Pair Difficulty Matters: Rethinking Pairwise LLM-as-a-Judge Evaluation and Consistency

Anonymous ACL submission

Abstract

Large Language Model judges are widely used to rank texts and text-generating systems through pairwise comparison, and their reliability is typically assessed via three proxies: position bias, transitivity, and pairwise agreement (self- or human-labeled). Because these proxies drive judge selection and benchmarking, a substantial literature reporting that judges perform poorly on them risks steering practitioners away from otherwise capable evaluators. We argue this assessment is misleading. Under the Bradley–Terry geometry underlying pairwise aggregation, each proxy is dominated by close-rank-gap pairs, where inconsistency is information-theoretically expected and individual verdicts contribute little to the aggregate ranking; far-gap pairs carry the ranking signal but barely move the proxies. We formalize this argument and validate it in a controlled simulation and on two human-rated corpora: the proxies correlate only weakly with ranking accuracy against gold, and their predictive component concentrates in the far-gap regime. Judges should therefore be assessed on rank-gap-conditional metrics, ideally against human rankings. Code at PLACEHOLDER.

1 Introduction

LLM-as-a-judge has become a central paradigm for evaluating generated text across many domains (Gu et al., 2025). However, a growing body of work has documented systematic problems with LLM-judges that fall broadly under the heading of inconsistency, and several distinct forms have been identified. Models exhibit run-to-run inconsistency, returning different verdicts across repeated queries under identical or near-identical settings. They exhibit position bias — by far the most-discussed failure mode — preferring one slot of a pairwise comparison at rates well above chance. They also exhibit non-transitivity at the level of the induced preference graph.

A separate line of work has compared LLM verdicts not to internal consistency criteria but to individual pairwise judgments, focusing on human–LLM divergence at the pair level (Fein et al., 2026; Zheng et al., 2023; Li et al., 2023, e.g.). These two strands — internal inconsistency on the one hand, and disagreement with humans on the other — are typically treated as distinct problems. However, the comparisons that dominate aggregate inconsistency metrics are often the comparisons that matter least for ranking recovery. We argue that they are closely connected: They are both affected by pair gap / difficulty, and because of this, are both unreliable proxies of judge quality when it comes to recovering human-derived rankings.

2 Related Work

Swapping two responses can flip a judge's verdict — a phenomenon documented early on by Wang et al. (2024) and Zheng et al. (2023), and usually addressed by balanced-position aggregation. Shi et al. (2025) provide the most thorough analysis to date and find that the gap between texts is a key factor: pairs close in quality are affected far more strongly than distant ones.

Xu et al. (2025) document non-transitivity in pairwise rankings and show that, like position bias, it concentrates on close pairs; they mitigate it through round-robin tournaments and Bradley-Terry (BT) aggregation. Wang et al. (2025) target inconsistency directly and document that 85–90% of transitivity violations are tie-driven and that judge capability does not monotonically reduce inconsistency. SAGE (Feng et al., 2025) evaluates judges without requiring human-labeled comparisons around exactly these two axes and finds that even top models fail on roughly a quarter of difficult cases. All of this work treats inconsistency as a defect that needs to be reduced rather than asking what its structure reveals about the judge.

Zheng et al. (2023) measure pair-level agreement between LLM judges and humans directly and find that this agreement is gap-conditional: the residual disagreement concentrates on close pairs where humans also disagree. Thakur et al. (2025) argue that aggregate alignment metrics are misleading proxies for judge usefulness: judges with substantially lower percent agreement can still produce ranking correlations with humans that are nearly identical, because consistent biases preserve relative ordering even when they distort absolute scores. We extend this line of argument along an orthogonal axis. While Thakur et al. decompose by use case (scoring versus ranking), we decompose by the rank gap of the items being compared, and show that aggregate inconsistency is dominated by close-pair behavior that a calibrated judge often produces.

