm originally developed for competitive gaming that estimates a latent skill score for each paper based on the full sequence of match outcomes. Unlike win rate, TrueSkill accounts for the strength of opponents and the uncertainty in each paper's estimated score, making it more robust to imbalanced matchups.[6]

In their respective samples, these three Kurate ranking methods produce the following cumulative distributions (Figure 3).

Figure 3: Cumulative distributions of Kurate rankings



2.4 Ground-Truth Rankings

SINGLE SCORE RANKINGS

The preferred ground-truth ranking is based on a combination of committee decisions and average reviewer score. Papers are first ranked according to their committee decision tier (Oral > Poster > Reject) and within each decision tier, papers are ranked according to average reviewer score. Note that ties are possible where papers share both a committee decision tier and an identical average reviewer score. In such cases, tied papers are assigned the maximum of their shared rank positions.

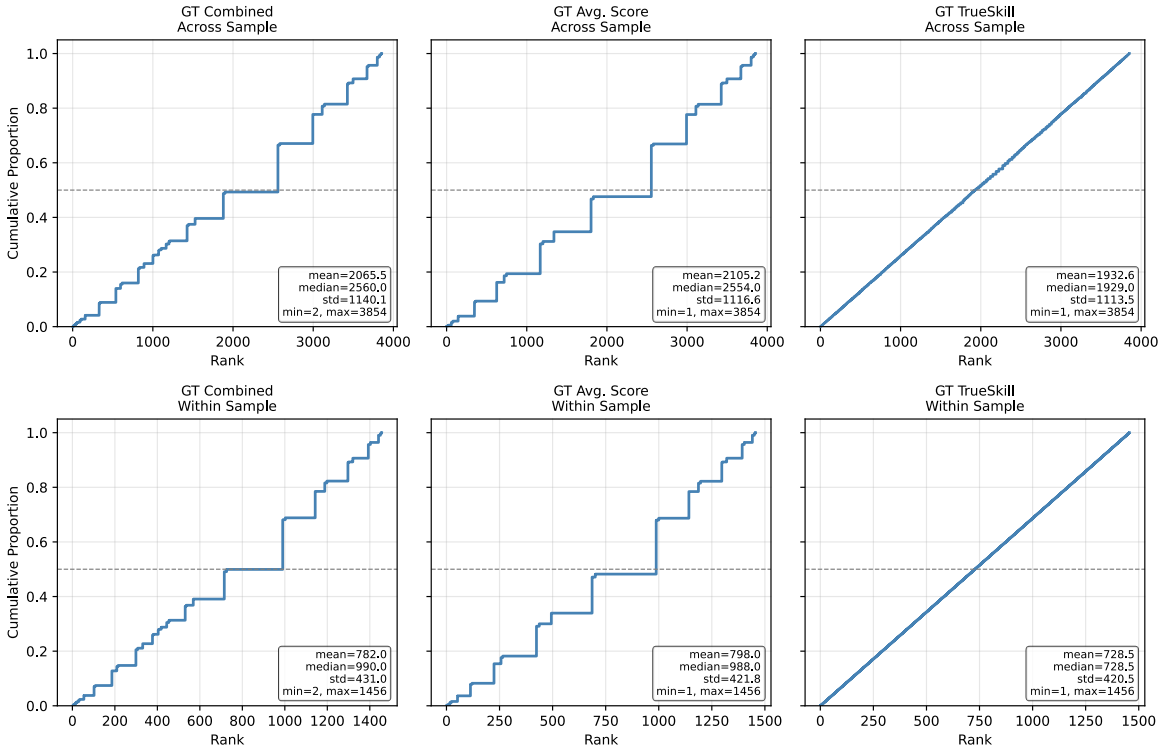
[6] In production, Kurate uses an adaptive matching algorithm based on TrueSkill match quality to increase the information gain per match. For more details, visit https://kurate.org/methodology.

TOURNAMENT RANKING

In addition to the rankings above, we construct a ranking using the TrueSkill algorithm applied to the full set of possible pairwise paper comparisons. Winners of pairwise comparisons are determined using committee decisions, with ties broken by average reviewer scores. Remaining ties are broken using a coin flip, i.e. the winner is chosen by a random draw.

A random draw is preferred over excluding tied pairs from pairwise comparison, since exclusion would introduce a systematic selection bias. Papers with more ambiguous assessments would contribute fewer matches to the ranking, causing their TrueSkill estimates to be based on a less representative and smaller set of opponents, which are potentially easily distinguishable. Random tie-breaking preserves the full match structure and ensures that all papers contribute equally to the ranking, while introducing only unsystematic noise that averages out across a sufficient number of matches.

Figure 4: Cumulative distribution of GT rankings



2.5 Measuring Ranking Performance

In the following, we employ two correlation metrics to measure the similarity between rankings.

- **Kendall's τ** measures the correlation between two rankings by comparing all possible pairs of items and counting how many pairs are in the same order (concordant) versus the opposite order (discordant) across the two rankings. It ranges from −1 (perfect disagreement) to +1 (perfect agreement), with 0 indicating no association.
- **Spearman's ρ** measures the correlation between two rankings by computing the standard Pearson correlation on the rank positions rather than the raw values. Like Kendall's Tau, it ranges from −1 to +1, but it is more sensitive to large rank differences. A single pair of items that are far apart in one ranking but close in another will have a stronger influence on Spearman's ρ than on Kendall's τ.

