VALIDATING RANKINGS FROM PAIRWISE COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pairwise human preferences are widely used to evaluate and rank large language models on open-ended tasks, as in Arena (Chiang et al., 2024). This raises a natural question: how can we determine whether a reported ranking is good? More specifically, which measures of ranking quality can be reliably validated when the ground truth is unknown and only finite comparison data are available? We formalize this question through statistical justifiability of an error metric, asking whether there exists a validation algorithm that can reliably distinguish outputs with small error from those with substantially larger error using finite comparison data. We first show that ranking error, a natural metric for the central objective of ranking accuracy, is not statistically justifiable. This motivates seeking alternative, statistically justifiable error metrics. We develop metrics based on tests with observable outcomes, tailoring the tests to the information available for validation. For example, when the ranking method also returns estimated probabilities that one model is preferred over another, these enable pairwise multicalibration tests, which assess agreement with observed preferences within slices of the data such as tasks or domains. When only a reported ranking is available, we measure its compatibility with calibrated predictions.

1 INTRODUCTION

Pairwise comparisons provide a natural way to evaluate large language models (LLMs), particularly on open-ended tasks with no single correct answer (Ouyang et al., 2022; Zheng et al., 2023; Chiang et al., 2024; Dubois et al., 2024). A prominent example is Arena (formerly Chatbot Arena) (Chiang et al., 2024; Dubois et al., 2024), where users compare responses from two anonymous models and indicate which response they prefer. A ranking method aggregates these judgments into a leaderboard, often through a model such as Bradley–Terry. Such leaderboards are easy to produce and interpret, but substantially harder to validate, as the population ordering they seek to estimate is never directly observed.

The central difficulty lies in the heterogeneity of preferences across users and tasks (Frick et al., 2025; Garbacea et al., 2026). Preferences may vary both in which model is favored and in how strongly it is favored (Gölz et al., 2025). With only a few comparisons per context, this variation can be difficult to infer without assumptions connecting preferences across contexts. Pooling comparisons reveals how often each model wins overall, but can obscure differences across contexts that matter for the intended ordering. Finite comparisons may thus be insufficient to recover the underlying preferences or determine the intended ranking. Therefore, we need to find measures of ranking quality that can still be reliably validated when such recovery is unavailable. This raises a fundamental question:

Which measures of ranking quality can be reliably validated through finite comparison data?

More specifically, we ask which error metrics admit a finite-sample procedure that uses fresh comparison data to reliably distinguish outputs with small error from those with substantially larger error. For ranking error, this would mean distinguishing a ranking that orders almost all model pairs correctly from one that misorders a substantial fraction. We formalize this requirement as *statistical justifiability* and study which measures of ranking quality satisfy it.

What can be validated also depends on the information available to the validator. A reported ranking alone provides no predicted win probabilities to check against comparison outcomes. If the procedure also reports its estimated win probabilities across contexts, the validator can assess whether these predictions agree with fresh observations. Access to the algorithm and the data used to produce the ranking further enables behavioral diagnostics, such as assessing the procedure's robustness to changes in its inputs.

Our contributions are as follows.

1. We propose *statistical justifiability* as a principle for selecting evaluation metrics for rankings: a metric should measure quality that can be reliably assessed from finite comparison data. We formalize this requirement through a validator that distinguishes outputs with small error from those with substantially larger error using fresh comparisons.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

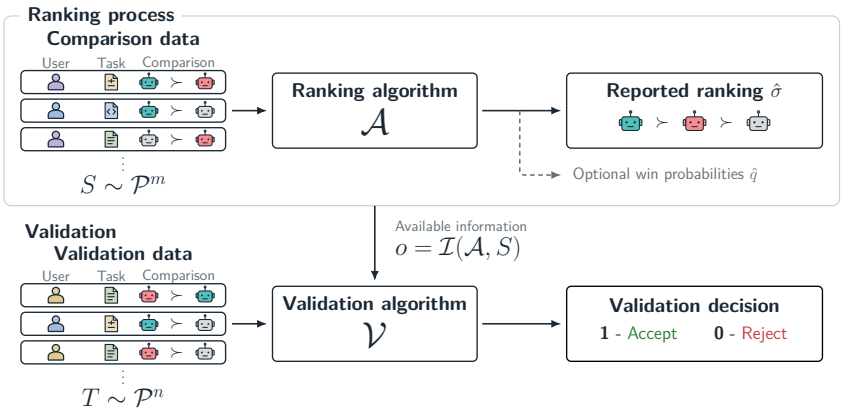


Figure 1: The validation pipeline. Given information $o = \mathcal{I}(\mathcal{A}, S)$ about the ranking process and independent comparison data T, the validation algorithm aims to distinguish error at most ε from error at least $\varepsilon + \gamma$.

2. We show that ranking error is not statistically justifiable in general. Near ties make opposite orderings difficult to distinguish from finite comparisons. Even when one model has a substantial average advantage, unrestricted variation across contexts can prevent uniform validation. Widely used methods, such as context-independent Bradley–Terry fitting, simplify this heterogeneous setting by pooling comparisons across contexts. We show that this simplification can discard information needed to identify the intended ordering, even with unlimited data.
3. We propose a general class of *test-based metrics* that assess explicit requirements on a ranking procedure's outputs and behavior. We instantiate this framework according to the information available for validation. With predicted win probabilities, we introduce pairwise multicalibration to test agreement with outcomes across context groups and prediction bins (Hébert-Johnson et al., 2018). With ranking-only access, we define a compatibility metric that asks whether calibrated predictions could induce the reported ordering. With algorithm access, we introduce evidence consistency to assess sensitivity to changes in the candidate set, drawing inspiration from independence of irrelevant alternatives (Arrow, 1950). The framework also supports tests that combine outcome and behavioral information. We establish a general validation guarantee: Lipschitz aggregations of finitely many bounded tests are statistically justifiable.
4. We relate these calibration metrics to log-loss and ranking error. Small excess log-loss controls calibration error and yields ranking guarantees under suitable conditions on the aggregation rule and separation between models. Following the perspective of outcome indistinguishability (Dwork et al., 2021), we also show that calibrated predictions minimize population log-loss in a comparison world that agrees with the true world on the selected test expectations.

We complement these results with controlled simulations examining ranking recovery under varying margins and an Arena case study of ranking stability and group calibration. Related work is discussed in Section A.

2 Problem Setup

We study a finite set of models $\mathcal{M}$, with $|\mathcal{M}| = M \geq 2$, evaluated across user–task contexts. A context $x = (i, t)$ belongs to $\mathcal{X} = \mathcal{U} \times \mathcal{T}$, and $\mathcal{D}$ describes how often such contexts occur in the population. For distinct models $a, b \in \mathcal{M}$, let $p_{ab}(x) \in [0, 1]$ be the probability that the user prefers a to b at x. We assume $p_{ba}(x) = 1 - p_{ab}(x)$ and that each p_{ab} is measurable. The comparisons reveal these preferences indirectly. To say whether a reported ranking is correct, we must first specify the population ranking they are meant to represent. We consider two modeling choices, with fixed deterministic tie-breaking in both.

Model-free setting. The first choice treats the win probabilities p as the primitive object, with no additional assumptions on it. Following the social-choice view of a ranking as an aggregation of individual preferences (Arrow, 1950), we fix an aggregation rule Aggr and define $\sigma^\star = \mathrm{Aggr}(p, \mathcal{D})$. The rule specifies how pairwise preferences are combined, while $\mathcal{D}$ determines the weight of each user–task context. For example, Borda aggregation (Shah and Wainwright, 2018) scores model m by its expected chance of beating a uniformly chosen opponent, $\frac{1}{M-1} \sum_{b \in \mathcal{M} \setminus \{m\}} \mathbb{E}_{x \sim \mathcal{D}} [p_{mb}(x)]$

and ranks models by decreasing score. In this setting, the target depends directly on the population's comparison probabilities.

Utility-based setting. The second choice starts from a latent utility $u^\star(x, m) \in [0, 1]$ for each model and context. Here the target $\sigma^\star$ orders models by their mean utilities $\mathbb{E}_{x\sim\mathcal{D}}[u^\star(x, m)]$. A specified response model F maps a utility function u to win probabilities $F(u)$, and we assume $p = F(u^\star)$. For the Bradley–Terry response model, $[F_{\mathrm{BT}}(u)]_{ab}(x) := \frac{1}{1+\exp\big(-\beta(u(x,a)-u(x,b))\big)}$ where $\beta > 0$ is fixed. Thus a comparison at x reflects a local utility difference, while the target ranking compares utilities averaged over $x \sim \mathcal{D}$.

