

Can LLMs Rank? A Tale of Triads and Triage

Gaurab Pokharel¹, Shafkat Farabi¹, Patrick J. Fowler², Sanmay Das¹

¹Virginia Tech

²Washington University in St. Louis

gaurab@vt.edu, mfarabi@vt.edu, pjfowler@wustl.edu, sanmay@vt.edu

Abstract

From housing allocation for households experiencing homelessness to triage in emergency departments, LLMs are increasingly being considered as judges of consequential decisions that require ranking people for scarce resources. Ranking large groups simultaneously is cognitively demanding and error-prone. A natural solution, drawing on decades of social choice theory, elicits pairwise comparisons and aggregates them into a total order. However, a fundamental question remains when LLMs serve as the pairwise judge: how can a practitioner tell, before committing to a ranking, whether the LLM's judgments are sufficiently consistent to trust the result? We discuss two different ways of identifying consistency. A classical diagnostic, the coefficient of consistency ζ , originally developed to measure judge reliability by counting circular triads in tournament graphs, provides a cheap, model-free measure of intra-run consistency. Various standard measures of distance between rankings, for example Kendall's τ , can measure inter-run variability. We show, in both theory and practice, that these measures are independently valuable, and advocate for using both to assess reliability of rankings. We demonstrate the practical importance of our results across two high-stakes prioritization tasks: homelessness service allocation and emergency department triage. Three different leading LLMs have considerably different performance profiles across these two axes of consistency. We provide guidelines for how practitioners could think about measuring and assessing consistency before committing to a model for ranking or prioritization.

1 Introduction

Ranking represents a ubiquitous challenge in high-stakes societal resource allocation settings. Caseworkers must prioritize households for receiving emergency housing, and triage nurses must prioritize patients in the emergency room, for example. These frontline workers – caseworkers, triage nurses, benefits administrators – typically apply professional judgment guided by standardized instruments. In homeless services, communities use different prioritization scores to rank households for housing interventions (for many years, they used the now-disfavored Vulnerability Index - Service Prioritization Decision Assistance Tool (VI-SPDAT),

for example (OrgCode Consulting Inc. and Community Solutions 2015b,a; OrgCode Consulting Inc. et al. 2015)). In emergency departments, the Emergency Severity Index (ESI) (Wolf et al. 2023) structures nurse assessments into a five-level acuity rating that determines who is seen first.

Interest in using large language models (LLMs) to assist with prioritization decisions continues to grow as LLMs demonstrate increasingly sophisticated reasoning capabilities (Wan et al. 2025; Drobotowicz et al. 2026; Williams et al. 2024). Interest aligns with the broader LLM-as-a-judge paradigm, gaining traction for tasks that once required trained human evaluators (Chiang et al. 2024; Liu et al. 2024). Recent work has examined LLM judgments across domains, such as social-service interventions and medical acuity, finding that while models can engage with the task, their assessments are often inconsistent and unreliable (Pokharel et al. 2025; Williams et al. 2024; Haim et al. 2024; Lee et al. 2025). If LLMs are to inform prioritization in such situations, how can a practitioner assess whether the resulting rankings are sufficiently consistent to trust?

To begin answering this crucial question, we first examine how LLMs are actually deployed in ranking tasks. It is well known that LLM outputs are sensitive to presentation order, list length, and prompt framing (Wang et al. 2024; Chiang et al. 2024; Saito et al. 2023), making them fragile for ranking a large number of cases at once. An alternative, grounded in social-choice theory (Condorcet 1785; Kendall and Smith 1940), is to elicit *pairwise* comparisons. This involves asking a model to repeatedly prioritize one out of two cases based on some criteria (e.g., urgency) and then aggregate the responses into a total order. Pairwise elicitation simplifies each judgment, controls context length, and allows principled aggregation methods such as Rank Centrality (Negahban, Oh, and Shah 2017) or Borda Count (Emerson 2013).

Two broad approaches have been used in prior literature to examine the reliability of such rankings. The first reviews intra-run consistency in the rankings, which can be quantified by a measure of transitivity violations (Kendall and Smith 1940). The alternative examines some measure of stability or variance of the ranking outcome, typically when the variance is over the subset of comparisons used in aggregation (Ameli et al. 2025; Zeng et al. 2026). However, to our knowledge, the empirical relationship between the two methods of measuring consistency has not yet been ex-

plored. Moreover, we are unaware of research that examines whether potential consistency or inconsistency in evaluation has implications for the reliability and trustworthiness of AI applications in high-stakes prioritization. These are the gaps that this paper explores.

1.1 Background and Motivation

In principle, one can construct a ranking method that is always internally consistent (no transitivity violations), yet has high variance. Consider a method that randomly permutes the candidates and reports all pairwise comparisons consistent with the chosen permutation. It is also possible that a system with low intra-run consistency nevertheless produces low-variance outcomes. An example of this would be any method that always outputs a fixed ranking (say, lexicographic) on the whole set of choices but encodes an arbitrary number of triad transitivity violations into its pairwise reporting. This makes clear that, at least in theory, intra-run consistency (IC) and inter-run variance (IV) are independent axes, necessitating a 2×2 taxonomy illustrated in Table 1.

		Inter-run variance (IV)	
		Low	High
Intra-run consistency (IC)	High	Borda, consistent triads	Random-permutation
	Low	Fixed-cycle	Noisy method, inconsistent triads

Table 1: Ranking methods by intra-run consistency (IC) and inter-run variance (IV).

The question we ask is where in this taxonomy a combination of an LLM-based (or other) approach and real-world data place us, and what factors drive intra-run consistency and inter-run variance. To measure intra-run consistency, we turn to the coefficient of consistency ζ . Introduced by Kendall and Smith (1940), ζ aims to measure judge reliability by counting circular triads in tournament graphs. Circular triads, i.e., triples $\{a, b, c\}$ for which $a \succ b \succ c \succ a$, are the minimal unit of intransitivity. These correspond to the curl component of the Hodge decomposition of Jiang et al. (2011); ζ is a scalar summary of its magnitude. We use ζ rather than the full decomposition because it distills inconsistency into a singular scalar that a practitioner can compute without specifying a generative model. We measure inter-run variance using Kendall’s τ across runs (Kendall 1938).¹

1.2 Related Work

Pairwise Comparison and Rank Aggregation The problem of aggregating pairwise comparisons into a global ranking has a long history in social choice and statistical decision theory. It dates back at least to Condorcet (1785)’s study

¹Our results are robust to the choice of a specific measure of distance between rankings, as we show in the appendix (Pokharel et al. 2026).

of majority preferences. Classical approaches include positional voting rules, such as Borda count (Emerson 2013) and probabilistic models for pairwise outcomes. In particular, Thurstone’s Case V model (Thurstone 1994) and the Bradley-Terry-Luce model (Bradley and Terry 1952; Luce 1959) provide foundational frameworks in which each item is associated with a latent score and pairwise choices are generated based on score differences. More recent work has developed scalable ranking algorithms for large pairwise-comparison datasets, including spectral methods such as Rank Centrality (Negahban, Oh, and Shah 2017). Rank Centrality (RC) constructs a Markov chain from observed pairwise comparisons and estimates item rankings from its stationary distribution. This makes RC a natural fit for our setting of pairwise prioritization judgments.

Consistency Diagnostics for Tournaments Kendall and Smith (1940) introduced the circular triad count and the coefficient of consistency ζ as a test of judge reliability in paired-comparison experiments; subsequent work developed distributional theory and extensions (David 1988; Alway 1962; Iida 2009). These classical diagnostics operate on tournament graphs and require no generative model. A more recent line of work takes a different approach. Jiang et al. (2011) decomposes the space of pairwise comparison data into three orthogonal components using the Hodge decomposition on graphs: a gradient component (consistent with a global ranking), a curl component (local cyclicity, the source of circular triads), and a harmonic component (global inconsistencies that cannot be resolved locally). The L_2 norm of the curl component quantifies the total cyclicity in the data, and in this sense ζ can be understood as a scalar summary of curl magnitude: a tournament with ζ near 1 has negligible curl, while one with ζ near 0 has maximum. Perotti (2025) study HodgeRank under noise, examining how the decomposition behaves when pairwise data are corrupted. We use ζ because it can be easily computed from the tournament adjacency matrix in closed form without solving an optimization problem and can be used by any practitioner as a summary diagnostic for inconsistency without depending on the specifics of the generative model.