3 Theory

We aggregate pairwise verdicts under the BT model (Bradley and Terry, 1952), the most robust ranking algorithm for most use cases (Daynauth et al., 2025). Under BT, each item i carries a latent strength θ_i, and the probability that i is preferred to j is $\Pr(i \succ j) = \sigma(\theta_i - \theta_j)$, where $\sigma(x) = (1 + e^{-x})^{-1}$ is the logistic function. In a pair with a true strength gap $\Delta = \theta_i - \theta_j$, the two possible outcomes have logarithmic probabilities $\log \sigma(\Delta)$ and $\log \sigma(-\Delta)$, which differ by exactly $|\Delta|$. Close pairs are near-coinflips, and the two outcomes are nearly equally consistent with the model; far pairs are lopsided, and one outcome is much more consistent than the other. A flipped verdict therefore changes that pair's log-likelihood contribution by $|\Delta|$:

- **Close pair** ($\Delta \approx 0$): both outcomes have nearly the same likelihood, so a flip is nearly free under the model and exerts little pull on $\hat{\theta}$.
- **Far pair** (large $|\Delta|$): the two outcomes differ in log-likelihood by $|\Delta|$, so a flip is strong evidence against the current $\hat{\theta}$ and pulls the MLE toward revising it.

This separates two failure modes that aggregate metrics conflate. Close-pair *self-inconsistency* — the judge flipping its own verdict across runs — is Bayes-optimal noise: the BT model itself assigns the two outcomes near-equal probability, so a flip is consistent with the model and the recovered ranking barely moves. Far-pair self-inconsistency, by contrast, is where calibration error becomes visible: the model strongly predicts one outcome, so absorbing a flip in the other direction forces $\hat{\theta}$ to revise what the latent strengths must be. Noise of the first kind is irreducible and does not systematically bias the recovered ranking; error of the second kind does. The same $|\Delta|$-weighting applies to judge–human *disagreement*: when the judge's verdict diverges from the human verdict, the induced shift in $\hat{\theta}$ is again largest in the high-$|\Delta|$ regime, while close-pair disagreement is bounded below by the noise human annotators themselves cannot avoid. Pair-level agreement metrics — Krippendorff's α, raw agreement rates, position-flip rates — weight every pair equally and therefore systematically overweight the regime that contributes least to ranking distance from gold, whether the comparison is judge-vs-judge or judge-vs-human.

This predicts the empirical pattern: aggregate reliability and position-bias scores should correlate weakly with ranking accuracy against human gold, with predictive signal concentrated in the far-pair tail. It also clarifies prior findings: position bias is empirically strongest on close pairs (Shi et al., 2025), exactly where its information cost is lowest, and human annotators themselves disagree most on close pairs (Zheng et al., 2023) — as BT predicts. Judges and humans agree on what the hard cases are; the question this paper takes up is whether the measures recognize that the hard cases are also the cheap ones.

4 Simulation: Human Agreement

To isolate the effect of disagreement distribution independently of model-specific behavior, we construct a controlled simulation in which judges differ only in where across the rank-gap spectrum their errors occur.

We simulate a dataset with normally distributed BT strengths and generate human comparisons from randomly drawing the winner of each comparison by using BT-probability. We then generate judges with identical agreement with the human pairs. The judge flips with a gap-dependent probability: it coin-flips (0.5) on pairs below its resolution threshold τ ("too close to call"), and above τ it errs at rate floor.

As Fig. 1 shows, agreement between simulated judges and humans is roughly equal across different floors — the only difference is that some judges agree less with humans on distant pairs, while oth-

