From Scores to Signs: Pairwise Win-Rate Estimation with Calibrated LLM Judges

Gaurav Sood*

March 2026

Abstract

LLM-as-judge scores are often unreliable as cardinal measurements. Small gaps on a 0–100 scale are hard to interpret, and absolute scales can drift across prompts, domains, and judge variants. Pairwise comparisons discard magnitude information, but that trade can still be attractive when the magnitude is mostly unstable or weakly calibrated. A pairwise win rate directly answers a common evaluation question: how often would policy A beat policy B on a random prompt? This paper formulates a CJE-style estimator for that target using a cheap judge everywhere and a small oracle-labeled slice for calibration. The judge signal can be either a direct pairwise score or an induced pairwise margin formed from pointwise scores. Under conditional mean calibration, averaging calibrated pairwise probabilities recovers the oracle win rate; with sample splitting, the plug-in estimator is consistent. We also discuss symmetry, heterogeneity, and calibration-aware uncertainty. In a reanalysis of the public CJE Arena sample, pointwise judge and oracle scores are only moderately aligned (Pearson $r = 0.390$), while induced pairwise comparisons are materially more useful as discriminators: across oracle-comparable labeled response pairs, the sign of the judge margin agrees with the oracle 73.6% of the time when the judge margin is non-zero, and same-prompt probe comparisons reach 84.6% accuracy when the judge does not tie. These results support treating LLM judges primarily as comparators rather than as cardinal rulers.

1 Introduction

LLM-as-judge has become a standard way to scale evaluation when human or expert labels are too expensive to collect broadly [Zheng et al., 2023, Chiang et al., 2024, Dubois et al., 2024, Landesberg, 2025]. The usual operational loop is simple: score every response with a cheap judge, label a small slice with a trusted oracle, and then use the judge as a proxy for what we actually care about. CJE systematizes this pattern by explicitly learning the judge-to-oracle mapping and propagating uncertainty from the calibration step [Landesberg, 2025].

The weakness of the usual pointwise formulation is that it treats the judge score as if it were a meaningful ruler. In practice, small cardinal gaps such as 0.81 versus 0.86 are often hard to interpret. Scales can compress near the top, drift over time, or move with nuisance factors such as response length [Zheng et al., 2023, Dubois et al., 2024, Jeong et al., 2025]. Yet the same judge may still be useful as a discriminator. It may be much easier for the judge to say that response A is better than response B than to place the two responses on a stable absolute scale.

*Gaurav can be reached at gsood07@gmail.com

This suggests a modest but important change in target. Rather than estimating an absolute oracle score, we target the *oracle win rate* of policy A against policy B. Pairwise win probabilities have been central objects in paired-comparison statistics since at least Thurstone [1927] and Bradley and Terry [1952], and modern LLM evaluation systems already rely heavily on pairwise preferences [Chiang et al., 2024, Dubois et al., 2024, Liusie et al., 2024, Liu et al., 2024]. The novelty here is not the estimand itself. The novelty is the combination of that estimand with a CJE-style calibration scheme that uses a cheap judge everywhere, an oracle on a small slice, and either a direct pairwise judge score or an induced pointwise margin.

There is a trade-off. Pairwise judgments do discard information. If a stable latent utility scale existed and the judge measured it well, then a cardinal score would tell us more than a binary or ternary comparison. But that extra information only matters when it is trustworthy. In many real LLM-evaluation settings, absolute levels contain a substantial amount of false precision. On balance, a calibrated win probability can be more useful because it is interpretable, more stable under scale drift, and directly tied to deployment decisions: how often would model A beat model B on the prompts we care about?

This paper makes four contributions.

- It formalizes a pairwise CJE target, the oracle win rate $\theta_{A,B}$, and shows how to estimate it from either a direct pairwise judge or an induced pointwise margin.
- It gives simple identification, symmetry, and consistency results for the plug-in estimator.
- It clarifies when pairwise comparisons are useful despite discarding magnitude information, and when they are not.
- It reanalyzes the public CJE Arena sample and finds a pattern that matches the motivating intuition: pointwise judge scores are only moderately aligned with oracle scores, while induced pairwise margins are much more useful as ordering devices.