2.6 Human Benchmark

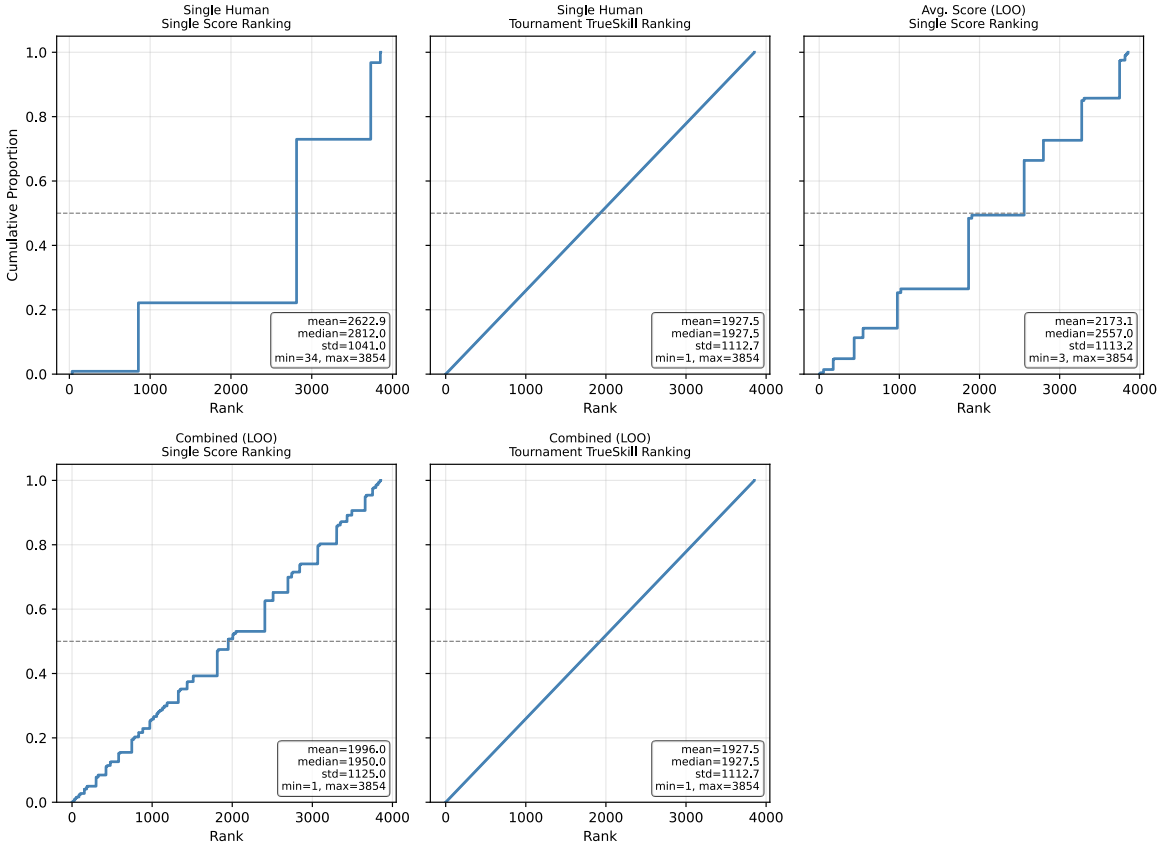
The primary objective of this report is to evaluate the ranking performance of the Kurate methodology relative to rankings derived from human expert reviews. To this end, Kurate rankings are compared directly against ground-truth rankings constructed from ICLR 2026 committee decisions and average reviewer scores. Because these ground-truth rankings aggregate the judgements of multiple reviewers, either through score averaging or through the committee acceptance decision, they reflect a form of expert consensus rather than any individual reviewer's assessment.

To provide an additional meaningful benchmark, we construct a measure of how well a single human reviewer agrees with these consensus-based ground-truth metrics. If Kurate achieves correlations with the ground truth that are comparable to those of an individual expert, this constitutes evidence that the AI-generated rankings approach human-level performance.

We estimate this single-expert benchmark using a leave-one-out (LOO) procedure. Each paper in the dataset is reviewed by multiple experts. For each paper, we randomly draw one reviewer score and set it aside as the "single expert" rating for that paper. The remaining reviewers' scores are averaged to construct a ground-truth ranking for that paper, from which the single expert was excluded. Where the combined GT ranking is used, this average score is then combined with the committee acceptance decision in the usual way. Repeating this across all papers yields two parallel rankings: one based on the randomly drawn single reviewer scores, and one based on the remaining reviewers' consensus. The correlation between these two rankings approximates how closely a single expert tracks the broader consensus, independently of their own

contribution to it. Figure 5 shows the cumulative distribution of the resulting benchmark rankings.

Figure 5: Cumulative distribution of benchmark human rankings, "Across Sample"



It should be noted that the LOO procedure does not fully remove the excluded reviewer's influence from some of the ground-truth rankings. In particular, the committee acceptance decision, which determines the tier ordering in the combined GT ranking, may itself have been influenced by the excluded reviewer's score. As a result, the correlation between the single-expert ranking and the combined GT ranking likely carries a modest upward bias. The benchmark therefore slightly overstates the agreement one would expect from a truly independent single reviewer.

3 Results

Table 3 reports Spearman and Kendall correlation coefficients between Kurate's three ranking methods and the three ground-truth rankings described above, separately for the Within and Across samples. Table 4 reports the correlation between a single human expert's score and the ground-truth rankings under a leave-one-out (LOO) procedure, representing the benchmark agreement one would expect from an individual human reviewer.

Across all ranking methods and ground-truth comparisons, Kurate achieves positive correlations with human expert rankings. Spearman coefficients range from approximately 0.50 to 0.71, and Kendall coefficients from 0.37 to 0.55, indicating a consistent but imperfect correspondence between AI-generated and human rankings.

A notable pattern in Table 3 is the consistently stronger performance of the Single Score method relative to the tournament-based Win Rate and TrueSkill rankings. Single Score achieves Spearman coefficients of 0.65–0.71 depending on the ground-truth ranking and sample, compared to 0.50–0.55 for both Win Rate and TrueSkill. This gap is consistent across all three ground-truth rankings and both samples, suggesting that directly eliciting a quality score from the LLM produces rankings that are more aligned with human judgement than aggregating rankings from pairwise comparisons.