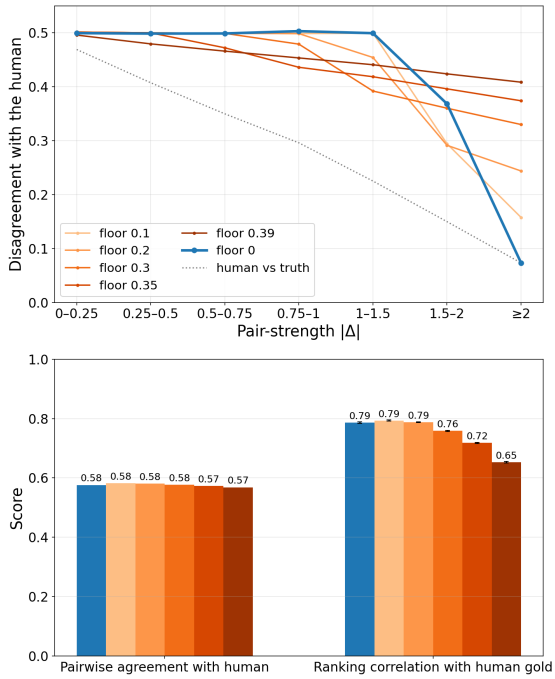

Figure 1: Simulation results for BT Judges with similar pairwise agreements. Top shows how disagreement between judges and humans is differently distributed, while overall agreement is roughly equal (bottom left). Despite this, ranking correlation with the human gold varies substantially depending on where disagreement occurs (bottom right).

ers more on close pairs. Despite that, agreement with gold falls sharply as judge errors increase at higher $|\Delta|$s, even if these errors are counterbalanced by higher accuracy on close pairs. Therefore, pairwise-level agreement is not the same as ranking correlation due to the spread of disagreement.

5 Experiment: Model Inconsistencies

Datasets CLEAR (Crossley et al., 2021) is a dataset of English school text excerpts rated on comprehensibility by teachers. The human rating was performed pairwise: teachers were shown two texts and asked to pick which one is easier for students to understand. We use 200 randomly selected texts. ASAP 2.0 (Crossley et al., 2025) is a corpus of student writing tests for grades 6 to 10, where students were tasked to argue for a position on a topic using a source text on that topic. The essays are human-rated on a scale of 1-6. We use 200 student essays from the test set on the topic of algorithmic facial emotion detection for grade 10.

Ranking human-labeled corpus texts rather than model outputs provides a clean gold standard for pairwise rank gaps, which is unavailable when the ranked items are the generating models themselves.

Position Bias We evaluate GPT-5.4-nano and -mini, Ministral-3-3B and -14B, as well as Gemma-3-4B and -12B (see App. A) on both the CLEAR and the ASAP 2.0 corpus. Each pair is evaluated twice with reversed answer orderings. We measure position bias using two standard metrics: (1) the consistency, i.e., the frequency with which the same verdict is retained after position reversal; and (2) inconsistent-pair primacy (IPP), the proportion of inconsistent pairs in which the judge prefers the first-shown response. An IPP of .5 indicates that inconsistent verdicts split evenly between positions.

Dividing positional consistency into different $|\Delta|$-quartiles shows that position bias exists mostly on close pairs. Given that a flipped rating produces a tie, we can say that this is the model being indecisive about very close pairs ("too close to score").

Crucially, consistency, IPP, and rank correlation with gold do not track each other (Tables 1 and 2). On CLEAR, the two GPTs flip at indistinguishable rates[1], yet the better-correlating one wins significantly on ρ. On ASAP the pattern reverses: nano is significantly more consistent while the two rank the field equally well; Ministral behaves similarly. Only Ministral on CLEAR aligns bias and correlation in the expected direction — evidence that aggregate bias metrics are weak predictors of ranking performance, not that bias is benign.

Gemma-3-4B is the limiting case: almost entirely indecisive on close pairs, yet its far-pair judgments suffice to recover a ranking comparable to much more consistent judges. Its ceiling is naturally capped — it cannot resolve the "middle pack" — but the bias metric overstates the damage. The mild top-quartile inconsistency present in every model points instead to genuine conceptual misalignment: some pairs humans rank far apart are treated by the model as close. Fixing that is the more promising lever for correlation gains.

Non-Transitivity We analyze non-transitivity only on triangles, distinguishing (1) strict ($A \prec B$, $B \prec C$, $C \prec A$), (2) mixed ($A \succ B$, $B \succ C$, $C \sim A$), and (3) inequality ($A \sim B$, $B \sim C$, $A \not\sim C$). Given the close-pair indecision documented above, inequality non-transitivity can be substantively valid: if $|\Delta|$ between A and C is large enough to score, the triangle is consistent with the model's own behavior.

[1] All pairwise significance claims use McNemar's test at $\alpha = 0.05$.