Pairwise comparison data generation. Both models use the same observation process. We first draw $x \sim \mathcal{D}$ and, independently, an ordered pair $(a, b) \sim \Pi$. We then draw $y \mid x, a, b \sim \mathrm{Ber}(p_{ab}(x))$, with $y = 1$ meaning that a wins. We fix Π and assume $\Pi(a, b) > 0$ for every ordered pair of distinct models. Write $\mathcal{P}_{\mathcal{D},p}$ for the resulting distribution of $W = (x, a, b, y)$. A training sample of n independent comparisons is $S = \{(x_j, a_j, b_j, y_j)\}_{j=1}^n \sim \mathcal{P}_{\mathcal{D},p}^n$. This common data model lets us ask what can be learned or validated under either definition of $\sigma^\star$.

2.1 Error Metric and Statistical Justifiability

Given a candidate set $\mathcal{M}$ and data S, a ranking algorithm $\mathcal{A}$ outputs a ranking $\widehat{\sigma}$ over $\mathcal{M}$, possibly together with a predicted win-probability function $\widehat{q}$. A reported ranking $\widehat{\sigma}$ can be checked against the target ordering $\sigma^\star$. If the algorithm also reports $\widehat{q}$, we can ask how well its probabilities predict comparisons.

For a reported ranking, a natural metric is the normalized Kendall–tau distance, which measures the fraction of model pairs ordered differently from the ground-truth ranking.

Definition 1 (Kendall–tau distance). *Let $M = |\mathcal{M}| \geq 2$. The normalized Kendall–tau distance between two rankings σ and $\sigma^\star$ over $\mathcal{M}$ is $d_{\mathrm{KT}}(\sigma, \sigma^\star) := \frac{1}{\binom{M}{2}} \sum_{\{a,b\}\in\binom{\mathcal{M}}{2}} \mathbb{1}\{\sigma_{ab} \neq \sigma^\star_{ab}\}$ where $\sigma_{ab} := \mathbb{1}\{\sigma(a) < \sigma(b)\}$ indicates whether σ ranks a above b. When the comparison distribution $\mathcal{P}$ determines a ground-truth ranking $\sigma^\star$, the corresponding ranking error is $f_{\mathrm{KT}}(\sigma, \mathcal{P}) := d_{\mathrm{KT}}(\sigma, \sigma^\star)$.*

A ranking tells us which model is placed first in each pair, but not how likely either model is to win. When $\widehat{q}$ is available, log-loss measures the quality of those probability predictions.

Definition 2 (Log-loss). *Given a win-probability function q, its population log-loss is $\ell(q, \mathcal{P}) := \mathbb{E}_{(x,a,b,y)\sim\mathcal{P}}\big[-y \log q_{ab}(x) - (1 - y) \log\big(1 - q_{ab}(x)\big)\big]$.*

Our aim is to validate a claim about the output of a ranking algorithm. A validator is a separate procedure that examines the available output and an independent sample of comparisons, then decides whether the claimed error indeed validate the ranking. What it can conclude depends on what the algorithm reveals. We model the information available about the ranking process through an access function $\mathcal{I}$. Let $o = \mathcal{I}(\mathcal{A}, S)$ denote the information revealed about an algorithm run $(\mathcal{A}, S)$. For example, o may be the reported ranking $\widehat{\sigma}$, the ranking together with an estimated win-probability function $\widehat{q}$, or the algorithm code $\mathcal{A}$ and data S. An error metric $f(o, \mathcal{P}) \in [0, \infty]$ measures the performance of the algorithm run as represented by o under the comparison distribution $\mathcal{P}$.

Formally, for fixed $\varepsilon, \gamma > 0$ and $\delta \in (0, 1)$, the validator $\mathcal{V}$ receives o and fresh data T, independent of S, and returns $\mathcal{V}(T, o) \in \{0, 1\}$, where 1 means accept. Its task is to accept when $f(o, \mathcal{P}) \leq \varepsilon$ and reject when $f(o, \mathcal{P}) \geq \varepsilon + \gamma$, with probability at least $1 - \delta$. The gap of width γ allows either decision for intermediate errors. Crucially, the validator does not know $\mathcal{P}$. It must infer enough from T to make this decision. Figure 1 illustrates this process

Definition 3 (Statistical justifiability). *Fix a candidate set $\mathcal{M}$, an access model $\mathcal{I}$, and a class of comparison distributions. For $\varepsilon, \gamma > 0$ and $\delta \in (0, 1)$, an error metric f is* $(\varepsilon, \gamma, \delta)$-statistically justifiable *if there exist a validation algorithm $\mathcal{V}$ and a finite sample bound $N_{\mathcal{M},f}(\varepsilon, \gamma, \delta)$ such that the following holds.*

For every $\mathcal{P}$ and algorithm run $(\mathcal{A}, S)$, let $o = \mathcal{I}(\mathcal{A}, S)$. Given independent validation data $T \sim \mathcal{P}^n$ with $n \geq N_{\mathcal{M},f}(\varepsilon, \gamma, \delta)$, with probability at least $1 - \delta$ over T and any randomness of $\mathcal{V}$, both implications hold:

$$f(o, \mathcal{P}) \leq \varepsilon \quad \Longrightarrow \quad \mathcal{V}(T, o) = 1, \qquad f(o, \mathcal{P}) \geq \varepsilon + \gamma \quad \Longrightarrow \quad \mathcal{V}(T, o) = 0.$$

We call f statistically justifiable *if this property holds for every $\varepsilon, \gamma > 0$ and $\delta \in (0, 1)$.*

With this definition, the rest of the paper asks which error metrics admit such a validator.

3 FAILURE OF RANKING ERROR AND ALTERNATIVE TEST-BASED METRICS

3.1 FAILURE OF RANKING ERROR AS A METRIC

We first ask whether comparison data can certify a reported ranking. The answer is negative in the utility-based setting, even when preferences follow the Bradley–Terry model. To isolate the issue, consider two models under ranking-only access. The validator receives $\widehat{\sigma}$ and an independent comparison sample T, but neither estimated win probabilities nor training data.

Fix $\mathcal{M} = \{a, b\}$, the context distribution $\mathcal{D}$, and a reported ranking $\widehat{\sigma}$. With two models, the Kendall–tau error is 0 if $\widehat{\sigma} = \sigma^\star$ and 1 otherwise. Setting $\varepsilon = \gamma = 1/2$ therefore reduces validation to a binary decision, i.e. accept precisely when $\widehat{\sigma}$ is the mean-utility ranking. For each possible utility function $u^\star$, let $\mathcal{P} := \mathcal{P}_{\mathcal{D}, F_{\mathrm{BT}}(u^\star)}$, and let $\sigma^\star$ denote the corresponding mean-utility ranking under $\mathcal{D}$. Define the validation error as

$$\mathrm{ValErr}_{u^\star, n}(\mathcal{V}, \widehat{\sigma}) := \mathbb{P}_{T \sim \mathcal{P}^n, \mathcal{V}} \left(\mathcal{V}(T, \widehat{\sigma}) \neq \mathbb{1}\{\widehat{\sigma} = \sigma^\star\} \right).$$

Statistical justifiability requires a single validator and finite sample bound that work for every admissible $u^\star$. For any $\delta < 1/2$, the error above must be at most δ whenever n exceeds this bound, which cannot depend on the utility function that generated the comparisons.

We now examine two reasons this uniform requirement fails. A near-tie can make opposite rankings statistically indistinguishable. Even after near-ties are excluded, arbitrary variation across contexts can hide a substantial average utility gap.

Small-margin sensitivity. Let $\widehat{\sigma} = (a \succ b)$ and consider two utility functions with opposite gaps of magnitude Δ. The correct validation decision changes between the two populations, but their Bradley–Terry comparison distributions converge as $\Delta \to 0$. The following proposition makes the resulting testing lower bound explicit.

Proposition 1 (Small-margin lower bound)**.** *Let $\mathcal{D}$ be concentrated on one context, let $\widehat{\sigma} = (a \succ b)$, and, for $\Delta \in (0, 1]$, define $u^\star_\pm(a) = \frac{1}{2} \pm \frac{\Delta}{2}$, $u^\star_\pm(b) = \frac{1}{2} \mp \frac{\Delta}{2}$. Under the Bradley–Terry response model with parameter $\beta > 0$, every possibly randomized validator $\mathcal{V}$ and every $n \geq 1$ satisfy $\max_{\omega \in \{+,-\}} \mathrm{ValErr}_{u^\star_\omega, n}(\mathcal{V}, \widehat{\sigma}) \geq \frac{1}{2} \left(1 - \min\left\{1, \frac{\beta\Delta\sqrt{n}}{2}\right\}\right)$. Consequently, Kendall–tau ranking error is not statistically justifiable uniformly over all positive utility gaps.*

An additive utility tolerance $\tau > 0$ removes this particular obstruction if an inverted pair $\{a, b\}$ is penalized only when $|\mathbb{E}_{x \sim \mathcal{D}} [u^\star(x, a) - u^\star(x, b)]| > \tau$. If $\Delta \leq \tau$, the reported ranking incurs no tolerant error in either case. The obstructions below nevertheless persist whenever τ is smaller than the fixed positive utility gaps in their constructions.