LLM-as-a-Judge – Reliability and Intransitivity As interest grows in using LLMs for pairwise prioritization tasks (Pokharel 2025), a parallel literature on the LLM-as-a-judge paradigm (Chiang et al. 2024) has identified systematic failure modes in LLM pairwise judgments, including sensitivity to presentation order (Wang et al. 2024; Chiang et al. 2024), verbosity bias (Saito et al. 2023), and self-preference (Panickssery, Bowman, and Feng 2024). Beyond these biases, recent work documents a subtler problem: LLM pairwise preferences are frequently intransitive. Xu et al. (2025) find that non-transitivity is more prevalent when the items being compared are close in quality. Zeng et al. (2026) count circular triads directly in passage-ranking tournaments ($n = 100$), Gupta and Kumar (2026) and Qian et al. (2026) report directed 3-cycle rates on text-evaluation tasks, and several other studies measure related constructs using k -subset violation ratios (Wang et al. 2026), strongly connected components (Yu et al. 2026), and intransitivity rates under various

aggregation schemes (Wang, Zhang, and Choi 2025). The consistent finding is that intransitivity is common, varies between models and tasks, and worsens for weaker judges. Crucially, however, all report consistency metrics and ranking accuracy in separate tables. None frames the question in terms of the two independent axes we identify here: intra-run consistency and inter-run variance.

AI in High-Stakes Resource Allocation Many public service allocation problems reduce to ranking under scarcity: caseworkers must decide which households receive limited housing interventions, triage nurses determine which patients are seen first, and child welfare workers choose which families receive intensive services (Cheng et al. 2022; Kube, Das, and Fowler 2023; Persad, Wertheimer, and Emanuel 2009). Standardized instruments, such as the VI-SPDAT (OrgCode Consulting Inc. and Community Solutions 2015b,a; OrgCode Consulting Inc. et al. 2015) in homelessness services and the Emergency Severity Index (Wolf et al. 2023) in emergency departments, exist precisely to impose consistency on these rankings. However, they have well-documented limitations, including racial bias (Cronley 2022; Wilkey et al. 2019) and weak predictive validity (Brown et al. 2018). Policies often mandate prioritizing the most vulnerable (U.S. Department of Housing and Urban Development, Office of Community Planning and Development 2017), while the outcomes on which workers are evaluated favor whoever would benefit most from the intervention (Kube, Das, and Fowler 2023; Persad, Wertheimer, and Emanuel 2009). Certainly, frontline workers exercise professional discretion to navigate this tension (Lipsky 1980; Pokharel, Das, and Fowler 2024). Algorithmic tools are increasingly deployed in these settings to assist or replace human judgment (Angelova, Dobbie, and Yang 2026; Saxena and Guha 2024; Starke et al. 2022), raising questions about whether automated rankings can be trusted when the consequences are borne by vulnerable populations.

Recently, Pokharel et al. (2025) explicitly examined LLM judgments in social service triage. They found that while models can engage meaningfully with prioritization tasks, their pairwise assessments are inconsistent and unreliable. In this paper, rather than asking whether LLMs can replicate expert rankings, we ask a more fundamental question: are LLMs internally consistent (a basic requirement for trustworthiness) in their prioritization rankings? The diagnostic framework we develop is domain-general, but the motivation is grounded in these high-stakes settings where an unreliable ranking could have direct consequences for people’s lives.

Although we study allocation settings where fairness is a central concern, this work is not specifically about group fairness: we do not measure disparities across protected groups, and our diagnostics cannot detect a judge that is biased in a perfectly consistent way. The fairness concern our results speak to is instead procedural. Consistency of treatment is a canonical criterion of procedural justice (Leventhal 1980), and this procedural lens has been extended to algorithmic decision-making (Grgić-Hlača et al. 2018; Lee et al. 2019). An individual whose priority is determined by a judge riddled with circular triads receives a rank that de-

pends on arbitrary features of how those contradictions happen to be resolved. Such arbitrariness becomes morally significant precisely when a single system’s decisions are applied at scale (Creel and Hellman 2022). This harm can persist even when standard group-fairness metrics are satisfied.

1.3 Contributions

We study the relationship between n , the number of candidates; ζ , the Coefficient of Consistency (measuring intra-run consistency); and τ , the rank distance (measuring inter-run variability) in both theory and practice.

1. We theoretically characterize the behavior of ζ as the number of candidates n increases. Of particular interest, we show that if the rate at which pairwise comparisons produce transitivity violations stays constant, ζ converges rapidly to a fixed floor and can be reliably estimated from small samples.
2. In a common stylized model (synthetic tournaments generated from Bradley-Terry-Luce (BTL) models (Bradley and Terry 1952; Luce 1959) with known ground-truth orderings), we empirically show that ζ increases linearly with n and tracks τ in tight lockstep.
3. In the same model, we demonstrate a divergence in the behaviors of τ and ζ when n is held constant, but the sample size of comparisons available to the rank aggregator changes. τ increases with this sample size, while ζ is, by definition, constant in expectation.
4. We show that empirical behavior on real data is markedly different as a function of n . Across a range of datasets that span three different populations experiencing homelessness, as well as patients in emergency rooms who need triage, we show that ζ stabilizes near its asymptotic floor as n increases, while τ continues to increase.
5. On these real-world datasets, we also show the differences between 3 major LLM models (LLaMA, Qwen, and DeepSeek) in terms of their performance on the ζ and τ dimensions. We show that the models differ considerably in terms of which is best on each of the two dimensions – Llama always “winning” on ζ while Qwen wins on τ , while DeepSeek often lands in the middle on both.
6. We discuss the implications of these results for real-world use of LLMs in prioritization and triage. Our results demonstrate clearly that our taxonomy has significant practical implications for measuring consistency and reliability of ranking. Our main recommendation is to analyze models carefully using both measures rather than relying exclusively on one of ζ or τ .

2 Preliminaries

Having introduced intra-run consistency and inter-run variance as independent axes of reliability in Table 1, we now develop the formal machinery needed to measure each one. For intra-run consistency, we need a measure that is computable from tournament data alone and that does not require external validation. The circular triad count and its normalization into ζ provide exactly this. For inter-run variance,

we use a standard measure of distance between rankings – Kendall’s τ (Kendall 1938).

2.1 Tournaments and Circular Triads

Given n items, a pairwise comparison procedure applied to every pair produces a tournament: a complete graph in which every pair of items is connected by exactly one directed edge indicating the preferred item. We represent a tournament by its adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, where $A_{ij} = 1$ if item i is preferred to item j and $A_{ij} = 0$ otherwise, with $A_{ij} + A_{ji} = 1$ for all $i \neq j$. The *out-degree* of item i is $s_i = \sum_{j \neq i} A_{ij}$, the number of items it is preferred to.

Definition 2.1 (Circular Triad). A circular triad in a tournament $\mathbf{A}$ is a triple $\{i, j, k\}$ such that $A_{ij} = A_{jk} = A_{ki} = 1$ or $A_{ik} = A_{kj} = A_{ji} = 1$.