	Judge	Consistency (1−flip)			IPP	% intransitive triads			ρ
		all	close (Q1 gap)	far (Q4 gap)	all	cyc	mix	eq	
CLEAR	GPT-5.4-nano	.78	.72	.90	.61	0.0	0.8	9.9	+0.691
	GPT-5.4-mini	.80	.66	.96	.97	0.0	0.4	9.5	+0.825
	Ministral-3-3B	.79	.70	.92	.89	0.0	0.4	6.2	+0.755
	Ministral-3-14B	.84	.73	.97	.89	0.0	0.8	4.1	+0.788
	Gemma-3-4B	.22	.09	.46	1.00	0.0	0.0	27.2	+0.676
	Gemma-3-12B	.76	.57	.96	.99	0.0	0.0	7.8	+0.830
ASAP	GPT-5.4-nano	.79	.74	.85	.21	0.0	4.9	10.7	+0.677
	GPT-5.4-mini	.73	.62	.84	.03	0.0	0.0	13.8	+0.674
	Ministral-3-3B	.78	.65	.94	.02	0.0	0.4	7.1	+0.726
	Ministral-3-14B	.83	.77	.94	.02	0.0	0.0	3.6	+0.690
	Gemma-3-4B	.34	.19	.53	1.00	0.0	0.0	29.3	+0.668
	Gemma-3-12B	.84	.78	.95	.86	0.0	0.9	4.9	+0.735

Table 1: Inconsistency as a close-pair phenomenon does not track ranking accuracy. Consistency is how often the judge chooses the same text across the swapped pairs; IPP is inconsistent-pair primacy, with .5 indicating perfect balance. CLOSE/FAR are the same rate restricted to the bottom/top quartile of the gold gap $|g_a - g_b|$. Triad-% is the proportion of all triads. ρ is the Spearman correlation of the LLM-derived BT skill against human gold.

Across all judges, non-transitivity is dominated by the inequality type and tracks consistency tightly (Pearson $r = .928$): it is largely a re-expression of the indecision already captured by position bias. Within each family the larger model produces fewer non-transitivities, so the metric mainly tracks model size — generally, but not reliably, a proxy for judge quality. Despite far higher non-transitivity, Gemma-3-4B does not correlate significantly worse with human gold than GPT-5.4-nano (Table 2), confirming that intransitivity is a weak model-selection signal.

Proxy	ρ_s vs ranking ρ	p_{Holm}
Consistency, all	+0.559	0.198
Consistency, close	+0.245	0.892
Consistency, far	+0.897	0.001*
Primacy (IPP)	+0.126	0.892
Intransitivity	−0.608	0.174

Table 2: Spearman correlation between inconsistencies and the model's ability to replicate human gold ranking, and the associated Holm-corrected significance. With N=12 the test is well-powered for strong effects but has limited power against moderate effects. We note this as a limitation but emphasize the practical implication: a metric whose correlation with gold ranking is too weak to survive correction at this scale is too noisy to be used as a model-selection signal in practice.

6 Conclusion

Internal consistency and pairwise agreement with humans are useful judge properties, but the central question is whether a judge recovers the same latent ranking structure as humans — so the most direct evaluation criterion is the agreement between human and LLM rankings, and imperfect local consistency should not be treated as disqualifying. Aggregate inconsistency metrics such as position bias should therefore be read in light of pair difficulty: a metric that weights all pairs equally rewards judges that excel on low-information comparisons and penalizes those better at the far ones. Where bias mitigation does help is in making an already well-aligned judge more decisive on close pairs, improving ranking granularity.

For benchmarks, two implications follow. Judges perform far better when item gaps are large — the mechanism exploited by ArenaHard (Li et al., 2024) — so items should be organized into rankings rather than isolated pairs wherever possible. Where full rankings are infeasible, pair difficulty should be made observable by other means, e.g. multiple human raters per pair with disagreement reported as a proxy. This reframing opens a more productive research agenda: instead of chasing ever-lower bias scores, the next generation of judge evaluations should target the regimes where ranking signal actually lives.