Despite their different underlying algorithms, Win Rate and TrueSkill produce nearly identical correlations with the ground truth in all cases. The difference between the two methods is negligible, suggesting that the choice of aggregation algorithm matters little in terms of ranking performance once the pairwise comparison format is adopted with sufficient matches to achieve stable rankings.

Table 3: Correlations between Kurate and GT rankings

		GT Avg. Score Ranking		GT Combined Ranking		GT TrueSkill Ranking	
Kurate Rankings	**Sample**	**Spearman**	**Kendall**	**Spearman**	**Kendall**	**Spearman**	**Kendall**
Win Rate	**Within**	0.505	0.368	0.537	0.389	0.531	0.372
	Across	0.516	0.379	0.546	0.400	0.543	0.385
TrueSkill	**Within**	0.505	0.365	0.536	0.385	0.531	0.368
	Across	0.522	0.381	0.554	0.402	0.551	0.387
Single Score	**Within**	0.646	0.497	0.707	0.545	0.702	0.523
	Across	0.653	0.504	0.710	0.549	0.706	0.528

Table 4 contextualizes these results by reporting the agreement between a single human reviewer and the ground-truth rankings. A single expert achieves Spearman

coefficients of 0.56–0.68 against the average score ranking and 0.67–0.68 against the combined ranking (combination of committee decision and average score). Single Score Kurate rankings outperform the human benchmark against the GT Average Score ranking by a clear margin (0.65 vs. 0.56) and perform comparably to or slightly above it against the GT Combined Ranking (0.71 vs. 0.68). Win Rate and TrueSkill fall somewhat below the human benchmark in both cases.

As noted above, the LOO benchmark may carry a modest upward bias, because the excluded reviewer likely influenced the committee decision that enters the combined GT ranking. The fact that Kurate's Single Score method matches or exceeds this benchmark despite the structural advantage held by human reviewers in this comparison therefore represents a stronger result than the raw figures suggest. Were the benchmark based on a truly independent external reviewer, the margin in Kurate's favor would likely be larger. This effect is reflected in the particularly large performance gap between single expert rankings and Kurate Single Score rankings when using the average score ranking as the ground-truth reference.

Table 4: Correlations between single expert and aggregated GT rankings

	Single Expert SS vs Avg. Score Ranking SS (LOO)		Single Expert SS vs Combined Ranking SS (LOO)		Single Expert TS vs Combined TS (LOO)	
Sample	**Spearman**	**Kendall**	**Spearman**	**Kendall**	**Spearman**	**Kendall**
Within	0.562	0.481	0.672	0.563	0.663	0.479
Across	0.569	0.486	0.677	0.568	0.679	0.493

The correlation coefficients above capture aggregate correspondence across the full ranking. In addition, the following figures examine whether this correspondence holds consistently from the highest- to the lowest-ranked papers. In particular, Figure 6 presents the distribution of ICLR committee decisions across deciles of Kurate's three ranking methods, where P1 represents the 10 percent highest-ranked papers and P10 the lowest-ranked. A well-calibrated ranking should show a monotonic decline in acceptance rates from left to right, with high-ranked papers more likely to be accepted and low-ranked papers more likely to be rejected.

All three Kurate methods display the expected pattern: acceptance rates are highest in P1 and decline broadly toward P10. This further confirms the above findings that Kurate's rankings carry meaningful signal about paper quality across all three methods. The bottom panel for the Kurate Single Score ranking shows the sharpest and most consistent monotonic decline across deciles. P1 contains a substantial share of Poster and Oral acceptances, and the transition toward rejection is relatively smooth. This is consistent with the correlation results in Table 3, where Single Score achieved the

highest correspondence with human ground-truth rankings. The top two panels display broadly comparable patterns to Single Score, but with a less smooth gradient.

Across all methods, the top 10% decile consistently stands out as the decile with the highest concentration of papers accepted for oral presentation, suggesting that Kurate reliably identifies a core group of high-quality papers regardless of the ranking algorithm used.