2 Related work

2.1 Paired-comparison statistics

Pairwise win probabilities are classical statistical objects. Thurstone [1927] introduced the law of comparative judgment, and Bradley and Terry [1952] developed the Bradley-Terry model for paired comparisons. Modern work continues to treat pairwise comparisons as the basic data structure for ranking problems, including computational and inferential advances for Bradley-Terry estimation [Newman, 2023]. In this literature, the primary inferential target is not an absolute score on an externally meaningful scale, but the probability that one item beats another or the latent strengths that generate those probabilities.

2.2 Preference learning and RLHF

Preference-based supervision in machine learning also starts from pairwise data. Christiano et al. [2017] learn reward models from human comparisons between trajectory segments, and Ouyang et al. [2022] collect rankings of model outputs as a central ingredient in RLHF for instruction-following language models. In both cases, pairwise preferences are attractive because humans can often compare alternatives more reliably than they can assign absolute scores.

2.3 LLM evaluation has already moved in a pairwise direction

Modern LLM evaluation platforms also lean heavily on pairwise judgments. Chatbot Arena is explicitly built around pairwise human preferences [Chiang et al., 2024]. AlpacaEval reports win-rate style metrics against a baseline and explicitly models confounders such as output length [Dubois et al., 2024]. Liusie et al. [2024] argue that comparative assessment often outperforms direct scoring for NLG evaluation, while Liu et al. [2024] formulate LLM evaluation as a ranking problem built from local pairwise comparisons. These papers establish that pairwise comparison is already a mainstream evaluation primitive.

2.4 Caveats for pairwise judging

Pairwise judgments are not a free lunch. Zheng et al. [2023] document position, verbosity, and self-enhancement biases in LLM judges. More recently, Jeong et al. [2025] argue that direct pairwise comparison can amplify certain judge biases relative to pointwise evaluation. This is one reason calibration, abstention, ties, and transportability diagnostics matter. The question is not whether pairwise comparisons are magically unbiased. The question is whether they are a better match to the information the judge can reliably provide.

2.5 Where this paper sits

CJE studies the general problem of learning a judge-to-oracle correction from a small labeled slice and then applying it at scale [Landesberg, 2025]. Our proposal is to specialize that logic to a pairwise target. The pairwise win-rate estimand is classical. What appears less standard is the specific package of: (i) a CJE-style oracle-slice calibration step, (ii) a pairwise target parameter, and (iii) the ability to work either with direct pairwise judgments or with induced margins built from pointwise judge scores.

3 Setup and target estimand

Let prompts be drawn from a target distribution $X \sim P_X$. Two policies π_A and π_B produce responses $R_A \sim \pi_A(\cdot \mid X)$ and $R_B \sim \pi_B(\cdot \mid X)$ on the *same* prompt. Let $Q(x, r)$ denote the oracle quality functional, such as a human preference score, an expert rubric score, or an expensive model score on a common scale.

Define the oracle pairwise label

$$Y_{A,B}(X) = \mathbf{1}\{Q(X, R_A) > Q(X, R_B)\} + \tfrac{1}{2}\mathbf{1}\{Q(X, R_A) = Q(X, R_B)\}. \tag{1}$$

Thus $Y_{A,B} \in \{0, 1/2, 1\}$ when the oracle is deterministic. More generally, if the oracle itself is stochastic or aggregated over raters, all results below continue to hold for any $Y_{A,B} \in [0, 1]$.

The estimand of interest is the oracle win rate

$$\theta_{A,B} = \mathbb{E}\left[Y_{A,B}(X)\right]. \tag{2}$$

Interpretationally, $\theta_{A,B}$ is the probability that policy A beats policy B on a random prompt, counting ties as one-half. This quantity directly answers a common evaluation question. If $\theta_{A,B} = 0.63$, then policy A beats policy B about 63% of the time on the target prompt distribution.

The pointwise alternative would target something like

$$\mu_A = \mathbb{E}\left[Q(X, R_A)\right], \tag{3}$$

which requires the judge to support a stable cardinal interpretation. The pairwise target makes a weaker claim: that the judge contains enough information to say which of two responses is more likely to win.