Circular triads are the minimal unit of intransitivity: three pairwise preferences that admit no coherent ordering of the triple. These arise naturally in voting data, in expert judgment, and in any setting where preferences are not perfectly transitive. Their presence indicates that the pairwise preference structure is not fully consistent with any single total order, and the more of them a tournament contains, the less cleanly that tournament can be summarized by a ranking. The total circular triad count in a tournament as defined above admits a closed-form expression (Kendall and Smith 1940; Alway 1962):

$$T_n = \binom{n}{3} - \sum_{i=1}^n \binom{s_i}{2}, \quad (1)$$

with maximum value

$$T_{\max} = \frac{1}{24} \cdot \begin{cases} n(n^2 - 1) & \text{if } n \text{ is odd,} \\ n(n^2 - 4) & \text{if } n \text{ is even.} \end{cases} \quad (2)$$

Definition 2.2 (Coefficient of Consistency). The coefficient of consistency (Kendall and Smith 1940; Alway 1962) is

$$\zeta = 1 - \frac{T_n}{T_{\max}} \in [0, 1]. \quad (3)$$

A value of $\zeta = 1$ indicates a perfectly transitive tournament (zero circular triads), while $\zeta = 0$ indicates maximum cyclicity. The intuition for why ζ is a good measure of ranking consistency is straightforward: any aggregation method must resolve circular preferences, and the more such resolutions it must make, the more the output depends on the particular method of resolution rather than on a shared signal roughly consistent with a total order.

Crucially, ζ can be computed from the tournament data $\mathbf{A}$ alone. This is precisely what a practitioner needs: after collecting pairwise comparisons and running an aggregator, they have a ranking but no way to assess its quality. Computing ζ from the same data that produced the ranking provides a diagnostic that requires no additional comparisons and no “ground truth” (in situations where that exists).

Kendall’s τ as a Measure of Inter-Run Variance To measure the second axis of our taxonomy, inter-run variance, we need a way to quantify how much two rankings of the same n items disagree. We use Kendall’s τ distance (Kendall 1938), which counts the number of pairwise disagreements between two rankings. Given two rankings of the same n items, a pair (i, j) is concordant if both rankings place i and j in the same relative order, and discordant otherwise. The normalized Kendall τ is

$$\tau = \frac{(\text{concordant pairs}) - (\text{discordant pairs})}{\binom{n}{2}},$$

which takes on values in $[-1, 1]$, with $\tau = 1$ indicating identical rankings, $\tau = -1$ indicating reversed rankings, and $\tau = 0$ indicating no more agreement than would be expected by chance. In the context of our work, given two independent runs of the same pairwise comparison and aggregation pipeline on the same set of items, τ measures inter-run agreement. A value near 1 indicates that the pipeline produces stable rankings regardless of which particular run is observed; a value near 0 indicates that the rankings are essentially unrelated across runs.

2.2 Size Dependence of ζ

The definition of ζ in Equation (3) combines two quantities that behave differently as the number of items n changes: the circular triad count T_n , which depends on the comparison data, and the normalization constant $T_{\max}$, which depends only on n . To understand how ζ evolves with n , it is useful to decompose it into a data-dependent term and a purely combinatorial term. This decomposition also has a practical benefit: it yields a natural estimator of ζ from incomplete tournament data, which we will need for our experiments.

Definition 2.3 (Inconsistency Rate). The inconsistency rate of a tournament $\mathbf{A}$ on n items is

$$r = \frac{T_n}{\binom{n}{3}}, \quad (4)$$

the fraction of all triplets that form circular triads.

Rewriting Equation (3) in terms of r gives

$$\zeta = 1 - r \cdot \frac{\binom{n}{3}}{T_{\max}}. \quad (5)$$

This separates ζ into two factors: the inconsistency rate r , which captures how intransitive the pairwise comparisons are, and the scaling factor $\binom{n}{3}/T_{\max}$, which is a known constant depending only on n . The inconsistency rate has a clean interpretation: it is the probability that a uniformly random triplet drawn from the tournament contains a circular triad. Substituting in Equation (2) and simplifying yields the scaling factor in closed form:

$$\frac{\binom{n}{3}}{T_{\max}} = 4 \cdot \begin{cases} \frac{(n-2)}{n+1} & \text{if } n \text{ is odd,} \\ \frac{(n-1)}{n+2} & \text{if } n \text{ is even.} \end{cases} \quad (6)$$

Estimation of ζ in sparse tournaments: The decomposition in Equation (5) has an immediate practical benefit. In many settings, not every pair of items can be compared: either the number of pairs is too large, or comparisons are expensive. When each pair $\{i, j\}$ is observed independently with probability p , a triplet $\{i, j, k\}$ is fully observed only if all three of its constituent pairs happen to be selected, which occurs with probability p^3 . Because the sampling mechanism is independent of the comparison outcomes, the circular triad rate among fully observed triplets is an unbiased estimate of r (and consequently of ζ). Let $\mathcal{O} = \{\{i, j, k\} : \text{all three pairs observed}\}$ denote the set of fully observed triplets, and let T_{obs} be the number of circular triads among them. Then

$$\hat{\zeta} = 1 - \frac{T_{\text{obs}}}{|\mathcal{O}|} \cdot \frac{\binom{n}{3}}{T_{\text{max}}} \quad (7)$$

is an unbiased estimator of ζ . The ratio $T_{\text{obs}}/|\mathcal{O}|$ estimates the inconsistency rate r , and the scaling factor $\binom{n}{3}/T_{\text{max}}$ is the same constant as before. The precision of the estimator depends on $|\mathcal{O}|$, which grows as $\binom{n}{3} \cdot p^3$; even at moderate sparsity the number of fully observed triplets is typically large enough for a reliable estimate.

The decomposition in Equation (5) also makes it clear how ζ depends on the number of items n . The scaling factor $\binom{n}{3}/T_{\text{max}}$ is an increasing function of n that converges to a finite limit, so the behavior of ζ is governed by how the inconsistency rate r changes with n .

Proposition 1 (Asymptotic behavior of the scaling factor). *Both cases of Equation (6) converge monotonically to 4 from below as $n \rightarrow \infty$.*

Proof. The odd case can be written as $4 \cdot (1 - \frac{2}{n})/(1 + \frac{1}{n})$; the n terms vanish as $n \rightarrow \infty$ and the ratio approaches 4. It is strictly less than 4 for all finite n because $n - 2 < n + 1$. The even case is analogous: $4 \cdot (1 - \frac{1}{n})/(1 + \frac{2}{n}) \rightarrow 4$. $\square$

Since the scaling factor increases toward 4, Equation (5) shows that the trajectory of ζ as n grows is entirely determined by the trajectory of r . Three regimes are possible:

Constant r : If the inconsistency rate does not depend on n , then $\zeta(n) = 1 - r \cdot f(n)$ where $f(n) = \binom{n}{3}/T_{\text{max}}$ is increasing in n . In this case ζ decreases monotonically toward the asymptote:

$$\lim_{n \rightarrow \infty} \zeta(n, r) = 1 - 4r. \quad (8)$$

Note that ζ cannot become negative: since $T_n \leq T_{\text{max}}$, the inconsistency rate is bounded by $r \leq 1/f(n)$, which converges to $1/4$, so $r \cdot f(n) \leq 1$ and $\zeta \in [0, 1]$ always. **Increasing r :** if r grows with n , both factors push ζ downward; this would happen if larger comparison sets introduced systematically harder distinctions. **Decreasing r :** If r decreases with n , the two forces oppose. Whether ζ falls, stays flat, or rises depends on whether r decreases slower than, at the same rate as, or faster than $f(n)$ grows.

This three-way characterization frames a central empirical question of the paper. The behavior of ζ as a function of n is not determined by the combinatorics of tournaments alone;

it depends on the inconsistency rate, which is a property of whatever process generates the pairwise comparisons. Different comparison processes (different LLMs, different noise models, different domains) may place the system in different regimes. In Section 3, we instantiate this framework with the Bradley-Terry-Luce model, where r can be derived analytically as a function of the preference strength parameter β . In Section 4, we measure r across several LLMs and datasets and identify each combination’s regime.