No free lunch without complexity control. A positive mean-utility margin alone does not ensure statistical justifiability. Even when contexts are observed, unrestricted utilities prevent uniform validation from finite comparison data.

Proposition 2 (No-free-lunch for validating ranking error)**.** *Suppose that $\mathcal{D}$ is nonatomic. There exists a constant $\gamma > 0$ such that, for every $n \geq 1$, every possibly randomized validation algorithm $\mathcal{V}$, and every reported ranking $\widehat{\sigma}$ over $\{a, b\}$, there exists a measurable utility function $u^\star : \mathcal{X} \times \{a, b\} \to [0, 1]$ satisfying $|\mathbb{E}_{x \sim \mathcal{D}} [u^\star(x, a) - u^\star(x, b)]| \geq \gamma$ and $\mathrm{ValErr}_{u^\star, n}(\mathcal{V}, \widehat{\sigma}) \geq \frac{1}{2}$. Consequently, ranking error is not statistically justifiable, even with a fixed positive mean-utility margin.*

The proof constructs a partition of the context space into many cells of equal probability and assigns each cell a utility gap of either 1 or $-1/2$. The assignment probabilities are chosen so that the pooled win probability is $1/2$ while a has a positive mean-utility advantage, and reversing every gap reverses this advantage. For any fixed sample size, the partition can be made fine enough that observing the same cell twice is unlikely. After averaging over the unknown cell assignments, outcomes from distinct cells appear as independent fair coins in both worlds, even though the population rankings are opposite with high probability. Observing the sampled contexts therefore reveals little about the unobserved assignments. The formal argument appears in Section B.2. We next consider a distinct source of information loss: a ranking method may discard context before the validator ever observes its output.

Limitations of pooled win-rate aggregation. Context-independent Bradley–Terry fitting and other pooled-win methods retain only $\overline{p}_{ab} := \mathbb{E}_{x \sim \mathcal{D}} [p_{ab}(x)]$, the marginal probability that a beats b. In the utility-based setting, however, the target order of a and b depends on the sign of $\mathbb{E}_{x \sim \mathcal{D}} [u^\star(x, a) - u^\star(x, b)]$. Because the Bradley–Terry link is nonlinear, the average win probability does not generally reveal the average utility gap.

Proposition 3 (Pooled comparisons do not identify the mean-utility ranking)**.** *Fix $\beta > 0$ and two models a, b. There exist two contextual Bradley–Terry populations, denoted $+$ and $-$, such that $\overline{p}^{+}_{ab} = \overline{p}^{-}_{ab} = \frac{1}{2}$, $\mathbb{E}_{x\sim\mathcal{D}}\left[u^{\star}_{+}(x,a) - u^{\star}_{+}(x,b)\right] = g$, $\mathbb{E}_{x\sim\mathcal{D}}\left[u^{\star}_{-}(x,a) - u^{\star}_{-}(x,b)\right] = -g$ for some $g > 0$. Hence, even exact knowledge of all pooled win probabilities does not identify the mean-utility ranking. Any possibly randomized ranking rule that depends only on the pooled distribution has misranking probability at least $1/2$ in one of the two populations. Likewise, for a fixed reported ranking, a validator with only pooled access has validation error at least $1/2$ in one population.*

The full proof is deferred in Section B.3. Indeed, under Bradley–Terry, the full conditional win probabilities identify each contextual utility gap through $u^{\star}(x,a) - u^{\star}(x,b) = \beta^{-1}\log\big(p_{ab}(x)/(1-p_{ab}(x))\big)$. Thus the full function $p_{ab}(x)$ identifies every contextual utility difference, whereas its average $\overline{p}_{ab}$ need not identify the average utility difference.

Proposition 3 is therefore an information-loss result. By contrast, Proposition 2 is a finite-sample result, as even when the full contextual distribution identifies the target, unrestricted variation across contexts prevents a uniform finite-sample validator from estimating it.

3.2 Test-based evaluation metrics

The preceding results show that finitely many comparisons may not suffice to certify agreement with the ground-truth ranking. Nevertheless, a ranking procedure can be assessed against explicit requirements for its outputs and behavior, such as agreement with observed preferences or consistency across candidate sets. Violations identify concrete failures without requiring the validator to recover the ground-truth ranking. We therefore define evaluation metrics through tests tailored to the information available to the validator. The chosen tests specify which failures the metric measures and hence what a validation guarantee establishes.

Formally, given the information o available to the validator and a comparison $z = (x, a, b, y)$, a *test* is a measurable score $\phi(o, z) \in [-1, 1]$ that the validator can compute. Its population value is $t_\phi(o, \mathcal{P}) := \mathbb{E}_{z\sim\mathcal{P}}\left[\phi(o, z)\right]$. A test may depend only on o, allowing checks of relations among the algorithm outputs available to the validator.

The validator estimates each test value by averaging its score over fresh comparisons. Independence from the algorithm run allows standard concentration bounds to control these estimates simultaneously for a fixed finite collection of tests. A Lipschitz aggregation of their values preserves this and give the following guarantee.

Theorem 1 (Finite-test metrics are justifiable)**.** *Suppose an error metric is determined by $l < \infty$ tests. Given access output o, the validator can identify and evaluate measurable query functions $\phi_1(o,\cdot), \ldots, \phi_l(o,\cdot)$ taking values in $[-1, 1]$. Let $t_j := \mathbb{E}_{z\sim\mathcal{P}}\left[\phi_j(o, z)\right]$ and suppose $f(o, \mathcal{P}) = g(t_1, \ldots, t_l)$ where g is a known aggregation function that is c-Lipschitz in the sup norm for some fixed $c < \infty$. Then f is statistically justifiable.*

We distinguish two categories of tests: *outcome tests*, which assess agreement with observed preferences, and *behavioral tests*, which assess properties of the ranking procedure. Which tests can be evaluated further depends on the information available to the validator. We develop a few metrics based on each category and then show how outcome and behavioral information can be combined to construct more informative tests.

Prediction access: multicalibration. Suppose the validator observes a predicted win-probability function q in addition to the reported ranking. A single overall average can hide errors that cancel across contexts or prediction levels. We therefore compare predictions with outcomes within specified context groups and ranges of predicted probabilities.

Let $\mathcal{G}$ be a collection of measurable context groups, representing, for example, tasks, prompt classes, domains, or user populations. Fix a finite partition $\mathcal{B}$ of $[0, 1]$ into prediction bins.

For distinct models a, b, a group $G \in \mathcal{G}$, and a bin $B \in \mathcal{B}$, define the cell $C^{q}_{a,b,G,B} := \{x \in G : q_{ab}(x) \in B\}$, consisting of contexts in G whose predicted pairwise win probability falls in B.

Following the probability-weighted formulation of calibration and multicalibration in (**?**, Sections 3.1 and 4.6.1), we weight conditional calibration errors by the probabilities of their corresponding events. This avoids requiring accurate conditional estimates on events with arbitrarily small probability. The probability-weighted pairwise multicalibration error is

$$\mathrm{MCE}_{\mathcal{G},\mathcal{B}}(q, \mathcal{P}) := \sup_{a\neq b,\ G\in\mathcal{G},\ B\in\mathcal{B}} \left|\mathbb{E}_{(x,a',b',y)\sim\mathcal{P}}\left[\mathbb{1}\{(a', b') = (a, b),\ x \in C^{q}_{a,b,G,B}\}\big(y - q_{ab}(x)\big)\right]\right|.$$

Each term weights the conditional calibration error by the probability of observing the corresponding model pair and context cell. Therefore, events of zero probability contribute zero automatically. When $\mathcal{G}$ is finite, this is the maximum absolute value of finitely many tests and is therefore statistically justifiable by Theorem 1.

Ranking access: compatibility. When only σ is reported, the validator has no probabilities to calibrate directly. It can instead ask whether the ranking has a calibrated explanation: is there a candidate win-probability function that induces σ and agrees with the validation comparisons on the chosen cells? Fix a class $\mathcal{Q}$ of candidate win-probability functions and an aggregation rule R, and define $\mathcal{Q}_\sigma := \{q \in \mathcal{Q} : R(q, \mathcal{D}) = \sigma\}$. We measure the ranking's compatibility with the calibration tests by

$$\mathrm{RankCalErr}_{\mathcal{G},\mathcal{B}}(\sigma, \mathcal{P}) := \inf_{q \in \mathcal{Q}_\sigma} \mathrm{MCE}_{\mathcal{G},\mathcal{B}}(q, \mathcal{P}),$$

setting the error to 1 if $\mathcal{Q}_\sigma$ is empty.