4 Pairwise calibration from direct judges or induced margins

4.1 Two surrogate constructions

Let $Z_{A,B}$ denote the judge-side signal used for calibration. There are two natural cases.

Direct pairwise judge. Query the judge on (X, R_A, R_B) and let it output a signed score $S_{A,B} \in \mathbb{R}$, such as a preference margin or a transformed probability.

Induced pointwise margin. When only pointwise judge scores are available, define

$$D_{A,B} = S(X, R_A) - S(X, R_B). \tag{4}$$

This case is attractive in CJE-like settings because it can be constructed from already logged pointwise judge scores, provided the two responses are evaluated on the same prompt.

In what follows, $Z_{A,B}$ stands for either $S_{A,B}$ or $D_{A,B}$.

4.2 The calibration map

Define the conditional mean calibration function

$$m(z) = \mathbb{E}\left[Y_{A,B} \mid Z_{A,B} = z\right]. \tag{5}$$

This is the pairwise analogue of the judge-to-oracle correction in CJE. A natural parametric choice is a symmetric logistic map

$$m_\beta(z) = \frac{1}{1 + e^{-\beta z}}, \tag{6}$$

which satisfies $m_\beta(-z) = 1 - m_\beta(z)$. This symmetry is attractive when swapping the responses flips the sign of the judge signal. A nonparametric alternative is isotonic regression or a symmetrized isotonic fit when calibration labels are scarce. The choice is practical rather than doctrinal. In one dimension, isotonic regression is data-efficient; a symmetric logistic or monotone spline is often cleaner when one wants interpretable pairwise probabilities.

4.3 The plug-in estimator

Suppose we observe unlabeled evaluation prompts $X_1, \ldots, X_n$ and their associated judge-side signals $Z_1, \ldots, Z_n$ for the policy pair (A, B). If $\hat{m}$ is learned on a separate oracle-labeled slice, the natural estimator is

$$\hat{\theta}_{A,B} = \frac{1}{n} \sum_{i=1}^{n} \hat{m}(Z_i). \tag{7}$$

For a collection of policies $1, \ldots, K$, one can estimate the full matrix $(\hat{\theta}_{a,b})_{a,b}$ and then summarize it using pairwise win rates directly, Copeland scores, or a Bradley-Terry model fit to the calibrated pairwise matrix.

4.4 Heterogeneity and transportability

A single one-dimensional map $m(z)$ may be too coarse when the judge-to-oracle relationship varies with task type, prompt distribution, or response characteristics. In that case the natural extension is

$$m(z, w) = \mathbb{E}\left[Y_{A,B} \mid Z_{A,B} = z, W = w\right], \tag{8}$$

where W collects covariates such as topic or response length. This mirrors the role of covariates in CJE's two-stage calibration variants [Landesberg, 2025]. The statistical point is simple: learning a 1D map is not the hard part; the hard part is making sure the map remains valid across the mixtures of prompts and styles that appear in deployment.

5 Theory

Proposition 1 (Identification)**.** *Let*

$$m(z) = \mathbb{E}\left[Y_{A,B} \mid Z_{A,B} = z\right]. \tag{9}$$

Then

$$\theta_{A,B} = \mathbb{E}\left[Y_{A,B}\right] = \mathbb{E}\left[m(Z_{A,B})\right]. \tag{10}$$

Proof. By the law of iterated expectations,

$$\mathbb{E}\left[Y_{A,B}\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{A,B} \mid Z_{A,B}\right]\right] = \mathbb{E}\left[m(Z_{A,B})\right].$$

□

The proposition is deliberately plain. The method lives or dies on whether the calibration function estimates the right conditional mean under the evaluation distribution.

Proposition 2 (Symmetry)**.** *Assume that $Y_{A,B} + Y_{B,A} = 1$ almost surely, $Z_{B,A} = -Z_{A,B}$ almost surely, and $m(-z) = 1 - m(z)$ for all z. Then*

$$\theta_{A,B} + \theta_{B,A} = 1. \tag{11}$$

Moreover, the plug-in estimator satisfies $\hat{\theta}_{A,B} + \hat{\theta}_{B,A} = 1$ whenever the same symmetry is imposed on $\hat{m}$.