3 A Synthetic Model

Whatever the aggregator – Rank Centrality (RC) (Negahban, Oh, and Shah 2017), Bradley-Terry maximum likelihood (BT-MLE) (Hunter 2004), or simple win-rate aggregation – a practitioner ends up with a ranking in hand and no way to assess its reliability. Theoretical guarantees, where they exist, are stated in terms of population-level quantities such as the spectral gap of a transition matrix (Negahban, Oh, and Shah 2017; Shah et al. 2016), which are unavailable from observed data. This section establishes the relationship between ζ and τ under controlled conditions, using a standard generative model with a known ground-truth ranking so that noise can be varied systematically.

3.1 Generative Model

We require a generative model for pairwise comparisons with three properties: (1) a known true ranking against which the aggregator’s output can be measured, (2) a single tunable parameter that controls comparison noise, and (3) rank-gap-dependent difficulty, so that close-in-rank items are harder to distinguish than distant ones. The Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952; Luce 1959) satisfies the first and third requirements by construction: each item $i \in [n]$ has a latent score $w_i > 0$, and the probability that item i defeats item j is $P(i \succ j) = w_i/(w_i + w_j) = \sigma(\theta_i - \theta_j)$, where $\theta_i = \log w_i$ and σ is the logistic function. To satisfy the second requirement, we impose a structure on the score vector that collapses the $n - 1$ free parameters into a single dimension.

Definition 3.1 (Equally-Spaced BTL Model). *Fix n items indexed $1, \dots, n$ in true ranking order. Assign log-scores $\theta_i = \beta(n - i)$ for $i = 1, \dots, n$, where $\beta \geq 0$ is the per-position discriminability. The pairwise comparison probability is*

$$P(i \succ j) = \sigma(\beta(j - i)) = \frac{1}{1 + \exp(-\beta(j - i))}. \quad (9)$$

β governs the entire noise structure. At $\beta = 0$, all comparisons are fair coin flips; as $\beta \rightarrow \infty$, all comparisons become deterministic. The probability that two adjacent items are ranked differently than in the true ranking is $\varepsilon(\beta) = 1/(1 + e^\beta)$, which decreases monotonically in β and represents the hardest-case pair-level error rate. This model has direct precedent in the original RC paper. In Negahban, Oh, and Shah (2017), the authors generate synthetic BTL data with geometric scores $w_i = b^{(2i-1-n)/2n}$, yielding a constant log-score gap of $(\log b)/n$ between consecutive items.

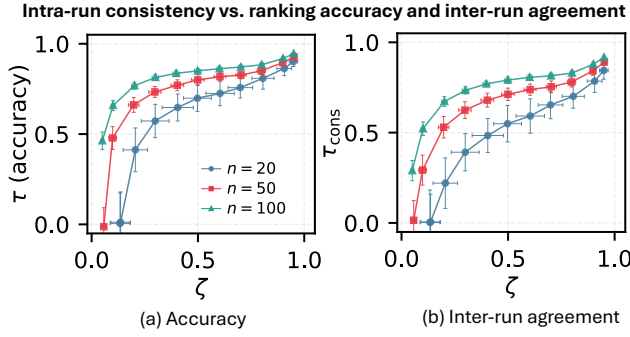


Figure 1: Intra-run consistency (ζ) vs. the two facets of ranking quality under the synthetic BTL model, for $n \in \{20, 50, 100\}$. **(a)** ζ vs. ranking accuracy (Kendall τ against the known ground-truth ordering). The relationship is monotonically increasing and largely converges across tournament sizes at high ζ , with greater separation at low consistency where smaller tournaments provide less information per item. **(b)** ζ vs. inter-run agreement (τ_{cons} , Kendall τ between two independent RC rankings from the same generative model). The curves separate by n : at a given ζ , larger tournaments yield higher inter-run agreement. This separation shows that intra-run consistency alone does not determine inter-run variance; the size of the comparison matrix also matters. Error bars represent ± 1 standard deviation across 100 trials.

This is the same one-parameter family as Definition 3.1 under the identification $\beta = (\log b)/n$. We prefer the β parameterization because it is directly interpretable: two experiments with the same β but different n share the same adjacent error rate $\varepsilon(\beta)$ and local comparison difficulty.

3.2 Experimental Design

The mathematical analysis of Section 2.2 showed that the behavior of ζ with n is governed by the inconsistency rate r , which under the model defined in Definition 3.1 is a function of both β and n . To disentangle the roles of these two parameters, we organize the synthetic experiments along two axes.

Results: Complete Tournaments In the first study, we fix a grid of target ζ values and, for each pair (ζ^*, n) with $n \in \{20, 50, 100\}$, solve for the value of β that produces expected consistency ζ^* at tournament size n . This inverts the natural direction of the model: rather than fixing the noise level and observing how consistent the resulting tournament is, we fix the consistency level and ask what noise level is required to achieve it. For each (ζ^*, n) pair, we generate 100 independent complete tournaments from the BTL model at the corresponding β , aggregate each via Rank Centrality (we show in Figure 6 in Appendix A.3² that these experiments are robust to the specific choice of rank aggregator used), and measure three quantities: the observed ζ of each tournament, the accuracy of the aggregated ranking (Kendall τ against the known ground-truth ordering), and the inter-run

²Full paper on arXiv (Pokharel et al. 2026)

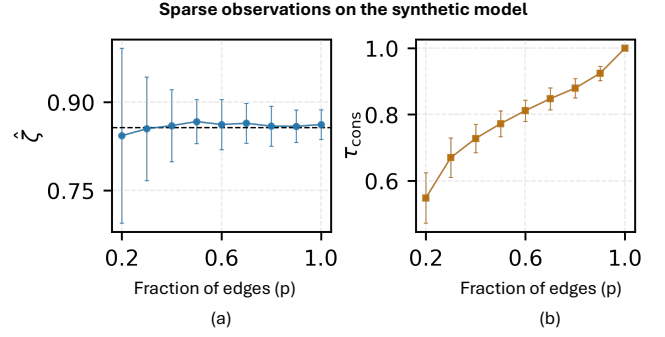


Figure 2: $n = 30$, $\beta = 0.3$, full-tournament $\zeta \approx 0.859$, marked by the dashed line. **(Left)** The sparse estimator $\hat{\zeta}$ is unbiased across the full range of p ; variance increases at low p but the mean remains stable. **(Right)** Inter-run agreement τ_{cons} increases monotonically with the observation fraction, from roughly 0.55 at $p = 0.2$ to perfect agreement at $p = 1.0$. Each point averages 100 independent subsampled tournaments. The contrast illustrates that ζ measures the consistency of the comparison signal while τ_{cons} reflects both consistency and quantity.

agreement (Kendall τ between rankings produced by independent runs of the same generative process, which we denote τ_{cons}).

Figure 1a (left panel) shows that ζ is a strong predictor of ranking accuracy. The relationship is monotonically increasing across all three tournament sizes: higher intra-run consistency corresponds to higher agreement with the ground-truth ordering. With increasing n , at the same level of consistency ζ , we observe a higher value of τ_{acc} . At high ζ (> 0.8), the three curves converge, indicating that when comparisons are sufficiently clean, even a modest number of items yields accurate rankings. Figure 1b (right panel) shows the relationship between ζ and inter-run agreement τ_{cons} . Here the curves separate similarly by n : at any given level of intra-run consistency, larger tournaments produce substantially more stable rankings across independent runs. At $\zeta = 0.5$, for instance, τ_{cons} ranges from roughly 0.55 at $n = 20$ to roughly 0.79 at $n = 100$. The explanation is that a complete tournament on n items contains $\binom{n}{2}$ comparisons, so each item participates in $n - 1$ comparisons. Increasing n increases the amount of data per item available to the aggregator, reducing estimation variance in the ranking even when the level of intransitivity (as captured by ζ) is unchanged.

These results show that, in this synthetic setting, intra-run consistency and inter-run variance are correlated, but they are not redundant: two tournaments with identical ζ can have very different levels of ranking stability depending on the size of the comparison matrix.