Figure 6: Concordance of Kurate rankings with GT committee decisions

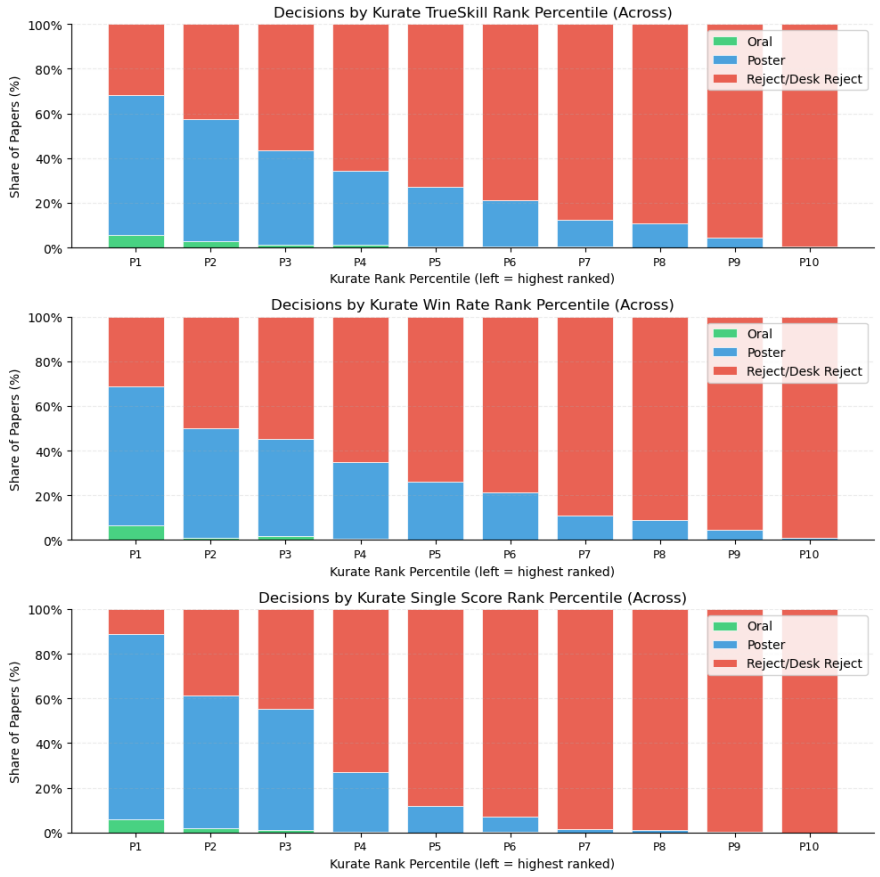
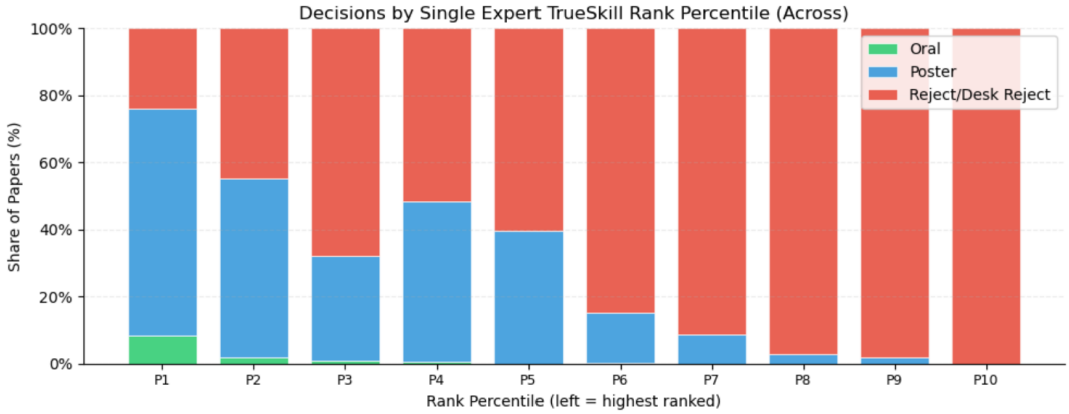


Figure 7 displays the same decile breakdown for a single human expert reviewer, serving as a benchmark for the level of agreement one would expect from an individual human. The similarity in overall pattern reinforces the aggregate correlation results and reiterates the conclusion that Kurate Single Score ranking outperforms a random single human expert reviewer. Unlike the Kurate figures, the single expert decile breakdown does not show a strictly monotonic decline in acceptance rates across all deciles, suggesting that individual human reviewers provide a somewhat less consistent quality signal than the Kurate rankings.[7]

[7] Note that the non-monotonic pattern in acceptance rates across deciles in 2026 should not be overinterpreted. This pattern does not appear to be systematic and is not observed in the analyzed samples of the 2025 dataset.

Figure 7: Concordance of single human rankings with GT committee decisions



4 Robustness Checks

The main results are subject to several potential concerns: that the findings are specific to the particular composition of the 2026 ICLR dataset and do not generalize across research subfields; that they reflect idiosyncrasies of the 2026 conference cohort rather than a robust property of the Kurate methodology; and that the pairwise ranking results depend on the number of matches conducted per paper, such that a higher match count might substantially alter the performance of these rankings compared to their single-item counterpart. The following checks address each concern in turn.