Proof. Using Proposition 1 and the assumed antisymmetry,

$$\theta_{B,A} = \mathbb{E}[m(Z_{B,A})] = \mathbb{E}[m(-Z_{A,B})] = \mathbb{E}[1 - m(Z_{A,B})] = 1 - \theta_{A,B}.$$

The same argument applies to the sample average with $\hat{m}$ in place of m. □

Proposition 3 (Consistency under sample splitting)**.** *Suppose the oracle-labeled calibration sample is independent of the evaluation sample, and $\hat{m}$ is estimated only on the calibration sample. If*

$$\mathbb{E}\,|\hat{m}(Z_{A,B}) - m(Z_{A,B})| \xrightarrow{p} 0, \tag{12}$$

and $m(Z_{A,B})$ is integrable, then

$$\hat{\theta}_{A,B} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(Z_i) \xrightarrow{p} \theta_{A,B}. \tag{13}$$

Proof. Write

$$\hat{\theta}_{A,B} - \theta_{A,B} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{m}(Z_i) - m(Z_i))}_{(I)} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} m(Z_i) - \mathbb{E}[m(Z_{A,B})]\right)}_{(II)}.$$

For term (I), conditional on the calibration sample,

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{m}(Z_i) - m(Z_i))\right| \middle| \hat{m}\right] \leq \mathbb{E}\left[|\hat{m}(Z_{A,B}) - m(Z_{A,B})||\hat{m}\right],$$

which converges to zero in probability by assumption. For term (II), the law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i) \xrightarrow{p} \mathbb{E}[m(Z_{A,B})] = \theta_{A,B}.$$

Hence both terms vanish in probability. □

Remark 1 (Uncertainty)**.** *If m were known, then $\mathrm{Var}(\hat{\theta}_{A,B}) = \mathrm{Var}(m(Z_{A,B}))/n$. In practice m is estimated on a small oracle slice, so standard errors that treat $\hat{m}$ as fixed are too optimistic. The natural operational analogue of CJE is to refit the calibrator across oracle folds and add a calibration-uncertainty term by a delete-one-fold jackknife or nested bootstrap [Landesberg, 2025].*

6 Why pairwise comparisons can be useful despite information loss

Moving from absolute scores to pairwise wins does destroy information. If a judge provided a trustworthy latent utility measure, then collapsing U_A and U_B to the sign of $U_A - U_B$ would throw away effect size information. One could no longer distinguish a narrow win from a blowout using the comparison alone.

The practical question is whether the lost information was real or illusory. Pairwise targets remain popular for at least four reasons.

First, they match how people and models often make judgments more reliably. It is usually easier to say that A is better than B than to decide that A deserves 0.81 and B deserves 0.86. This intuition is explicit in both classical paired-comparison work and recent LLM evaluation papers [Thurstone, 1927, Liusie et al., 2024].

Second, pairwise win rates live on an interpretable scale. A win rate of 0.65 means one system beats another about two-thirds of the time. That remains meaningful even if the underlying judge scale compresses, shifts, or changes units.

Third, pairwise targets line up with common decisions. In many deployment settings the relevant question is whether a candidate model beats the baseline often enough to justify adoption. A calibrated win probability directly addresses that question.

Fourth, if one does need a global ranking or a latent-strength summary, pairwise win rates can still be fed into Bradley-Terry style models. The resulting latent score is model-based and comparative. It is not the same as pretending the raw judge score was already a meaningful ruler.

This is not an argument that pairwise is always superior. If one needs absolute guarantees, safety thresholds, or cost-benefit calculations on a common cardinal scale, then pairwise wins are insufficient on their own. The argument is narrower: when the judge is much more reliable as a comparator than as a ruler, the estimand should respect that fact.