Results: Sparse Observation In the second study, we fix $n = 30$ and $\beta = 0.3$ (corresponding to $\zeta \approx 0.859$ in the full tournament) and vary the fraction of pairwise comparisons observed. For each observation probability $p \in \{0.2, 0.3, \dots, 1.0\}$, we subsample edges uniformly at ran-

dom, aggregate only the observed comparisons via RC, and measure both $\hat{\zeta}$ (using the sparse estimator from Equation (7)) and τ_{cons} (between rankings from independent subsampled runs). This isolates the effect of comparison budget on inter-run variance while holding the underlying noise process constant.

Figure 2a confirms that the sparse estimator $\hat{\zeta}$ is unbiased across the full range of observation probabilities. At every value of p from 0.2 to 1.0, the mean of $\hat{\zeta}$ falls on or near the dashed line marking the full-tournament value $\zeta = 0.859$. The variance of the estimator increases at low p , as expected given that fewer fully observed triplets are available, but there is no systematic bias. Figure 2b, on the other hand, tells a different story for inter-run agreement. τ_{cons} increases steadily from roughly 0.55 at $p = 0.2$ to 1.0 at $p = 1.0$. When only 20% of pairs are observed, two independent subsampled runs see largely different comparisons, and their aggregated rankings diverge substantially. As the observation fraction increases, the runs share more data and the rankings converge. Taken together, Figure 2 provides a clean illustration of the distinction between our two axes. The underlying comparison process is identical across all values of p : the same β , the same items, the same pairwise noise. The sparse estimator correctly reports this, returning the same $\hat{\zeta}$ regardless of how many comparisons are observed. But inter-run variance depends not only on the consistency of the comparisons but on the quantity. A practitioner who observes low τ_{cons} but high $\hat{\zeta}$ can diagnose the problem as insufficient comparison budget rather than unreliable judgments, and the remedy is to elicit more comparisons rather than to distrust the comparisons themselves.

4 Evaluating LLMs

Section 3 established that the two axes of our taxonomy are correlated but not redundant even under controlled conditions: ζ captures comparison consistency independently of how many comparisons are collected, while inter-run variation also depends on the comparison budget and aggregation method. The question is whether real-world pairwise comparisons generated by LLMs respect this relationship, or depart from it in informative ways.

Proposition 2. Fix $\beta > 0$. In the equally-spaced BTL model as defined in Definition 3.1, the expected number of circular triads in a complete tournament satisfies $E[T_n] = \Theta(n)$, and $T_{\text{max}} = \Theta(n^3)$. Hence:

$$E[1 - \zeta] = \frac{E[T_n]}{T_{\text{max}}} = \Theta\left(\frac{1}{n^2}\right).$$

Proof sketch. For a triple $i < j < k$ with rank gaps $a = j - i$ and $b = k - j$, the upset probability at gap d is $q_d = \sigma(-\beta d)$, which decays exponentially in d . A circular triad requires either one long-edge upset or two short-edge upsets, so the cycle probability for a given triple is dominated by $e^{-\beta(a+b)}$. Summing over all triples, the contribution from close neighbors (small rank gaps) grow linearly in n , while contributions from distant items are exponentially suppressed. This gives $E[T_n] = \Theta(n)$. Since $T_{\text{max}} = \Theta(n^3)$, the ratio

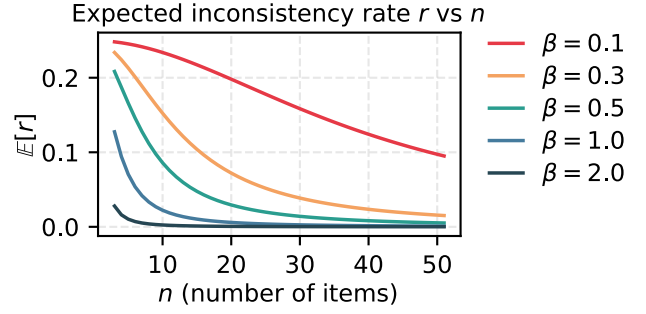


Figure 3: Expected inconsistency rate $E[r]$ as a function of n for several values of β in the equally-spaced BTL model. For every fixed $\beta > 0$, the inconsistency rate decreases with n , confirming Proposition 2.

vanishes as $\Theta(1/n^2)$. The full proof is in Appendix A.2 (Pokharel et al. 2026). $\square$

Figure 3 confirms Proposition 2 visually. For every fixed β , the inconsistency rate $E[r]$ decreases with n and approaches zero. The intuition is that in the equally-spaced BTL model, the vast majority of triples involve items that are far apart in rank, and those comparisons are nearly deterministic: the upset probability $q_d = \sigma(-\beta d)$ decays exponentially with rank gap d . Circular triads then arise almost exclusively from close neighbors, and the number of such close triples grows only linearly in n , while the total number of triples grows cubically. So r shrinks as $\Theta(1/n^2)$, driving ζ toward 1.

In the language of Section 2.2, this places the BTL model firmly in the decreasing- r regime, and the decrease is fast enough that ζ not only remains stable but actually increases. However, we note that the scaling of ζ with n in practice may differ substantially from the stylized model. If LLM pairwise comparisons behaved like this BTL process with some fixed preference strength, we would expect ζ to be very high in the tournaments we study (n in the hundreds). As we will soon see, this is not what we observe. The ζ values in our real-world experiments remain roughly constant across different values of n , suggesting that the inconsistency rate r is approximately constant rather than decreasing. Further, in real prioritization tasks, asking an expensive method like an LLM to perform the full $\binom{n}{2}$ comparisons may be infeasible. We now evaluate the relationship between ζ and the inter-run variability τ_{cons} .

Datasets We evaluate the diagnostic framework on real-world prioritization tasks spanning two domains: homelessness housing allocation and emergency department triage.

Homelessness housing allocation. We use household-level vulnerability assessments from St. Louis coordinated entry records (2021-2024) (Pokharel et al. 2025). Three population-specific instruments are treated as separate datasets: VI-SPDAT for single adults ($n = 325$), VI-F-SPDAT for families ($n = 698$), and TAY-VI-SPDAT for transition-age youth ($n = 561$). Each household has a total acuity score used in practice for prioritization and is assigned to a priority band (low, medium, high) based on es-