The infimum chooses the most calibrated explanation among the candidates that induce the reported ranking. A large value means that every such explanation has a large probability-weighted calibration error for at least one model pair and context cell. If $p \in \mathcal{Q}$, then $\mathrm{RankCalErr}_{\mathcal{G},\mathcal{B}}(\sigma^\star, \mathcal{P}) = 0$ for $\sigma^\star = R(p, \mathcal{D})$, because p is a feasible, perfectly calibrated explanation.

Validation requires simultaneous estimation of the test values across candidate functions. The following finite-class condition suffices.

Corollary 1. *Suppose $\mathcal{Q}$ and $\mathcal{G}$ are finite. Assume that, given σ, the validator can identify $\mathcal{Q}_\sigma$ and evaluate each of its candidate functions on observed contexts. Then* $\mathrm{RankCalErr}_{\mathcal{G},\mathcal{B}}$ *is statistically justifiable from ranking access.*

The corollary follows from Theorem 1. For nonempty $\mathcal{Q}_\sigma$, the error is a minimum of maxima of absolute values of at most $|\mathcal{Q}|M(M-1)|\mathcal{G}||\mathcal{B}|$ test values, and this aggregation is 1-Lipschitz in the sup norm. If $\mathcal{Q}_\sigma$ is empty, the error is the known constant 1.

Algorithm access: evidence consistency. Outcome tests assess agreement with observed preferences. Algorithm access also allows us to test behavioral properties, such as robustness to changes in the candidate set. Inspired by independence of irrelevant alternatives (IIA),[1] we focus on *evidence consistency*: whether the relative order of two models remains unchanged when other candidates and their associated comparisons are added or removed, while the direct comparison evidence for the pair is held fixed. A reversal reveals sensitivity to the surrounding candidates.

Suppose the validator observes $o = (\mathcal{A}, S)$ and can rerun $\mathcal{A}$ on restrictions of S. Let $\mathcal{K}$ be a finite collection of candidate sets. For each $K \in \mathcal{K}$, define $S_K := \{(x, a, b, y) \in S : a, b \in K\}$ and write $\sigma^S_{\mathcal{A},K} := \mathcal{A}(K, S_K)$ for the resulting ranking. We fix tie-breaking and any random seeds used in the reruns. Define $\mathcal{K}_{ab} := \{K \in \mathcal{K} : \{a, b\} \subseteq K\}$. The evidence-consistency error on S is

$$\mathrm{ECErr}_{\mathcal{K}}((\mathcal{A}, S), \mathcal{P}) := \max_{\substack{a \neq b \\ K, L \in \mathcal{K}_{ab}}} \left|(\sigma^S_{\mathcal{A},K})_{ab} - (\sigma^S_{\mathcal{A},L})_{ab}\right|.$$

The error is zero exactly when every pair retains the same relative order across the specified candidate sets containing it. Because each rerun uses a restriction of the same dataset, the direct comparison evidence for a shared pair is unchanged. The metric measures consistency on the supplied evidence S and is independent of $\mathcal{P}$.

Each discrepancy in the maximum is a bounded test that depends only on $o = (\mathcal{A}, S)$. Given access to $(\mathcal{A}, S)$, the validator can compute these discrepancies and their maximum exactly by rerunning $\mathcal{A}$ on each (K, S_K). Thus, $\mathrm{ECErr}_{\mathcal{K}}$ is statistically justifiable with no validation samples.

Outcome and behavioral information can also be combined. Algorithm access allows us to refine calibration cells using predictions across candidate sets. Joint candidate-set multicalibration (JMCE) tests each prediction on contexts selected by the entire pattern of predictions. These tests can expose errors that cancel when each candidate set is evaluated separately.

Given $o = (\mathcal{A}, S)$, suppose each rerun on (K, S_K) also returns a win-probability function $q^{o,K}$, with any randomness fixed as above. For distinct models a, b, a group $G \in \mathcal{G}$, and a vector of bins $\mathbf{B} = (B_L)_{L \in \mathcal{K}_{ab}} \in \mathcal{B}^{\mathcal{K}_{ab}}$, define the joint cell $C^o_{a,b,G,\mathbf{B}} := \left\{x \in G : q^{o,L}_{ab}(x) \in B_L \text{ for every } L \in \mathcal{K}_{ab}\right\}$. For example, a joint cell may select contexts where one candidate set yields a high predicted probability that a beats b, while another yields a low probability. The probability-weighted joint candidate-set multicalibration error is

$$\mathrm{JMCE}_{\mathcal{K},\mathcal{G},\mathcal{B}}((\mathcal{A}, S), \mathcal{P}) := \sup_{\substack{a \neq b,\ G \in \mathcal{G} \\ \mathbf{B} \in \mathcal{B}^{\mathcal{K}_{ab}},\ K \in \mathcal{K}_{ab}}} \left|\mathbb{E}_{(x,a',b',y)\sim\mathcal{P}}\left[\mathbb{1}\{(a', b') = (a, b),\ x \in C^o_{a,b,G,\mathbf{B}}\}\big(y - q^{o,K}_{ab}(x)\big)\right]\right|.$$

[1] Arrow's IIA compares preference profiles on a fixed candidate set; our condition compares restrictions of a fixed dataset. Arrow's impossibility theorem rules out jointly satisfying IIA, Pareto efficiency, and nondictatorship on an unrestricted preference domain. It does not preclude consistency on a given dataset and collection of candidate sets. In the utility-based setting, ordering models by their population-average utilities provides a candidate-set-independent target, although an estimated ranking need not share this property.

Each term compares one candidate-set prediction with outcomes within a joint cell, weighted by the probability of observing the corresponding pair and cell. Events of zero probability contribute zero automatically. The predictions and cells are determined by o, while independent validation comparisons allow us to estimate their calibration errors.

Corollary 2. *Suppose $\mathcal{K}$, $\mathcal{G}$, and $\mathcal{B}$ are finite. Assume that, given $o = (\mathcal{A}, S)$, the validator can obtain and evaluate each $q^{o,K}$ and the resulting joint-cell indicators. Then $\mathrm{JMCE}_{\mathcal{K},\mathcal{G},\mathcal{B}}$ is statistically justifiable from algorithm-and-prediction access.*

For fixed o, each pair, group, joint bin vector, and candidate set specifies a bounded test. The metric is the maximum absolute value of finitely many such test expectations, a 1-Lipschitz aggregation in the sup norm. The result therefore follows from Theorem 1, without a lower bound on joint-cell probabilities.

4 The Relationship between Log-loss, Ranking Error, and Test-based Metrics

The previous section developed test-based metrics for validation. Yet their relationship to standard quantities such as log-loss and Kendall–tau distance is not immediate. We now ask when a guarantee for one of these metrics implies a guarantee for the others.

4.1 Implications of Maximum Likelihood

We first examine log-loss, the objective minimized by maximum-likelihood estimation of comparison probabilities. Low log-loss indicates that the predicted and true win probabilities are close on average over the contexts and model pairs that occur. Write $\mathcal{P} = \mathcal{P}_{\mathcal{D},p}$ for the true comparison distribution. The population excess log-loss of a prediction q is the expected Bernoulli KL divergence $\ell(q,\mathcal{P}) - \ell(p,\mathcal{P}) = \mathbb{E}_{x\sim\mathcal{D},\ (a,b)\sim\Pi}\left[\mathrm{KL}\big(\mathrm{Ber}(p_{ab}(x)) \,\big\|\, \mathrm{Ber}(q_{ab}(x))\big)\right]$.

Proposition 4 (Consequences of small excess log-loss)**.** *Let $\mathcal{P} = \mathcal{P}_{\mathcal{D},p}$ and suppose $\ell(q,\mathcal{P}) - \ell(p,\mathcal{P}) \le \varepsilon$. Then:*

(i) the mean-squared prediction error satisfies $\mathbb{E}_{x\sim\mathcal{D},\,(a,b)\sim\Pi}\left[(p_{ab}(x) - q_{ab}(x))^2\right] \le \frac{\varepsilon}{2}$.

(ii) for every finite $\mathcal{G}$ and $\mathcal{B}$, $\mathrm{MCE}_{\mathcal{G},\mathcal{B}}(q,\mathcal{P}) \le \sqrt{\frac{\varepsilon}{2}}$.

(iii) for a fixed pair (a,b) and measurable C with $\mathcal{D}(C) > 0$, $|\mathbb{E}_{x\sim\mathcal{D}}\left[p_{ab}(x) - q_{ab}(x) \mid x \in C\right]| \le \sqrt{\frac{\varepsilon}{2\Pi(a,b)\mathcal{D}(C)}}$.