7 Empirical reanalysis of the public CJE sample

7.1 Data

The public CJE repository ships an Arena sample dataset with 1,000 base-policy responses, of which 480 carry oracle labels, plus fresh judge-scored draws for several target policies and small oracle-labeled probe slices for transportability checks [cimo-labs, 2026a,b]. The repository states that judge scores in this sample come from GPT-4.1-nano and oracle labels come from GPT-5 [cimo-labs, 2026b]. This is a public demonstration sample, not the full 4,961-prompt Arena benchmark analyzed in the CJE paper [Landesberg, 2025].

7.2 Pointwise judge scores are compressed and only moderately aligned with the oracle

Table 1 reports descriptive statistics from the 480 oracle-labeled base responses in the public sample. The raw pointwise relationship between judge and oracle is moderate rather than strong: Pearson $r = 0.390$, Spearman $\rho = 0.446$, and Kendall $\tau = 0.360$. The identity map from judge score to oracle score also performs poorly, with a negative R^2.

The score distribution reveals why. The judge is highly compressed near the top of the scale. In this sample, 86.3% of judge scores are at least 0.85, while only 47.9% of oracle labels are. Fully 61.0% of judge scores are exactly 0.85. That is useful if the judge is screening for broadly acceptable answers, but not if one wants to interpret the raw score as a precise cardinal measurement.

7.3 Induced pairwise margins are materially better discriminators

We next form all unordered oracle-comparable pairs from the same 480 labeled responses. For each pair (i, j) with distinct oracle labels, define the induced judge margin $D_{ij} = S_i - S_j$ and let the oracle indicate which response truly wins. Table 2 reports descriptive pairwise statistics.

The key pattern is much healthier than in the pointwise analysis. When the judge margin is positive, the oracle agrees with the sign 76.7% of the time. When the judge margin is zero, the oracle is

Table 1: Pointwise judge-oracle alignment on the 480 oracle-labeled base responses in the public CJE Arena sample.

Metric	Value
Oracle-labeled base responses	480
Pearson correlation	0.390
Spearman correlation	0.446
Kendall tau	0.360
MAE of raw identity map	0.155
RMSE of raw identity map	0.230
R^2 of raw identity map	-0.026
Mean judge score	0.840
Mean oracle score	0.754
Judge standard deviation	0.132
Oracle standard deviation	0.227
Share with judge score ≥ 0.85	0.863
Share with oracle label ≥ 0.85	0.479
Share with judge score exactly 0.85	0.610

Table 2: Descriptive pairwise metrics from all unordered oracle-comparable pairs formed from the 480 oracle-labeled base responses. Because the same responses appear in many pairs, these figures are descriptive summaries rather than iid estimates.

Metric	Value
Oracle-comparable unordered pairs	111,400
Pr(oracle higher $\mid D_{ij} > 0$)	0.767
Pr(oracle higher $\mid D_{ij} = 0$)	0.525
Pr(oracle higher $\mid D_{ij} < 0$)	0.294
Accuracy when judge margin is non-zero	0.736
Tie-aware concordance index	0.641
Tie-aware concordance index for oracle gap > 0.3	0.710
Accuracy for oracle gap > 0.3 and non-zero judge margin	0.820
Symmetric logistic slope on D_{ij}	4.260
AUC of symmetric logistic fit	0.707
Brier score of symmetric logistic fit	0.223

nearly a coin flip. Overall, among pairs with non-zero judge margin, the sign of the judge agrees with the oracle 73.6% of the time. The tie-aware concordance index is 0.641, and it rises to 0.710 when the oracle gap exceeds 0.3.

These numbers should be read descriptively rather than inferentially. The all-pairs construction reuses the same 480 responses many times, so the resulting pair indicators are strongly dependent. That dependence does not invalidate the qualitative pattern, but it does mean that naive binomial confidence intervals would exaggerate precision.

7.4 Same-prompt probe comparisons remain useful, but ties matter

The probe slices in the public sample allow a smaller but cleaner same-prompt comparison. Matching probe responses to base-policy responses on the same prompt yields 49 oracle-labeled base-versus-target comparisons across the clone, parallel-universe-prompt, and unhelpful policies. Table 3

Table 3: Same-prompt base-versus-target comparisons from the oracle-labeled probe slices. The Wilson interval is reported only for the non-tied same-prompt comparisons, where the unit of analysis is a prompt-level pair.