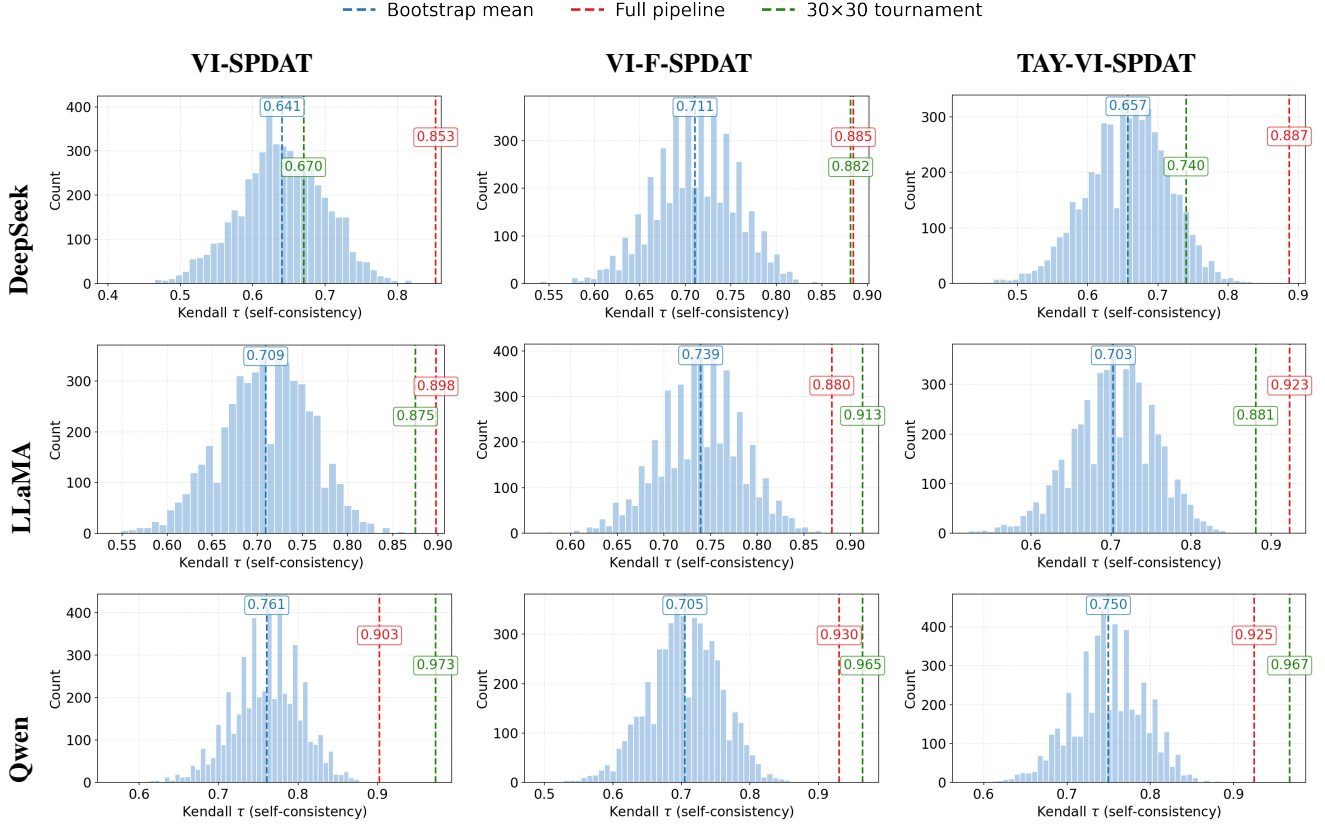


Figure 4: Each panel shows the empirical distribution of τ_{cons} from 100 bootstrap subsamples of a single 30×30 complete subtournament (at $p = 0.4$), with the full-pipeline and complete 30×30 tournament τ_{cons} overlaid (see legend). Rows correspond to LLM judges (DeepSeek, LLaMA, Qwen); columns correspond to homelessness datasets (VI-SPDAT, VI-F-SPDAT, TAY-VI-SPDAT). In every case, the full-pipeline τ_{cons} falls well to the right of the bootstrap distribution, illustrating the distinction between our two reliability axes: the subtournament correctly estimates comparison quality (ζ , see Figure 5), but inter-run agreement (τ_{cons} , this Figure) depends on both the quality and the quantity of comparisons collected.

established scoring rubrics. For the Stage 1 subtournament, we select $m = 30$ households per instrument with proportional representation across priority bands.

Emergency department triage. We use ED visits from MIMIC-IV (Johnson et al. 2024) ($N = 180,057$ total encounters), restricted to encounters with complete triage features and valid acuity scores. Each visit is characterized by vital signs, chief complaint, medication history, and demographics recorded at arrival. The ground-truth label is the Emergency Severity Index (ESI), a five-level acuity score assigned by the triage nurse (Wolf et al. 2023). For Stage 1, we select $m = 30$ patients with proportional representation across ESI levels, matching the distribution in the full dataset. For Stage 2, we sample $n = 500$ again matching proportional representation across ESI levels from the eligible pool. Full inclusion criteria and feature definitions are in Appendix B (Pokharel et al. 2026).

LLMs, Prompting and Implementation Details We use three LLMs for our experiments: *Meta-Llama-3-8B-Instruct*, *DeepSeek-R1-Distill-Llama-8B*, and *Qwen2.5-7B-Instruct*. Due to the sensitive nature of the data, we deploy

these open source models locally with weights downloaded from Hugging-face. The prompts were iteratively designed to try and achieve high consistency, and are provided in Appendix C (Pokharel et al. 2026). We ran these models on one NVIDIA A100 GPU with 80 GB of memory and sampled outputs with 0.1 temperature.

4.1 Experimental Results

Subtournament Evaluation For each dataset, we select $m = 30$ representative items and run all $\binom{30}{2} = 435$ pairwise comparisons per LLM judge, forming a complete subtournament. Item order in each comparison is chosen at random. We compute ζ exactly from the out-degree formula (Equation 1) and measure self-consistency τ_{cons} by running the aggregator on two independent sets of comparisons drawn from the same judge. In addition, we apply the following bootstrap procedure. We repeatedly subsample 40% of the edges of this subtournament. This matches the observation probability $p = 0.4$ in what we call the *full pipeline*, where all nodes are known, 40% of edges are sampled, and a full ranking constructed. For each run, we record ζ , and in