(iv) under Borda aggregation, with $\pi_{\min} := \min_{a\neq b}\Pi(a,b)$, $|s_m^{\mathrm{B}}(q,\mathcal{D}) - s_m^{\mathrm{B}}(p,\mathcal{D})| \le \sqrt{\frac{\varepsilon}{2(M-1)\pi_{\min}}}$ for every m. Hence the Borda ranking induced by q equals the true Borda ranking whenever every pair of true scores differs by more than $\sqrt{2\varepsilon/((M-1)\pi_{\min})}$.

Proof. Pinsker's inequality applied pointwise to the Bernoulli distributions gives (i). Part (ii) follows by bounding each probability-weighted residual with Cauchy–Schwarz. Applying the same inequality after restricting to the event $\{(a',b') = (a,b), x \in C\}$ gives (iii). For (iv), apply Cauchy–Schwarz first across the $M-1$ opponents and then use $\Pi(m,b) \ge \pi_{\min}$. Each predicted score is within the displayed radius of its true value, so a true gap larger than twice that radius cannot reverse. □

The bounds clarify how rare pairs and cells affect the guarantees. Global MCE remains controlled without a minimum cell probability because it is probability weighted, whereas the unweighted conditional residual in part (iii) deteriorates as either the pair or cell becomes rare. Ranking requires the additional score-gap condition in part (iv), because ordering is discontinuous at a tie.

These observations apply to any fitted model with small excess log-loss. For a finite, well-specified class $\mathcal{H}$ whose predictions lie in $[\eta, 1-\eta]$, empirical log-loss minimization over n samples yields, with probability at least $1-\delta$, $\ell(\widehat{q},\mathcal{P}) - \ell(p,\mathcal{P}) \le \frac{2\log(|\mathcal{H}|/\delta)}{\eta(1-\eta)n}$. The calibration and ranking implications above then hold with this value of ε. Under misspecification, however, even a population log-loss minimizer need not recover the true win probabilities. The following result shows that this failure can affect both calibration and ranking.

Theorem 2 (Failure under misspecification)**.** *There exists a contextual Bradley–Terry instance with three models and uniform pair sampling, and a misspecified class $\mathcal{H}$ of context-independent Bradley–Terry win-probability functions, such that a population log-loss minimizer over $\mathcal{H}$ has pairwise multicalibration error greater than 0.035, even for $\mathcal{G} = \{\mathcal{X}\}$. Ranking models by its fitted scores has normalized Kendall–tau error $1/3$ relative to the true mean-utility ranking.*

The failure arises from fitting a single score vector to heterogeneous preferences. Although comparisons follow a Bradley–Terry model within each context, their pooled win probabilities need not admit a context-independent Bradley–Terry representation. Population log-loss minimization then finds a compromise across model pairs, which can leave nonzero pairwise calibration errors even over the entire context population. Moreover, win probabilities depend nonlinearly on contextual utility differences, so fitting their averages need not preserve the ordering of mean utilities. The example exhibits both failures: persistent calibration error and one reversed pair among three models, giving normalized Kendall–tau error $1/3$.

4.2 Calibration implies maximum likelihood in an indistinguishable world.

We next examine calibration, which only constrains the tests selected by $\mathcal{G}$ and $\mathcal{B}$. A small MCE therefore does not imply that q is close to the true win probabilities everywhere. It can only validate that q agrees with the truth on those selected tests.

This narrower claim has a useful interpretation. Given a prediction q, form the prediction-induced world $\mathcal{P}_{\mathcal{D},q}$ by keep the same context and pair distributions, but generate each outcome with probability $q_{ab}(x)$. In that world q minimizes population log-loss by construction. Small MCE means that the specified tests cannot distinguish this world from the true one. The following statement makes the comparison quantitative.

Theorem 3 (Calibration yields an indistinguishable world)**.** *Suppose that* $\mathrm{MCE}_{\mathcal{G},\mathcal{B}}(q) \le \alpha$*. Then, for every distinct* $a, b \in \mathcal{M}$ *and every calibration cell* $C = C^{q}_{a,b,g,B}$ *with* $\mathcal{D}(C) > 0$*,* $\left|\mathbb{E}_{\mathcal{P}_{\mathcal{D},p}}[y_{ab} \mid x \in C] - \mathbb{E}_{\mathcal{P}_{\mathcal{D},q}}[y_{ab} \mid x \in C]\right| \le \alpha$*. Moreover,* q *has zero population excess log-loss in* $\mathcal{P}_{\mathcal{D},q}$*,* $\ell(q, \mathcal{P}_{\mathcal{D},q}) - \inf_r \ell(r, \mathcal{P}_{\mathcal{D},q}) = 0$*, where the infimum ranges over all win-probability functions* r*. Thus exact pairwise multicalibration gives a world that agrees exactly with the true world on these calibration tests and in which* q *minimizes population log-loss.*

Proof. The distributions $\mathcal{P}_{\mathcal{D},p}$ and $\mathcal{P}_{\mathcal{D},q}$ have the same context and pair distributions. Since the conditional outcome mean under $\mathcal{P}_{\mathcal{D},q}$ is $q_{ab}(x)$, $\mathbb{E}_{\mathcal{P}_{\mathcal{D},q}}[y_{ab} \mid x \in C] = \mathbb{E}_{\mathcal{P}_{\mathcal{D},p}}[q_{ab}(x) \mid x \in C]$. Consequently, $\left|\mathbb{E}_{\mathcal{P}_{\mathcal{D},p}}[y_{ab} \mid x \in C] - \mathbb{E}_{\mathcal{P}_{\mathcal{D},q}}[y_{ab} \mid x \in C]\right| = \left|\mathbb{E}_{\mathcal{P}_{\mathcal{D},p}}[y_{ab} - q_{ab}(x) \mid x \in C]\right| \le \alpha$. For any win-probability function r, the excess log-loss identity gives $\ell(r, \mathcal{P}_{\mathcal{D},q}) - \ell(q, \mathcal{P}_{\mathcal{D},q}) = \mathbb{E}_{x\sim\mathcal{D},\ (a,b)\sim\Pi}\left[\mathrm{KL}(\mathrm{Ber}(q_{ab}(x)) \,\|\, \mathrm{Ber}(r_{ab}(x)))\right] \ge 0$. Since $r = q$ attains equality, q minimizes population log-loss in $\mathcal{P}_{\mathcal{D},q}$. □

Agreement on the selected tests does not identify the full comparison distribution, so the two worlds may still differ outside the tested cells. In particular, even if $q = F(u)$ for an admissible utility function u, the prediction-induced world may have a different target ranking from the true world. Recovering the ranking from these tests therefore requires additional structure. The experiments below compare prediction, calibration, and ranking assessments for fitted Bradley–Terry models.

5 Experiment

We study fitted Bradley–Terry (BT) models first in a resampling experiment with a known population ranking, with Arena comparisons.

5.1 Ranking

Recovery under shrinking margins

Our theory predicts that small prediction error preserves ranking only when true scores are sufficiently separated. To vary this separation while holding the comparison design fixed, we build a Mallows model from Chatbot Arena outcomes.

Table 1: Results across 1,000 runs. Parentheses report the Monte Carlo standard error of the Kendall error. "Top" and "Exact" denote recovery probabilities. Each normalized-error entry reports the mean, followed in brackets by the fraction of runs below one.

Δ_θ	Kendall error	Between-fit	Top	Exact	NMCE
0.002947	0.2794 (0.0026)	0.3757	0.274	0.000	9.286 [0.000]
0.005894	0.1641 (0.0018)	0.2374	0.439	0.000	4.686 [0.000]
0.011786	0.0737 (0.0010)	0.1146	0.645	0.004	2.425 [0.003]
0.023558	0.0233 (0.0005)	0.0409	0.883	0.190	1.336 [0.183]
0.047011	0.0024 (0.0002)	0.0047	0.982	0.848	0.814 [0.848]

After discarding ties and `both_bad` outcomes, we retain the 12 most compared models and the $n = 12{,}852$ decisive comparisons among them, covering all $\binom{12}{2} = 66$ pairs. An unregularized BT fit on these outcomes yields scores $\widehat{s} \in \arg\max_{\sum_m s_m = 0} \sum_{a<b} [w_{ab} \log \sigma(s_a - s_b) + (n_{ab} - w_{ab}) \log \sigma(s_b - s_a)]$, where w_{ab} and n_{ab} are the win and comparison counts. Sorting $\widehat{s}$ gives the central ranking σ_0.