7 Limitations

Our rank-gap analysis requires a notion of latent strength on each corpus. CLEAR comes with BT-scores pre-computed. On ASAP 2.0 we use pointwise-averaged human ratings as a strength proxy; this is a noisier estimator of $\hat{\theta}^{\text{human}}$ than pairwise-derived strengths and weakens within-bucket contrasts, so the reported pattern on ASAP 2.0 is a lower bound.

We restrict our attention to pairwise judging. The theoretical argument is specific to binary comparison, and listwise judging introduces additional position-bias structure (Shi et al., 2025) that the present framework does not address.

Finally, our experiments assume the goal of LLM-as-a-judge is to recover a human gold standard that is usually somewhat subjective. Human raters are themselves imperfect estimators of any underlying quality, and what they rate may diverge from what we actually care about estimating — on CLEAR, for instance, teacher ratings of text comprehensibility are not the same construct as comprehension measured directly from student readers. In settings with an objective gold and well-separated candidates, such as verifiable math or code tasks with skill-stratified models, the close-pair regime our analysis emphasizes becomes sparse and the practical importance of the misweighting effect diminishes, though the underlying information theory remains unchanged.

8 Use of AI Assistants

AI writing assistance was used for sentence-level paraphrasing and polishing of author-written text; it was not used to generate research ideas, technical claims, related-work positioning, or citations. AI-assisted coding tools used during analysis script development are documented in the released code repository's README.

References

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.

Scott Crossley, Aron Heintz, Joon Choi, Jordan Bachelor, Mehrnoush Karimi, and Agnes Malatinszky. 2021. The CommonLit Ease of Readability (CLEAR) Corpus. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM)*.

Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A large-scale corpus for assessing source-based writing quality: ASAP 2.0. *Assessing Writing*, 65:100954.

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.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2026. LitBench: A Benchmark and Dataset for Reliable Evaluation of Creative Writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7740–7755, Rabat, Morocco. Association for Computational Linguistics.

Yuanning Feng, Sinan Wang, Zhengxiang Cheng, Yao Wan, and Dongping Chen. 2025. Sage: A Scalable Framework for Evaluating LLM-as-a-Judge Without Human Effort.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative Judge for Evaluating Alignment. *Preprint*, arXiv:2310.05470.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and BenchBuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

OpenAI. 2026. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. *Preprint*, arXiv:2406.07791.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin,

Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xinyu Dai, Yue Zhang, Cunxiang Wang, Wei Ye, and Shikun Zhang. 2025. TrustJudge: Inconsistencies of LLM-as-a-Judge and How to Alleviate Them. In *The Fourteenth International Conference on Learning Representations*.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating Non-Transitivity in LLM-as-a-Judge. *Preprint*, arXiv:2502.14074.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Preprint*, arXiv:2306.05685.

A Models

We evaluate six judges spanning three model families and two size tiers per family. Table 3 summarizes parameter counts, licenses, and access modalities. All model use is consistent with the intended research and evaluation use stated by each provider. We focus on smaller models (3B–14B parameters) because the variance-limited regime our analysis targets is most pronounced at this scale; larger models exhibit ceiling effects on the corpora we use.

Note that this generation of OpenAI models does not support deterministic output via greedy decoding (which we used for the open weight models), which may slightly inflate the reported inconsistency numbers and prevent exact replication. Nevertheless, these models are state-of-the-art (e.g., mini's top performance on CLEAR) and widely used in production, so we felt it was important to include them in order to situate API-based models within our broader thesis.

Model	Params	Access
GPT-5.4 nano (OpenAI, 2026)	undisclosed	API
GPT-5.4 mini (OpenAI, 2026)	undisclosed	API
Ministral 3 3B (Liu et al., 2026)	3B	Open weights
Ministral 3 14B (Liu et al., 2026)	14B	Open weights
Gemma 3 4B (Team et al., 2025)	4B	Open weights
Gemma 3 12B (Team et al., 2025)	12B	Open weights

Table 3: Judge models.

All open-weights inference was run locally on Apple Silicon using llama.cpp with Q8_0 quantization, which is deterministic under greedy decoding and introduces minimal precision loss relative to FP16. Total wall-clock across the four open-weights judges and two corpora was approximately 7–12 hours.