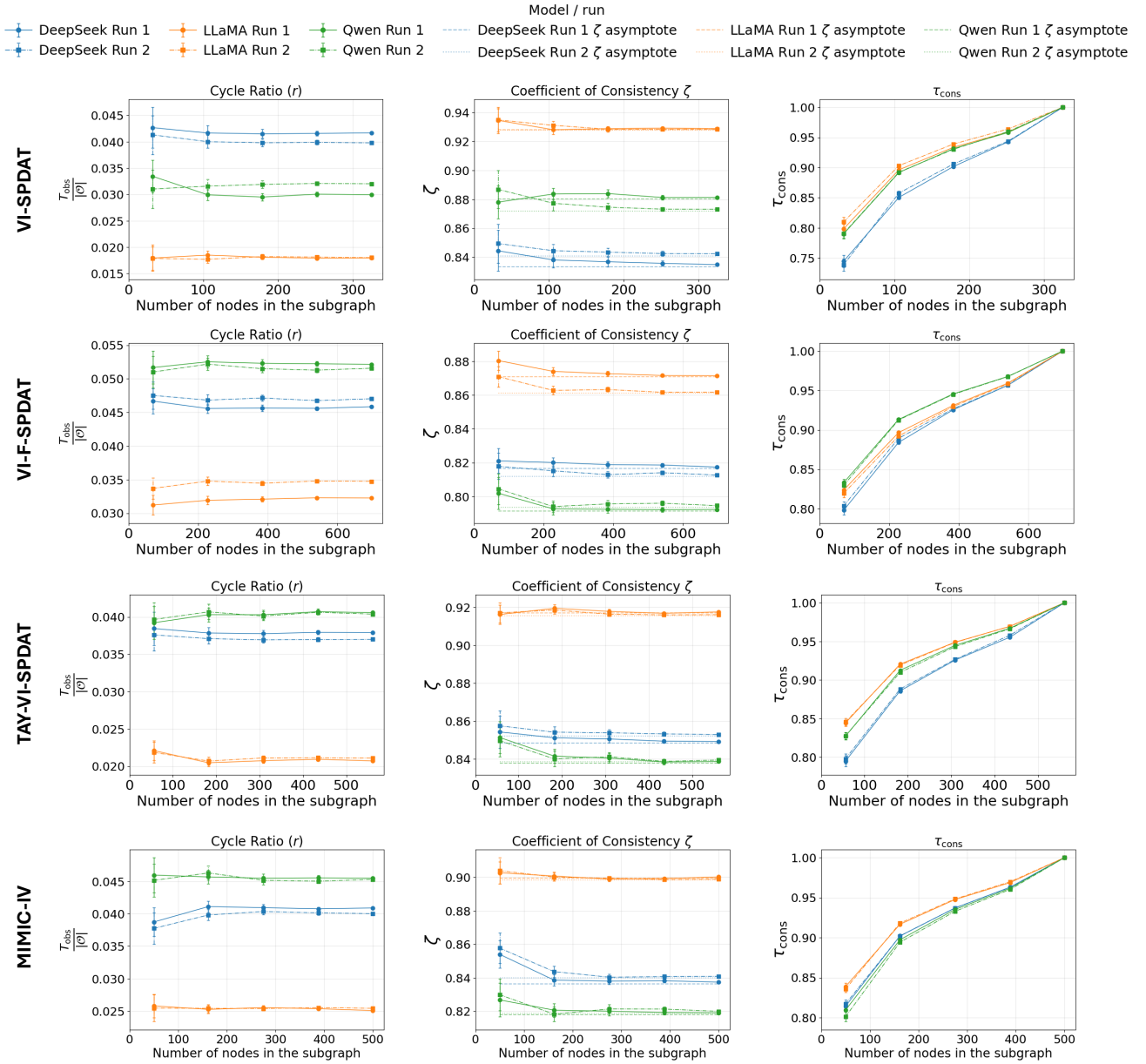


Figure 5: Cycle ratio, ζ , and Rank Centrality (RC) Kendall τ as functions of the number of nodes sampled in the subgraph. Each model uses a consistent color across runs, while separate runs are distinguished by line style and marker. The two left columns show that $r = \frac{T_{obs}}{|\mathcal{O}|}$ and ζ computed from the comparison graphs corresponding to a sub-sample of items remain relatively unchanged, making it a good estimator of LLM comparison consistency even with low comparison samples. The right column shows the same cannot be stated for inter-rank consistency (Kendall’s τ) between generated ranks using rank centrality from same comparison graphs, making it a relatively poor estimator of consistency without sufficiently large sample size.

addition the $\binom{100}{2} = 4,950$ pairwise τ_{cons} values form an empirical distribution of self-consistency.

Figure 4 shows the result of this experiment for inter-run variability. In almost every case, the full-pipeline τ_{cons} , and the τ from the Stage 1 fall well to the right of the bootstrap distribution: the full pipeline is considerably more self-consistent than the subtournament. Meanwhile, the subtournament $\hat{\zeta}$ correctly estimates comparison quality (when scaled appropriately by Equation 5), as in the synthetic

sparse-observation experiment (Figure 2): $\hat{\zeta}$ is unbiased regardless of observation budget. Figure 5 confirms this in the real data: as the number of items grows, ζ behaves as predicted by Equation (8) (the constant r regime) while τ_{cons} increases steadily across all datasets and LLM judges. The full pipeline observes more comparisons per item than any single bootstrap subsample, so its rankings are more stable even though the underlying comparison noise is the same.

Dataset	DeepSeek		LLaMA		Qwen	
	$\bar{\zeta}$	τ_{cons}	$\bar{\zeta}$	τ_{cons}	$\bar{\zeta}$	τ_{cons}
TAY-VI-SPDAT	o	—	+	o	—	+
VI-F-SPDAT	o	o	+	—	—	+
VI-SPDAT	—	—	+	o	o	+
MIMIC-IV	o	o	+	—	—	+

Table 2: Relative model ranking by average ζ and Kendall rank correlation (τ_{cons}). + indicates best, o middle, and — worst among the three models for that dataset and measure.

Full Sparse Pipeline: ζ and τ_{obs} as Functions of n For each dataset, we run the full pipeline on all n items with edge sampling probability $p = 0.4$. This is to keep the number of calls to the LLM tractable. We estimate $\hat{\zeta}$ from the fully-observed triplets in the sparse data (Equation (7)) and measure self-consistency by running the pipeline twice on independent edge samples. To estimate how our two measures scale with n , for each dataset, we take the comparisons performed above, randomly sample fractions of nodes $\in \{0.1, 0.325, 0.55, 0.775\}$, and construct a comparison graph with only the edges that had been sampled. We repeat this 100 times and empirically compute the expected τ, ζ and $r = \frac{\tau_{\text{obs}}}{|\mathcal{O}|}$. The results are provided in Figure 5.

As n increases, the inter-run variability of the ranks produced also decreases. This is because τ_{cons} depends on the number of effective comparisons per item, which increases as we get more nodes with the same sampling ratio of edges. However, the expected inconsistency rate r remains approximately constant across values of n , placing these datasets in the constant- r regime of the decomposition in Equation (5). This is qualitatively different from the BTL model, where Proposition 2 predicts $r = \Theta(1/n^2)$ and $\zeta \rightarrow 1$.

4.2 Consequences

The discussion above demonstrates in practice that ζ and τ_{cons} are different ways of measuring consistency of a ranking method, and they can behave quite differently. Another natural question is whether a method that is good along one dimension is naturally also good along another. Our experiments also allow us to examine this question in the context of the three LLMs that we use. Table 2 shows which of the models performs best/middle/worst along each of these dimensions across the four datasets we look at.³ We find that no model Pareto-dominates across both dimensions. In fact, Qwen is reliably best in terms of inter-run variability, but also reliably worse in terms of intra-run consistency! Meanwhile, LLaMA does best across all the datasets in terms of intra-run consistency, but it is twice worst and twice in the middle in terms of inter-run variability. DeepSeek typically occupies the middle position along both axes. Which dimension is more important? This will of course come down to a question of what a stakeholder values in a particular domain. But our results illustrate clearly the value of evaluating ranking along both these dimensions of consistency.

³For complete results see Appendix B.3 (Pokharel et al. 2026)

5 Discussion

The central finding of this paper is that intra-run consistency and inter-run variance are genuinely independent axes of ranking reliability, not just in theory (Table 1) but in practice. The synthetic calibration shows that even under the idealized BTL model, where a single parameter governs all comparison noise, the two axes are correlated but not redundant: tournaments with identical ζ can have very different τ_{cons} depending on the number of comparisons per item. The real-world experiments sharpen this separation further. The inconsistency rate r remains approximately constant as n grows, placing all four datasets in the constant- r regime of our decomposition and confirming that LLM comparison noise does not have the rank-gap-dependent structure of the BTL model (where Proposition 2 predicts $r = \Theta(1/n^2)$). Most strikingly, no LLM dominates across both dimensions: Qwen produces the most stable rankings across runs but the least internally consistent comparisons, while LLaMA exhibits the opposite profile. This means that choosing a model on the basis of only one diagnostic can be actively misleading. A practitioner who evaluated only inter-run variance would select Qwen; one who evaluated only intra-run consistency would select LLaMA. Neither choice is wrong, but neither is complete.

Our results suggest a simple diagnostic protocol. First, run a small complete subtournament and compute ζ exactly; because ζ is unbiased regardless of observation budget, this inexpensive step reliably estimates comparison quality before committing to the full pipeline. If ζ is low, the comparisons themselves are inconsistent in the sense of transitivity, and no amount of additional data will “fix” that issue. If ζ is high but τ_{cons} is low, the comparisons are consistent but too few have been collected, and the remedy is to increase the sampling fraction rather than change the model. This diagnostic separation between signal quality and signal quantity is the practical payoff of the 2×2 taxonomy.

Ethics Statement

Datasets used in this paper contain sensitive records of real individuals. The MIMIC-IV ED data were accessed under the PhysioNet credentialed data use agreement. The coordinated entry assessment records were used as part of an ongoing community-based research partnership with the St. Louis County, MO, Continuum of Care. In accordance with the data use agreements which prohibit sharing record-level data with third-party services, records were not transmitted to a hosted or commercial LLM API; all LLM judges were open-weight models run entirely on local, access-controlled infrastructure. No record-level data appear in the paper, and all reported results are aggregate statistics.

Acknowledgments

We are grateful for support from NSF Award 2533162. We also thank the various community partners who helped conceptualize the challenges facing the delivery of homeless services, as well as their ongoing efforts to support local families.