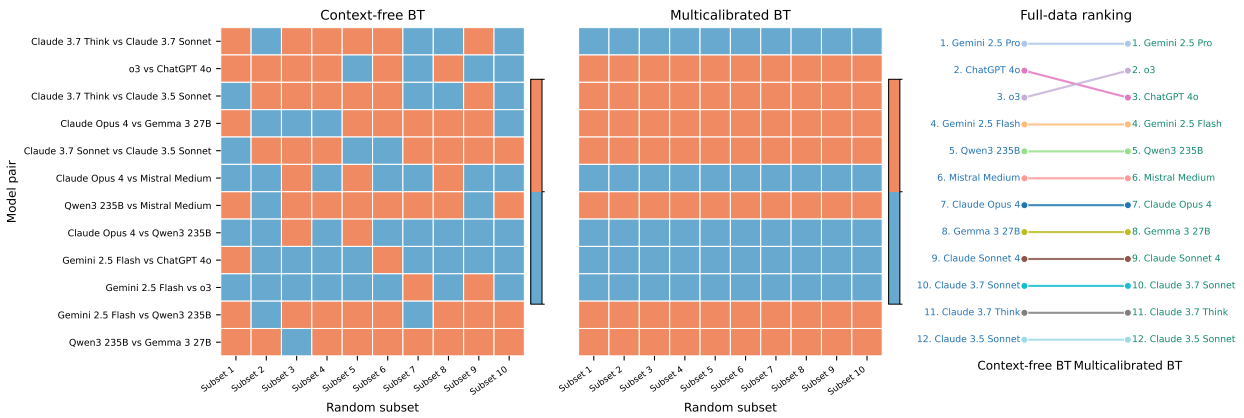


Figure 3: Pairwise ordering of selected model pairs across ten stratified random subsets. The context-free BT model is refitted on every subset (left); the multicalibrated reference ranking is held fixed while its probabilities are recalibrated (middle). Orange and blue indicate opposite relative orders. Right: full-data rankings under the two methods.

We then use the Kendall–Mallows distribution $P_\theta(\pi \mid \sigma_0) = Z_M(\theta)^{-1} \exp\{-\theta d_K(\pi, \sigma_0)\}$, fitting θ by pairwise marginal likelihood with σ_0 held fixed. Setting $\theta_c = c\widehat{\theta}$ for $c \in \{1/16, 1/8, 1/4, 1/2, 1\}$ yields populations that share the same ranking σ_0 but have progressively smaller pairwise margins $\Delta_{\theta_c} := \min_{a \neq b} \left| p_{ab}^{\theta_c} - \frac{1}{2} \right|$. For each concentration and each of $R = 1{,}000$ repetitions, we keep the Arena counts n_{ab} and redraw $W_{ab}^{(r)} \sim \mathrm{Binomial}\big(n_{ab}, p_{ab}^{\theta_c}\big)$, then refit BT. With $\widehat{q}_{ab}^{(r)}$ the fitted pairwise probability, the normalized multicalibration error is $\mathrm{NMCE}^{(r)} = \max_{a<b} \frac{\left| p_{ab}^{\theta_c} - \widehat{q}_{ab}^{(r)} \right|}{\left| p_{ab}^{\theta_c} - 1/2 \right|}$. Table 1 also reports mean Kendall error against σ_0, between-fit disagreement, and top-model and exact-ranking recovery rates.

As predicted by Propositions 1 and 4, recovery improves sharply with margin. At the smallest margin, mean Kendall error is 0.2794 and the exact ranking is never recovered; at the largest, these become 0.0024 and 84.8%. The fraction of runs with $\mathrm{NMCE} < 1$ tracks exact recovery, since $\mathrm{NMCE} < 1$ keeps every pairwise probability on the correct side of $1/2$.

5.2 Evidence consistency across subsets

5.3 Group calibration

On the original Arena data the population ranking is unknown, so we instead probe evidence consistency. We form ten stratified random subsets and refit a context-free BT model on each. As a comparison, a multicalibrated predictor keeps the full-data ranking fixed and only recalibrates probabilities on each subset.

Figure 3 shows that context-free BT reverses several pairs across subsets, including highly ranked models. The fixed multicalibrated reference has no subset reversals by construction. The two full-data rankings also reverse ChatGPT 4o and o3, so calibration can change the fitted order, but this alone does not establish greater evidence consistency under full refitting.

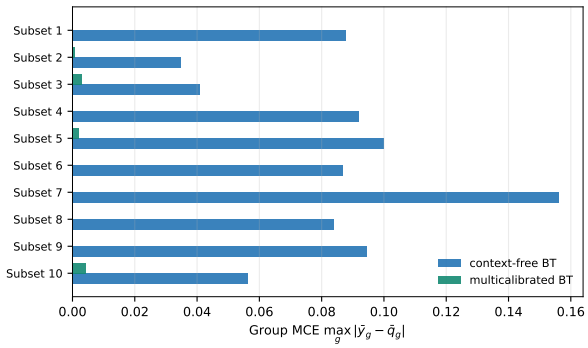


Figure 2: Empirical maximum group calibration residual for context-free BT and the subset-recalibrated predictor on the same ten subsets.

Figure 2 reports the empirical maximum group residual $\widehat{\mathrm{GCE}} := \max_{g:\, n_g \geq 25} |\overline{y}_g - \overline{q}_g|$ on the same subsets. Context-free BT shows residuals of several percentage points on most subsets. The comparison predictor, recalibrated on each subset from fixed full-data scores, is nearly calibrated there by construction. The gap shows that a globally fitted score model can remain poorly calibrated on particular population slices.

AI USE STATEMENT

Generative AI tools were used to assist with writing and proofreading mathematical arguments, implementing the experimental methods, writing and editing experiment code, and revising the manuscript for readability. They were not used to generate research ideas, to identify research topics or gaps, or to conduct the research autonomously. The authors reviewed all AI-assisted contributions: we checked the proofs, verified and tested the experiment code, and reviewed the revised manuscript. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ameli, S., Zhuang, S., Stoica, I., and Mahoney, M. W. (2025). A statistical framework for ranking LLM-based chatbots. In *International Conference on Learning Representations*.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346.

Bong, H. and Rinaldo, A. (2022). Generalized results for the existence and consistency of the MLE in the Bradley–Terry–Luce model. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2160–2177.

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

Chen, P., Gao, C., and Zhang, A. Y. (2022). Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Daynauth, R., Clarke, C., Flautner, K., Tang, L., and Mars, J. (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.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. (2024). Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling*.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. (2021). Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108.

Fan, J., Hou, J., and Yu, M. (2024). Uncertainty quantification of MLE for entity ranking with covariates. *Journal of Machine Learning Research*, 25(358):1–83.

Frick, E., Chen, C., Tennyson, J., Li, T., Chiang, W.-L., Angelopoulos, A. N., and Stoica, I. (2025). Prompt-to-leaderboard: Prompt-adaptive LLM evaluations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 17672–17689.

Garbacea, C., Wang, H., and Tan, C. (2026). Personalized benchmarking: Evaluating LLMs by individual preferences. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 639–675. Association for Computational Linguistics.

Gölz, P., Haghtalab, N., and Yang, K. (2025). Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems*, volume 38.

Haghtalab, N., Procaccia, A. D., Shao, H., Wang, S. L., and Yang, K. (2026). Pluralistic leaderboards.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948.

Huang, J. Y., Shen, Y., Wei, D., and Broderick, T. (2026). Dropping just a handful of preferences can change top large language model rankings. In *International Conference on Learning Representations*.

Liu, Z., Li, J., Zhuang, Y., Liu, Q., Shen, S., Ouyang, J., Cheng, M., and Wang, S. (2025). am-ELO: A stable framework for arena-based LLM evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 38857–38868.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York.

Noothigattu, R., Peters, D., and Procaccia, A. D. (2020). Axioms for learning from pairwise comparisons. In *Advances in Neural Information Processing Systems*, volume 33.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Oyarhoseini, H., Lin, J., and Karimi, A.-H. (2026). A unified perturbation framework for analyzing leaderboard stability and manipulation.

Seshadri, A. and Ugander, J. (2019). Fundamental limits of testing the independence of irrelevant alternatives in discrete choice. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 65–66.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. J. (2016). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47.

Shah, N. B. and Wainwright, M. J. (2018). Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Related Work

Ranking from pairwise comparisons is approached in mainly two ways: parametric estimation and social-choice methods. Parametric estimation includes the Bradley-Terry model and Luce's choice rule Bradley and Terry (1952); Luce (1959). Subsequent work characterizes how estimation accuracy depends on the comparison design and establishes conditions for maximum-likelihood estimation under sparse or irregular observations Shah et al. (2016); Bong and Rinaldo (2022). Under the Bradley-Terry-Luce model, the minimax rate of full-ranking recovery under Kendall tau loss transitions between an exponential and a polynomial regime Chen et al. (2022). The asymptotic distribution of the maximum-likelihood estimator has been derived when the socre has a covariate term Fan et al. (2024). Social-choice methods define a ranking by aggregating pairwise preferences,rather than by estimating latent scores. Borda count is one such method: it scores each item by its expected win count, and the resulting ranking has robust recovery guarantees Shah and Wainwright (2018). Arrow (1950) characterizes the axiomatic limits of preference aggregation and Noothigattu et al. (2020) studies the properties of rankings learned from pairwise data. Closest to our focus on validation, Seshadri and Ugander (2019) establish finite-sample limits for testing independence of irrelevant alternatives.