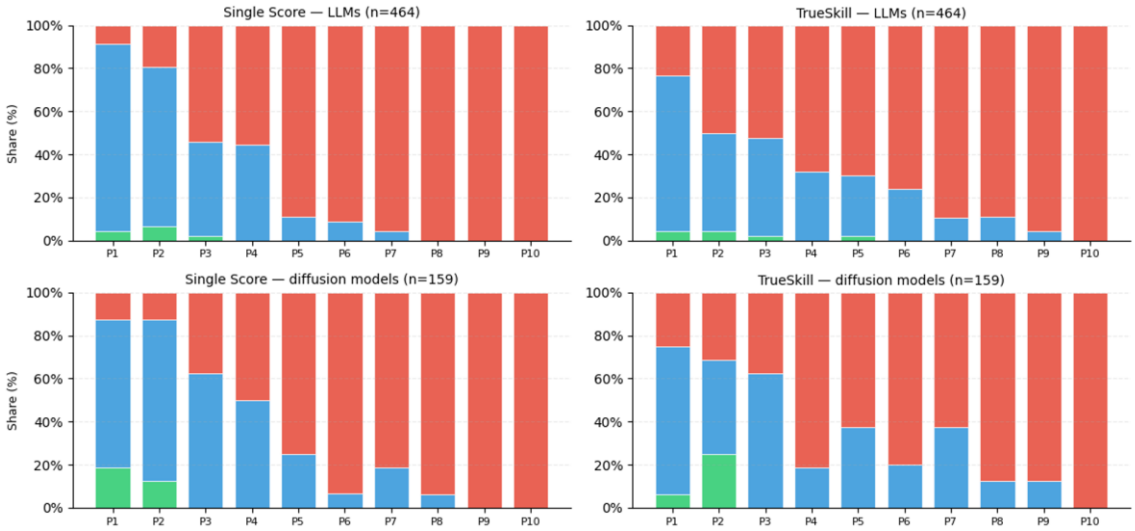
4.1 Label-Specific Results

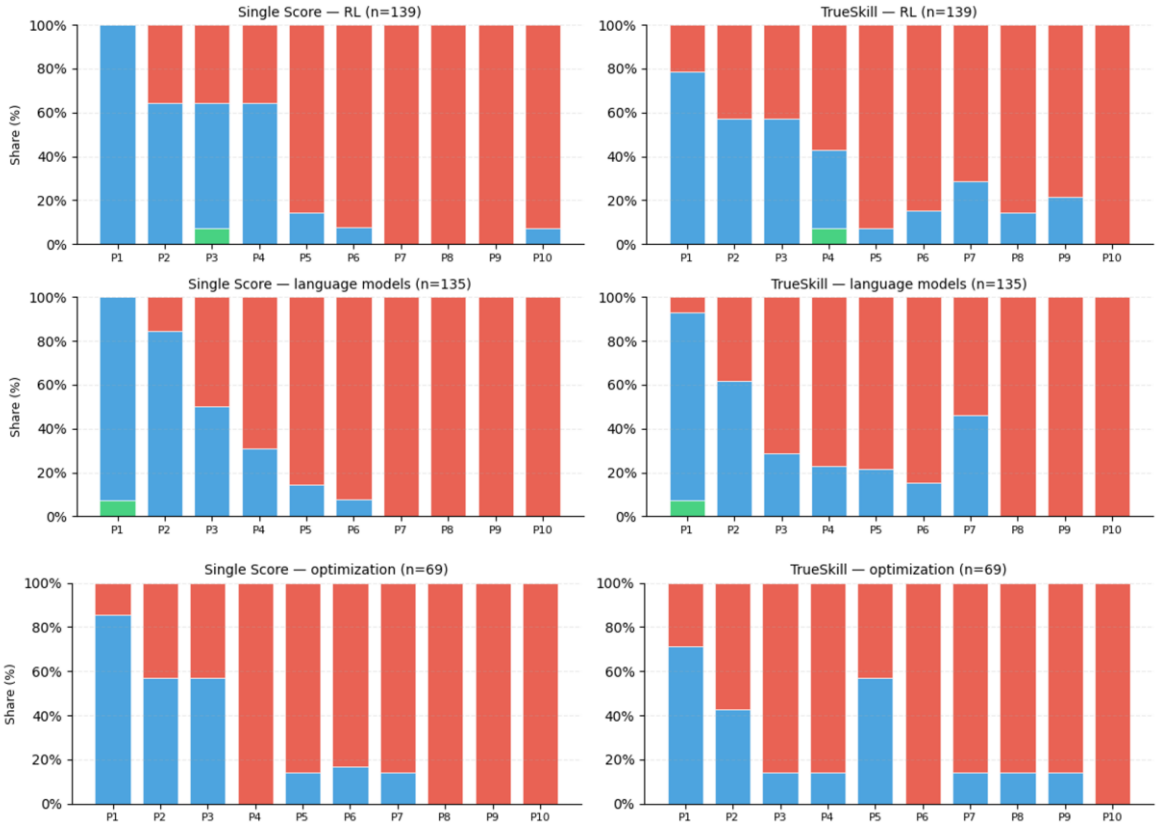
Figure 8 disaggregates the above decile analysis by research field, assessing whether Kurate's ranking performance is consistent across different areas of machine learning research.

High acceptance rates in P1 declining toward near-zero acceptance in P9–P10 are present across all subfields and both ranking methods. The Single Score Kurate ranking consistently displays a steeper and more monotonic decline than TrueSkill, in line with the aggregate results. The advantage of Single Score over the tournament-based TrueSkill ranking therefore holds across subfields, rather than being driven by any particular research area.

Variability in the middle deciles is more pronounced in smaller subfields such as optimization (n=69), where each decile contains fewer papers and individual placement decisions carry more weight. This is an expected consequence of sample size rather than a systematic weakness in the rankings. Across the larger subfields, the gradient is notably smoother and the ranking signal more apparent.

Figure 8: Concordance between Kurate rankings and committee decisions, by label





4.2 2024/2025 Sample

Figure 9 and Figure 10 provide an out-of-sample validation of Kurate's ranking performance by comparing results across two distinct conference cohorts. The left panels show decile acceptance distributions based on a sample of papers from ICLR 2024/25 and topic labels “LLM” and “Optimization”, while the right panels show the corresponding results for the 2026 sample used throughout the main analysis.

The key finding is that the decile patterns are broadly consistent across the two cohorts. In both years and for both TrueSkill and Single Score, acceptance rates are highest in P1 and decline toward near-complete rejection in the lower deciles. This consistency across conference years suggests that Kurate's rankings are not specific to the 2026 dataset used in the main analysis, but reflect a more general ability to distinguish paper quality that persists across cohorts.

Importantly, the advantage of Single Score over the tournament-based TrueSkill ranking is visible in both the 2024/25 and 2026 samples. In both cohorts, Single Score displays a steeper and more consistent decline across deciles, with a higher concentration of accepted papers in the top deciles relative to TrueSkill. This cross-cohort replication strengthens the main conclusion that Single Score is the preferred ranking method, suggesting it is not a finding specific to the 2026 data but a robust feature of the two approaches.