Metric	Value
Oracle-labeled same-prompt pairs	49
Judge tie rate	0.469
Non-tied judge comparisons	26
Accuracy on non-tied judge comparisons	0.846
95% Wilson interval	[0.665, 0.938]
Tie-aware concordance index	0.684

Policy	n	Tie rate	Non-ties	Accuracy	Base win rate	Judge diff.	Oracle diff.
clone	14	0.857	2	0.000	0.571	-0.018	0.004
parallel universe prompt	15	0.600	6	0.667	0.400	0.007	-0.087
unhelpful	20	0.100	18	1.000	1.000	0.543	0.645

reports the results.

Here the pattern is even clearer. When the judge does not tie, the same-prompt accuracy is 84.6%. But ties are common: 46.9% of the matched policy pairs have zero judge margin. For the clone policy, the tie rate is 85.7%; for the unhelpful policy, it is only 10%. In other words, the judge is perfectly willing to discriminate when differences are obvious, but it becomes hesitant when policies are close. That is not a bug. It is precisely the sort of information a pairwise method should preserve rather than hide.

7.5 Takeaway from the reanalysis

The public sample is small, and the all-pairs analysis is descriptive rather than fully inferential. But the central qualitative message is robust: the judge behaves much more plausibly as a comparator than as a cardinal ruler. That is exactly the use case for a pairwise estimand.

8 Discussion

The proposed estimand is intentionally conservative. It does not claim that pairwise judging solves bias, transportability, or calibration. It claims only that if the judge is better at ordering than at measuring, then pairwise win rate is the more defensible target.

There are several limitations.

First, pairwise targets are still vulnerable to judge bias. Position effects, verbosity bias, and stylistic preference can all distort pairwise outcomes [Zheng et al., 2023, Jeong et al., 2025]. Calibration helps only if the judge-to-oracle relationship is stable enough to learn and transport.

Second, a single scalar signal Z may be too crude. If the judge-to-oracle map varies strongly across prompt families, a global calibration can fail even when it looks good on average. Covariate-aware calibration is the natural next step.

Third, pairwise win rate is not the right target for every problem. Absolute safety thresholds or business KPIs may require cardinal quantities, not just win probabilities. Pairwise CJE is best

viewed as a targeted tool for pairwise model comparison and ranking.

Finally, the empirical section here uses the public CJE sample, not the full Arena benchmark. The descriptive reanalysis is meant to illustrate the phenomenon, not to supersede the larger experiments in Landesberg [2025].

9 Conclusion

Pairwise win probabilities are not a new invention. They are classical statistical objects and already sit underneath much of modern LLM evaluation. The contribution of this paper is narrower and more practical: it adapts the CJE idea of calibrating a cheap judge to a small oracle slice so that the primary target is a pairwise oracle win rate rather than an absolute oracle score.

The theoretical story is simple. If one can estimate the conditional mean oracle preference given the judge signal, then averaging that calibrated probability over unlabeled comparisons recovers the oracle win rate. The empirical story is equally simple. In the public CJE sample, raw pointwise scores are only moderately aligned with the oracle and appear heavily compressed, while induced pairwise margins are substantially more useful as ordering devices.

In plain language, the judge looks less like a ruler and more like a referee. Once that is true, the estimand should change with it.

References

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

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

cimo-labs. CJE: Causal judge evaluation. GitHub repository, 2026a. https://github.com/cimo-labs/cje, accessed March 2, 2026.

cimo-labs. Arena sample dataset for CJE. GitHub repository documentation, 2026b. https://raw.githubusercontent.com/cimo-labs/cje/main/examples/arena_sample/README.md, accessed March 2, 2026.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2024.

Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee, and Jaegul Choo. The comparative trap: Pairwise comparisons amplify biased preferences of LLM evaluators. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 79–108, 2025.

Eddie Landesberg. Causal judge evaluation: Calibrated surrogate metrics for LLM systems, 2025.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators, 2024.

Adian Liusie, Potsawee Manakul, and Mark Gales. 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 112–130, 2024.

M. E. J. Newman. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

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. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023.