These approaches have been applied to ranking LLMs from human preferences. Chatbot Arena established large-scale pairwise human evaluation with Bradley-Terry aggregation Chiang et al. (2024), and follow-up work has developed improved treatments of ties and uncertainty Ameli et al. (2025), prompt-adaptive leaderboards Frick et al. (2025), comparisons of alternative ranking procedures Daynauth et al. (2025), and models of annotator reliability Liu et al. (2025). Leaderboard positions can change after a small number of targeted vote deletions Huang et al. (2026), removing a candidate can reorder the remaining models Oyarhoseini et al. (2026), and global aggregation can obscure systematic heterogeneity across users Haghtalab et al. (2026). These findings motivate formal criteria for determining which claims about a leaderboard are supported by observable evidence.

B Omitted Technical Details and Proofs

B.1 Small-margin sensitivity

Fix $\mathcal{M} = \{a, b\}$, $\beta > 0$, and $\widehat{\sigma} = (a \succ b)$. Let $\mathcal{D}$ be concentrated on one context, whose argument we omit. For $\Delta \in (0, 1]$, define $u^\star_\pm(a) = \frac{1}{2} \pm \frac{\Delta}{2}$ and $u^\star_\pm(b) = \frac{1}{2} \mp \frac{\Delta}{2}$, and write $\mathcal{P}_\pm := \mathcal{P}_{\mathcal{D},F(u^\star_\pm)}$. We show that, for every $n \geq 1$ and every possibly randomized validation algorithm $\mathcal{V}$,

$$\max_{\omega \in \{+,-\}} \mathrm{ValErr}_{u^\star_\omega, n}(\mathcal{V}, \widehat{\sigma}) \geq \frac{1}{2}\left(1 - \min\left\{1, \frac{\beta\Delta\sqrt{n}}{2}\right\}\right).$$

Proof. For the ordered pair (a, b), the comparison probabilities in the two worlds are $(1 + \exp(-\beta\Delta))^{-1}$ and $(1 + \exp(\beta\Delta))^{-1}$. For (b, a), these probabilities are swapped. The conditional Bernoulli KL divergence is the same for both ordered pairs. Since the context and pair distributions agree between the two worlds,

$$\mathrm{KL}(\mathcal{P}_+ \| \mathcal{P}_-) = \beta\Delta\left(\frac{1}{1 + \exp(-\beta\Delta)} - \frac{1}{1 + \exp(\beta\Delta)}\right) \leq \frac{\beta^2\Delta^2}{2},$$

where the inequality uses the fact that the logistic function has derivative at most $1/4$.

By Pinsker's inequality and additivity of KL divergence for independent samples,

$$\mathrm{TV}(\mathcal{P}_+^n, \mathcal{P}_-^n) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_+^n \| \mathcal{P}_-^n)} = \sqrt{\frac{n}{2}\mathrm{KL}(\mathcal{P}_+ \| \mathcal{P}_-)} \leq \frac{\beta\Delta\sqrt{n}}{2}.$$

The reported ranking $\widehat{\sigma}$ is correct in world $+$ and incorrect in world $-$. Thus its validation errors are the rejection probability in world $+$ and the acceptance probability in world $-$. The difference between the two acceptance probabilities is at most $\mathrm{TV}(\mathcal{P}_+^n, \mathcal{P}_-^n)$, so

$$\mathrm{ValErr}_{u^\star_+, n}(\mathcal{V}, \widehat{\sigma}) + \mathrm{ValErr}_{u^\star_-, n}(\mathcal{V}, \widehat{\sigma}) \geq 1 - \mathrm{TV}(\mathcal{P}_+^n, \mathcal{P}_-^n).$$

This also holds for randomized validators, since applying the same randomized decision rule cannot increase total variation. At least one error is at least half their sum. Combining the preceding inequalities with the bound $\mathrm{TV}(\mathcal{P}_+^n, \mathcal{P}_-^n) \leq 1$ proves the claim.

Finally, fix $\delta \in (0, 1/2)$. For every finite n, choosing $\Delta > 0$ sufficiently small that $\beta\Delta\sqrt{n}/2 < 1 - 2\delta$ makes the lower bound strictly greater than δ. With $\varepsilon = \gamma = 1/2$, the two ranking errors, zero and one, require opposite validation decisions. Therefore no validation algorithm admits a finite sample bound that guarantees error at most δ uniformly over these utilities, establishing failure of statistical justifiability. □

B.2 Proof of Proposition 2

Proof of Proposition 2. Fix n, $\mathcal{V}$, and $\widehat{\sigma}$. Assume without loss of generality that $\widehat{\sigma} = (a \succ b)$.

For r near $1/2$, define

$$\alpha(r) := \frac{r - (1 + \exp(\beta/2))^{-1}}{(1 + \exp(-\beta))^{-1} - (1 + \exp(\beta/2))^{-1}}.$$

At $r = 1/2$,

$$\alpha(1/2) = \frac{e^{\beta} + 1}{2(e^{\beta} + e^{\beta/2} + 1)} \in (1/3, 1/2).$$

Set $\gamma := (3\alpha(1/2) - 1)/4 > 0$.

Consider two distributions of utility gaps. In the positive construction, the gap equals 1 with probability $\alpha(r)$ and $-1/2$ otherwise. In the negative construction, it equals -1 with probability $\alpha(1 - r)$ and $1/2$ otherwise. Both constructions have expected Bradley–Terry win probability r. Their expected utility gaps are, respectively,

$$\frac{3\alpha(r) - 1}{2} \qquad \text{and} \qquad -\frac{3\alpha(1 - r) - 1}{2}.$$

At $r = 1/2$, these equal 2γ and -2γ. By continuity, there is an open interval containing $1/2$ on which both constructions are well-defined and their expected gaps are greater than γ and less than $-\gamma$, respectively.

Since $\mathcal{D}$ is nonatomic, partition $\mathcal{X}$ into N measurable cells of probability $1/N$. Independently assign a gap Δ_i to each cell using either the positive or the negative construction. On cell i, define $u(x, a) = (1 + \Delta_i)/2$ and $u(x, b) = (1 - \Delta_i)/2$. These utilities are measurable and take values in $[0, 1]$. Their mean-utility gap is $N^{-1} \sum_{i=1}^{N} \Delta_i$. For each fixed r in the interval above, the law of large numbers therefore gives

$$\mathbb{P}_+ \left(\mathbb{E}_{x \sim \mathcal{D}} \left[u(x, a) - u(x, b) \right] \geq \gamma \right) \longrightarrow 1,$$
$$\mathbb{P}_- \left(\mathbb{E}_{x \sim \mathcal{D}} \left[u(x, a) - u(x, b) \right] \leq -\gamma \right) \longrightarrow 1$$

as $N \to \infty$, where these probabilities are over the utility assignments.

Let $A_n(r)$ denote the acceptance probability of $\mathcal{V}(T, \widehat{\sigma})$ when contexts and model pairs follow $\mathcal{D}$ and Π, while the conditional win probability of a against b is the constant r. The probability that two of the n sampled contexts belong to the same cell is at most $\binom{n}{2}/N$. Conditional on visiting distinct cells, averaging over the utility assignments makes the comparison outcomes independent, with probability r that a wins, in both constructions. Consequently, under either construction, the acceptance probability averaged over utility assignments converges to $A_n(r)$.

Suppose, toward a contradiction, that $\mathrm{ValErr}_{u,n}(\mathcal{V}, \widehat{\sigma}) < 1/2$ for every measurable utility function with mean-utility gap of magnitude at least γ. For every such utility with positive gap, the acceptance probability must exceed $1/2$; for every such utility with negative gap, it must be below $1/2$. Applying these inequalities to the two random constructions and letting $N \to \infty$ gives $A_n(r) \geq 1/2$ and $A_n(r) \leq 1/2$. Thus $A_n(r) = 1/2$ throughout the open interval.

Finally, $A_n(r)$ is a polynomial in r of degree at most n. Indeed, after conditioning on the contexts and model pairs, each vector of outcomes indicating whether a wins has probability $r^k(1 - r)^{n-k}$ for some k. Summing over outcome vectors and averaging over contexts, pairs, and the validator's randomness preserves this polynomial form. Since $A_n(r) = 1/2$ on an open interval, it equals $1/2$ everywhere.