Figure 9: Concordance between Kurate and GT, 2024/2025, "LLM" label

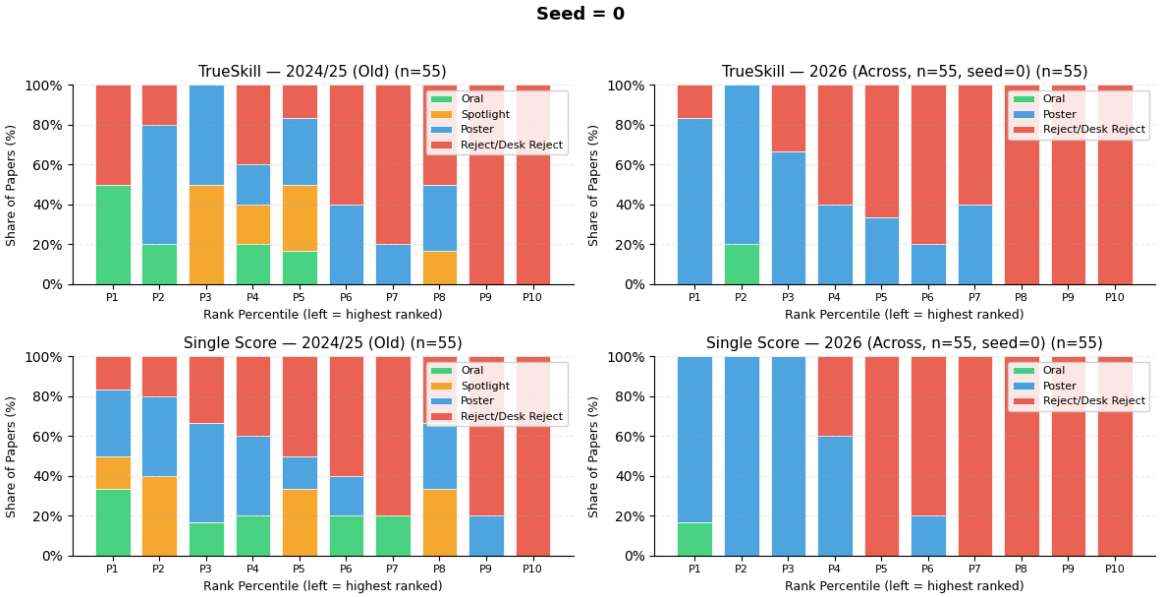
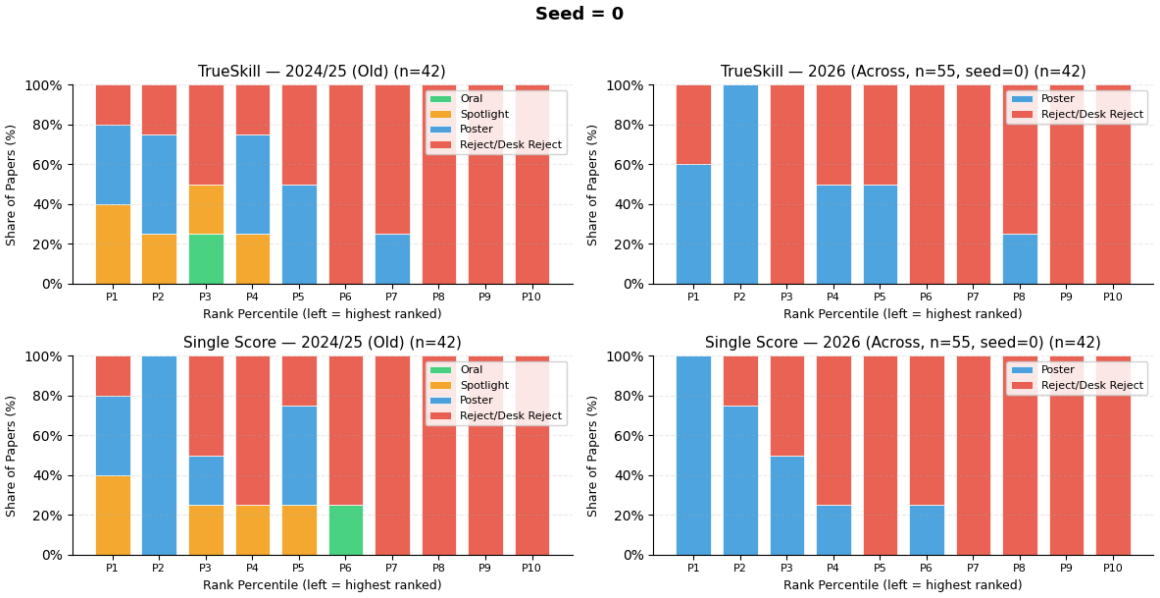


Figure 10: Concordance between Kurate and GT, 2024/2025, "Optimization" label

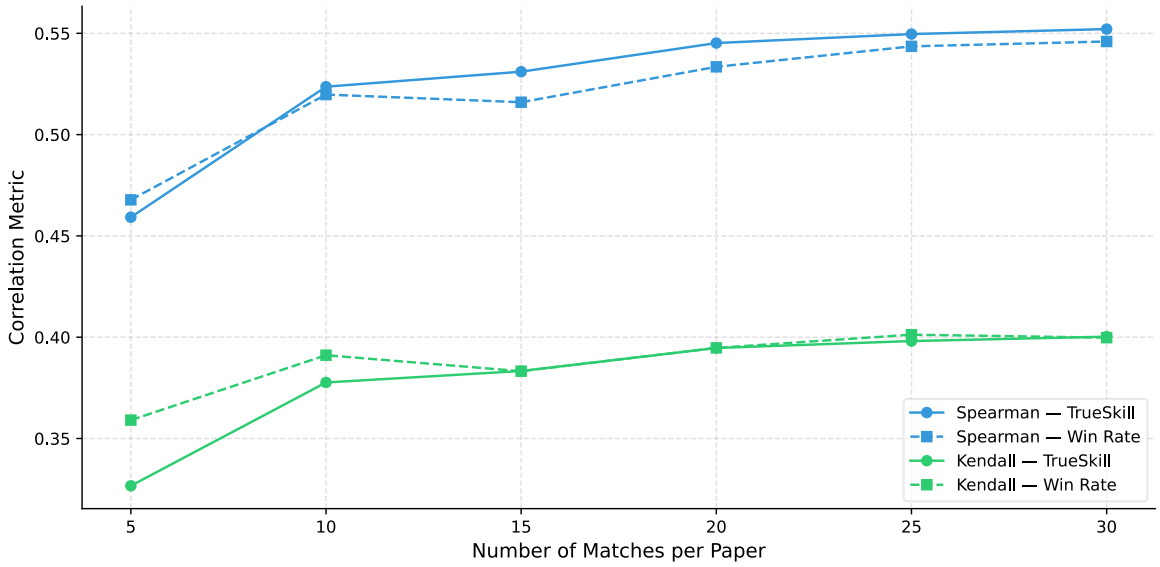


4.3 Convergence of Pairwise Kurate Rankings

A potential concern for the pairwise ranking methods is that the correlation results depend on the number of matches conducted per paper. If rankings based on 30 matches per paper differ substantially from those based on fewer matches, this would suggest that the results reported in the main analysis are sensitive to this design choice, and that expanding the methodology to even higher match counts might yield materially different conclusions, in particular regarding the relative performance of the pairwise rankings compared to the Single Score ranking.

Figure 11 examines this by plotting the correlation between Kurate's pairwise rankings and the combined ground-truth ranking as a function of the number of matches per paper, ranging from 5 to 30. Both Spearman and Kendall correlations are shown.

Figure 11: Convergence of Kurate Rankings by Number of Matches per Paper



The results indicate that both methods converge relatively quickly. The largest gains in correlation occur between 5 and 10 matches per paper, after which the improvement flattens considerably. Beyond 15 matches, additional comparisons yield only marginal gains, and correlations at 30 matches are close to those already achieved at 20. This pattern holds for both TrueSkill and Win Rate and for both correlation metrics. This is somewhat surprising, but may be due to the fact that the matches were drawn randomly rather than using TrueSkill's own match quality metric, as is the case for Kurate's production version of the TrueSkill ranking. The results reported in the main analysis, which are based on 30 matches per paper, therefore reflect a stable ranking result rather than a point of ongoing improvement, and are unlikely to change substantially with additional matches.

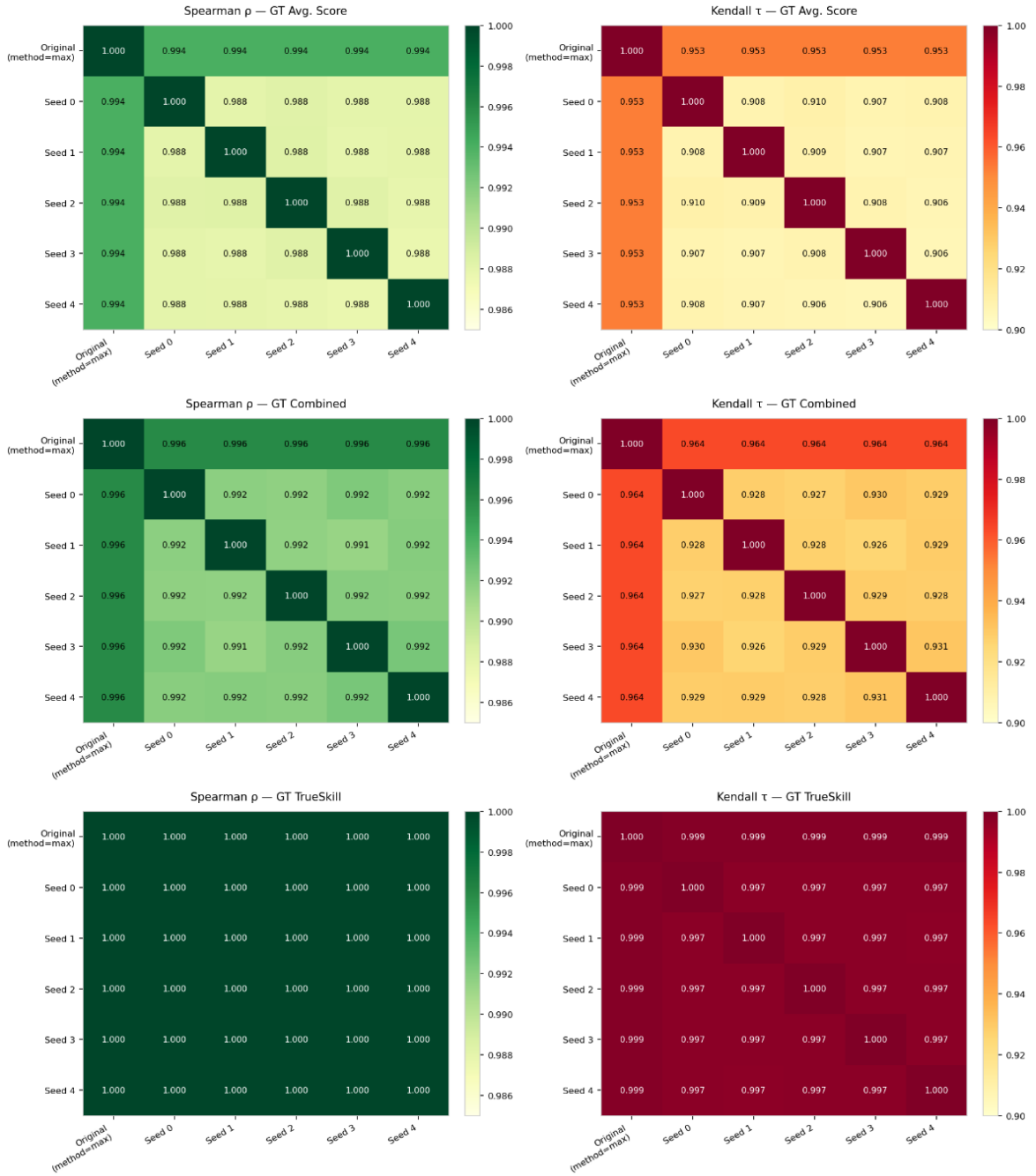
4.4 Tie-Breaking in Ground-Truth Rankings

The preferred ground-truth rankings based on discrete reviewer scores in some instances are substantially coarser than Kurate rankings. When a granular ranking is correlated against a coarse one, the maximum achievable correlation may be bounded below 1, not due to disagreement, but simply due to the mismatch in granularity.