In particular, consider the constant utilities $u(x, a) = 1$ and $u(x, b) = 0$. Their mean-utility gap is 1, and

$$\mathrm{ValErr}_{u,n}(\mathcal{V}, \widehat{\sigma}) = 1 - A_n\big((1 + \exp(-\beta))^{-1}\big) = \frac{1}{2},$$

contradicting the assumed strict inequality. Hence a utility function satisfying the proposition exists. □

B.3 Proof of Proposition 3

Proof of Proposition 3. Fix $\beta > 0$ and define

$$A := (1 + e^{-\beta})^{-1}, \qquad B := (1 + e^{\beta/2})^{-1}, \qquad \alpha := \frac{1/2 - B}{A - B}.$$

Since $\beta > 0$, one has $0 < B < 1/2 < A < 1$, so $\alpha \in (0, 1)$. Set $g := (3\alpha - 1)/2$. Writing $t := e^{\beta/2} > 1$ gives $A = t^2/(t^2 + 1)$ and $B = 1/(1 + t)$, and therefore

$$\alpha - \frac{1}{3} = \frac{(t - 1)^3}{6(A - B)(t^2 + 1)(1 + t)} > 0.$$

Hence $\alpha \in (1/3, 1/2)$ and $g > 0$.

Let $\mathcal{D}$ place mass α on a context x_1 and mass $1-\alpha$ on a context x_2. Define utility gaps $\Delta_+(x_1) = 1$, $\Delta_+(x_2) = -1/2$, $\Delta_-(x_1) = -1$, and $\Delta_-(x_2) = 1/2$, and set

$$u^\star_\omega(x, a) = \frac{1 + \Delta_\omega(x)}{2}, \qquad u^\star_\omega(x, b) = \frac{1 - \Delta_\omega(x)}{2}$$

for $\omega \in \{+, -\}$. These utilities take values in $[0, 1]$. Their mean gaps are

$$\mathbb{E}_{x\sim\mathcal{D}}\left[u^\star_+(x, a) - u^\star_+(x, b)\right] = \alpha \cdot 1 + (1-\alpha)\cdot\left(-\tfrac{1}{2}\right) = g$$

and $\mathbb{E}_{x\sim\mathcal{D}}\left[u^\star_-(x, a) - u^\star_-(x, b)\right] = -g$.

Under the Bradley–Terry response model, the conditional win probabilities are $p^+_{ab}(x_1) = A$, $p^+_{ab}(x_2) = B$, $p^-_{ab}(x_1) = 1 - A$, and $p^-_{ab}(x_2) = 1 - B$. By construction of α,

$$\overline{p}^+_{ab} = \alpha A + (1-\alpha)B = \frac{1}{2},$$

and the pointwise identity $p^-_{ab} = 1 - p^+_{ab}$ yields $\overline{p}^-_{ab} = 1/2$. For two models these are all of the pooled win probabilities, so the two populations induce the same pooled distribution. Their mean-utility rankings are nevertheless opposite.

Any ranking rule that depends only on the pooled distribution therefore induces the same distribution over rankings in both populations. If q denotes the probability of outputting $(a \succ b)$, the misranking probabilities are $1 - q$ and q, so at least one of them is at least $1/2$.

Finally, fix a reported ranking $\widehat{\sigma}$. A validator with only pooled access may depend on a validation sample only through the model pairs and outcomes, and not through the contexts. The marginal law of (a, b, y) agrees in the two populations: $(a, b) \sim \Pi$ and $\mathbb{P}(y = 1 \mid a, b) = \overline{p}_{ab} = 1/2$ in both. The acceptance probabilities therefore coincide. Since $\widehat{\sigma}$ is the mean-utility ranking in exactly one of the two populations, the two validation errors sum to one, and at least one is at least $1/2$. The same bound holds for a randomized validator, because randomness shared across the two populations cannot separate identical observation laws. □

B.4 Proofs of the statistical-justifiability results

Proof of Theorem 1. It suffices to consider $\varepsilon < 1$ and $\varepsilon + \gamma \le 1$, since otherwise the verifier that always accepts works. Fix such $\varepsilon, \gamma, \delta$ and take

$$n \ge \frac{36}{\rho\gamma^2} \log \frac{4L}{\delta}.$$

Condition on o; independence of the validation data makes the test family, events, and scores fixed under this conditioning. For $T = (W_i)_{i=1}^n$ and every $j \le L_o$, define

$$N_j := \sum_{i=1}^n \mathbf{1}\{W_i \in A_j\}, \qquad \widehat{\mu}_j := \begin{cases} \dfrac{1}{N_j}\displaystyle\sum_{i=1}^n \mathbf{1}\{W_i \in A_j\} Z_j(W_i), & N_j > 0, \\ 0, & N_j = 0. \end{cases}$$

For every positive-probability event, $N_j \sim \mathrm{Bin}(n, \pi_j)$ with $\pi_j \ge \rho$. A multiplicative Chernoff bound and a union bound give

$$\mathbb{P}\left(\min_{j:\pi_j>0} N_j < \frac{n\rho}{2} \,\middle|\, o\right) \le L e^{-n\rho/8} \le \frac{\delta}{2}.$$

Conditional on $N_j = m$, the selected scores are independent, lie in $[-1, 1]$, and have mean μ_j. Thus Hoeffding's inequality gives

$$\mathbb{P}\left(|\widehat{\mu}_j - \mu_j| > \frac{\gamma}{3} \,\middle|\, N_j = m, o\right) \le 2e^{-m\gamma^2/18}.$$

On the preceding count event, another union bound bounds the probability of any such deviation by $2Le^{-n\rho\gamma^2/36} \le \delta/2$. A zero-probability event has $N_j = 0$ almost surely and $\widehat{\mu}_j = \mu_j = 0$. Consequently, with probability at least $1 - \delta$,

$$\|\widehat{\mu} - \mu\|_\infty \le \frac{\gamma}{3} \quad \text{and hence} \quad |F_o(\widehat{\mu}) - f(o, \mathcal{P})| \le \frac{\gamma}{3}.$$

The validator accepts exactly when $F_o(\widehat{\mu}) \le \varepsilon + \gamma/2$. It therefore accepts if $f \le \varepsilon$ and rejects if $f \ge \varepsilon + \gamma$, on the same event of probability at least $1 - \delta$. □

Proof of **??**. Use one test for each (a, b, g, B), with

$$A_{a,b,g,B} = \{(x, a', b', y) : (a', b') = (a, b),\ x \in C^q_{a,b,g,B}\}, \qquad Z_{a,b,g,B} = y - q_{ab}(x),$$

and $F((v_j)_j) = \max_j |v_j|$. This map is 1-Lipschitz in the sup norm, and there are at most $M(M-1)|\mathcal{G}||\mathcal{B}|$ tests. Theorem 1 applies. □

Proof of Corollary 1. If $\mathcal{Q}_{\widehat{\sigma}}$ is empty, its error is known to equal one. Otherwise, use one test for each (q, a, b, g, B) with $q \in \mathcal{Q}_{\widehat{\sigma}}$, using the events and scores from the preceding proof, and set

$$F((v_{q,a,b,g,B})_{q,a,b,g,B}) = \min_{q \in \mathcal{Q}_{\widehat{\sigma}}} \max_{a \neq b, g, B} |v_{q,a,b,g,B}|.$$

The minimum of these sup-norm errors is 1-Lipschitz, and the test family is known to the validator and has size at most $|\mathcal{Q}|M(M-1)|\mathcal{G}||\mathcal{B}|$. Theorem 1 now applies. □

Proof of **??**. Fix $\varepsilon, \gamma, \delta \in (0, 1)$ with $\varepsilon + \gamma \leq 1$, and define

$$\mathcal{V}_{\mathrm{EC}}(T, o) := \mathbf{1}\{\mathrm{ECerr}_{\mathcal{K}}(\mathcal{A}) \leq \varepsilon\}.$$

Every indicator $r^{\mathcal{A},K}_{ab}$, and hence their maximum absolute difference, is a deterministic function of the rankings contained in o. Thus, if $\mathrm{ECerr}_{\mathcal{K}}(\mathcal{A}) \leq \varepsilon$, the verifier accepts, whereas if $\mathrm{ECerr}_{\mathcal{K}}(\mathcal{A}) \geq \varepsilon + \gamma$, it rejects. These conclusions hold with probability one independently of T, so the sample bound may be taken to be zero. □

Proof of Corollary 2. Use one test for each $(a, b, g, \mathbf{B}, K)$ with $K \in \mathcal{K}_{ab}$:

$$A_{a,b,g,\mathbf{B},K} = \{(x, a', b', y) : (a', b') = (a, b),\ x \in C^{\mathcal{A}}_{a,b,g,\mathbf{B}}\}, \qquad Z_{a,b,g,\mathbf{B},K} = y - q^{\mathcal{A},K}_{ab}(x).$$

Take $F((v_j)_j) = \max_j |v_j|$. The number of tests is at most

$$\sum_{a \neq b} |\mathcal{G}|\,|\mathcal{K}_{ab}|\,|\mathcal{B}|^{|\mathcal{K}_{ab}|},$$

so Theorem 1 applies. □