Figure 12 provides an empirical approximation of this ceiling by reporting correlations between ground-truth variants with randomly shuffled ties. For GT TrueSkill, which produces continuous scores and is effectively tie-free, all correlations are at or near 1 and no structural ceiling applies. For GT Average Score and GT Combined, Spearman

correlations between shuffled variants remain high at ρ ≈ 0.988–0.992, suggesting that the Spearman results in Table 3 can be interpreted largely at face value. Kendall correlations between these shuffled variants are lower at τ ≈ 0.906–0.931, indicating a more meaningful ceiling on the Kendall results. This is expected: unlike Spearman, which is based on squared rank distances, Kendall's τ counts pairwise ordinal agreements directly, making it inherently more sensitive to the local swaps that random tie-breaking introduces. As a result, part of the gap between the observed τ values in Table 3 and the theoretical maximum of 1 may reflect the coarseness of the ground-truth data rather than genuine disagreement between Kurate and expert rankings.

Figure 12: Effects of tie-breaking in ground-truth rankings



5 Discussion

The results presented above support several conclusions about the validity and practical utility of Kurate's AI-generated paper rankings.

- **Kurate successfully distinguishes paper quality:** Across all ranking methods, ground-truth comparisons, and research fields, Kurate consistently tends to assign higher ranks to papers that are ultimately accepted to ICLR 2026 and lower ranks to papers that are rejected from ICLR 2026. This holds both in aggregate and within individual subfields, suggesting that the signal Kurate extracts from LLM-based evaluations is generalizable rather than an artefact of a particular comparison set or subject area.
- **Performance exceeds individual human reviewers**: The correlation between Kurate's rankings and ground-truth rankings is broadly in line with, and for the best-performing method above, the agreement observed between single human experts and the ground truth. This implies that Kurate does not merely produce plausible-looking rankings, but does so at a higher level of accuracy than that one would expect from qualified human reviewers. Notably, this comparison is conservative, since the human benchmark carries an upward bias from the LOO procedure. Accounting for this, Kurate's performance relative to a truly independent human reviewer is stronger than the raw figures suggest. Given that Kurate operates at a fraction of the time and cost of human reviewers, this represents a compelling benchmark for automated paper evaluation.
- **Single Score is the preferred ranking method:** Among the three methods evaluated, the Single Score ranking consistently achieves the strongest correspondence with human ground-truth rankings, both in terms of aggregate correlation coefficients and the steepness of the decile acceptance gradient. In several comparisons, Single Score meets or exceeds the human benchmark, performing at least as well as individual expert ratings when evaluated against the preferred ground-truth measure. This is a notable result, as it suggests that directly eliciting a quality judgement from the LLM is more effective than aggregating outcomes from pairwise comparisons.
- **Pairwise methods are not recommended as a starting point:** Despite the intuitive appeal of pairwise comparison approaches, which avoid the need for absolute quality scores, the results provide no evidence that tournament-based Win Rate or TrueSkill rankings offer advantages over Single Score rankings. Both pairwise methods perform consistently below Single Score across all conditions, with no subfield or ground-truth comparison where they clearly outperform it. Based on the evidence presented in this report, Single Score should be the preferred Kurate ranking method. Pairwise approaches may require further development or a compelling application-specific rationale to justify their use.

Several methodological limitations should be considered:

- **The ground-truth rankings are themselves noisy:** The above analysis treats committee decisions and average reviewer scores in the ICLR dataset as a reliable proxy for paper quality, but these are likely imperfect measures. This is particularly the case if the human review process is subject to biases (seniority, familiarity with subfield, author reputation). The correlations reported here should therefore be interpreted as measuring agreement with an observed expert consensus rather than with an objective quality standard. To the extent that the observed expert consensus is itself noisy or biased, Kurate's true performance in terms of identifying paper quality may be even higher than the reported figures above suggest.
- **Sampling variability:** Two sources of randomness are present in the analysis. The human benchmark draws a single reviewer score per paper at random, and the tournament-based ground-truth rankings resolve ties via coin flip. This means that the results presented above to some degree reflect a single random draw rather than an expected value. Repeating the analysis over many draws and reporting confidence intervals, would quantify this uncertainty and allow more precise comparisons between Kurate and the human benchmark.
- **Stability of pairwise rankings:** A further concern could be whether the pairwise ranking results are sensitive to the number of matches conducted per paper. The convergence analysis above indicates that this is not a material concern. Correlations between pairwise Kurate rankings and the ground-truth rankings stabilize after approximately 15–20 matches per paper, with only marginal gains observed beyond that point. The main results, based on 30 matches, are therefore unlikely to be meaningfully affected by the choice of match count, and increasing the number of matches further would not be expected to alter the conclusions presented above.
- **External validity:** The present validation draws exclusively on ICLR 2026, a top-tier ML conference with a specific submission culture and reviewer pool. It is not self-evident that findings generalize to the broader range of fields Kurate covers, including Economics, Game Theory, and Robotics, where methodological standards are more heterogeneous and LLM familiarity with subject matter may differ. Furthermore, conference acceptance conflates novelty and presentation quality with longer-term scientific impact, the dimension Kurate explicitly aims to capture. Extending the validation to additional conferences and fields, and exploring alternative quality proxies, would substantially strengthen the generalizability of the above validation results.