References

- Alway, G. G. 1962. The Distribution of the Number of Circular Triads in Paired Comparisons. *Biometrika*, 49(1/2): 265–269.
- Ameli, S.; Zhuang, S.; Stoica, I.; and Mahoney, M. W. 2025. A Statistical Framework for Ranking LLM-based Chatbots. In Yue, Y.; Garg, A.; Peng, N.; Sha, F.; and Yu, R., eds., *International Conference on Learning Representations*, volume 2025, 9505–9540.
- Angelova, V.; Dobbie, W.; and Yang, C. S. 2026. Algorithmic Recommendations and Human Discretion. *Review of Economic Studies*, 93(4): 2250–2283.
- 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.
- Brent, R. 1973. *Algorithms for minimization without derivatives*. Prentice-Hall. ISBN 0-13-022335-2.
- Brin, S.; and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30: 107–117.
- Brown, M.; Cummings, C.; Lyons, J.; Carrión, A.; and Watson, D. P. 2018. Reliability and validity of the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) in real-world implementation. *Journal of Social Distress and the Homeless*, 27(2): 110–117.
- Cheng, H.-F.; Stapleton, L.; Kawakami, A.; Sivaraman, V.; Cheng, Y.; Qing, D.; Perer, A.; Holstein, K.; Wu, Z. S.; and Zhu, H. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.
- 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*, 8359–8388. PMLR.
- Condorcet, M. d., Marie Jean Antoine Nicolas de Caritat. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.
- Creel, K.; and Hellman, D. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52(1): 26–43.
- Cronley, C. 2022. Invisible intersectionality in measuring vulnerability among individuals experiencing homelessness – critically appraising the VI-SPDAT. *Journal of Social Distress and Homelessness*, 31(1): 23–33.
- David, H. A. 1988. *The method of paired comparisons*. Griffin's statistical monographs and courses. New York: Oxford Univ. Press, 2. ed., rev edition. ISBN 9780195206166.
- Drobotowicz, K.; Ylipulli, J.; Varanasi, U. S.; and Mäkitalo, H. S. 2026. Automate, Assist, Avoid: Caseworkers' Perspectives on Applying Large Language Model-Based Assistance in Public Sector Decision-Making Processes. CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.
- Emerson, P. 2013. The original Borda count and partial voting. *Social Choice and Welfare*, 40(2): 353–358.
- Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18, 51–60. AAAI Press. ISBN 978-1-57735-800-8.
- Gupta, M.; and Kumar, D. 2026. Diagnosing LLM Judge Reliability: Conformal Prediction Sets and Transitivity Violations. In *ICML 2026 Workshop on Combining Theory and Benchmarks: Towards A Virtuous Cycle to Understand and Guarantee Foundation Model Performance*.
- Haim, G. B.; Saban, M.; Barash, Y.; Cirulnik, D.; Shaham, A.; Eisenman, B. Z.; Burshtein, L.; Mymon, O.; and Klang, E. 2024. Evaluating Large Language Model-Assisted Emergency Triage: A Comparison of Acuity Assessments by GPT-4 and Medical Experts. *Journal of Clinical Nursing*.
- Hunter, D. R. 2004. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1): 384–406.
- Iida, Y. 2009. The Number of Circular Triads in a Pairwise Comparison Matrix and a Consistency Test in the AHP. *Journal of the Operations Research Society of Japan*, 52(2): 174–185.
- Jiang, X.; Lim, L.-H.; Yao, Y.; and Ye, Y. 2011. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1): 203–244.
- Johnson, A.; Bulgarelli, L.; Pollard, T.; Celi, L. A.; Mark, R.; and Horng, S. 2023. MIMIC-IV-ED. *PhysioNet*. Version 2.2.
- Johnson, A.; Bulgarelli, L.; Pollard, T.; Gow, B.; Moody, B.; Horng, S.; Celi, L. A.; and Mark, R. 2024. MIMIC-IV. *PhysioNet*. Version 3.1.
- Kendall, M. G. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1/2): 81–93.
- Kendall, M. G.; and Smith, B. B. 1940. On the method of paired comparisons. *Biometrika*, 31(3–4): 324–345.
- Kube, A. R.; Das, S.; and Fowler, P. J. 2023. Fair and efficient allocation of scarce resources based on predicted outcomes: implications for homeless service delivery. *Journal of Artificial Intelligence Research*, 76: 1219–1245.
- Lee, J.; Shang, T.; Baik, J. Y.; Duong-Tran, D. A.; Yang, S.; Li, L.; and Shen, L. 2025. Investigating LLMs in Clinical Triage: Promising Capabilities, Persistent Intersectional Biases. *ArXiv*, abs/2504.16273.
- Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26.

- Leventhal, G. S. 1980. *What Should Be Done with Equity Theory?*, 27–55. Boston, MA: Springer US. ISBN 9781461330899.
- Lipsky, M. 1980. *Street-level bureaucracy: Dilemmas of the individual in public services*. New York: Russell Sage Foundation.
- Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; and Collier, N. 2024. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In *First Conference on Language Modeling*.
- Luce, R. D. 1959. *Individual choice behavior*, volume 4. Wiley New York.
- Negahban, S.; Oh, S.; and Shah, D. 2017. Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research*, 65(1): 266–287.
- OrgCode Consulting Inc.; and Community Solutions. 2015a. Family Service Prioritization Decision Assistance Tool (F-SPDAT): U.S. Version 2.0. <https://everyonehome.org/wp-content/uploads/2016/02/F-SPDAT-2.0-Families.pdf>. Accessed May 17, 2025.
- OrgCode Consulting Inc.; and Community Solutions. 2015b. Vulnerability Index–Service Prioritization Decision Assistance Tool (VI-SPDAT): Prescreen Triage Tool for Single Adults. <https://everyonehome.org/wp-content/uploads/2016/02/VI-SPDAT-2.0-Single-Adults.pdf>. Accessed May 17, 2025.
- OrgCode Consulting Inc.; Corporation for Supportive Housing; Community Solutions; and Eric Rice. 2015. Next Step Tool for Homeless Youth (TAY-VI-SPDAT): U.S. Version 1.0. <https://letsendhomelessness.org/wp-content/uploads/2018/07/TAY-VI-SPDAT-v1-0-w-Intro-Script.pdf>. Accessed May 17, 2025.
- Panickssery, A.; Bowman, S.; and Feng, S. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems 37*, 68772–68802. Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 9798331314385.
- Perotti, J. I. 2025. Analysis of the inference of ratings and rankings in complex networks using discrete exterior calculus on higher-order networks. *Physical Review E*, 111(3): 034306.
- Persad, G.; Wertheimer, A.; and Emanuel, E. J. 2009. Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661): 423–431.
- Pokharel, G. 2025. Beyond Automation: Understanding Fairness, Ethics, and Human Discretion in AI-driven Societal Decisions. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2918–2920.
- Pokharel, G.; Das, S.; and Fowler, P. J. 2024. Discretionary trees: understanding street-level bureaucracy via machine learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI’24*, 22303 – 22312. AAAI Press. ISBN 978-1-57735-887-9.
- Pokharel, G.; Farabi, S.; Fowler, P. J.; and Das, S. 2025. Street-Level AI: Are Large Language Models Ready for Real-World Judgments? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2043–2054.
- Pokharel, G.; Farabi, S.; Fowler, P. J.; and Das, S. 2026. Can LLMs Rank? A Tale of Triads and Triage. *arXiv:2606.30412*.
- Qian, M.; Sun, G.; Gales, M.; and Knill, K. 2026. Who can we trust? LLM-as-a-jury for Comparative Assessment. In *Forty-third International Conference on Machine Learning*.
- Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity Bias in Preference Labeling by Large Language Models. *arXiv*.
- Saxena, D.; and Guha, S. 2024. Algorithmic Harms in Child Welfare: Uncertainties in Practice, Organization, and Street-level Decision-making. *ACM Journal on Responsible Computing*, 1(1): 1–32.
- Shah, N.; Balakrishnan, S.; Guntuboyina, A.; and Wainwright, M. 2016. Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 11–20. New York, New York, USA: PMLR.
- Starke, C.; Baleis, J.; Keller, B.; and Marcinkowski, F. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2).
- Thurstone, L. L. 1994. A law of comparative judgment. *Psychological Review*, 101(2): 266–270.
- U.S. Department of Housing and Urban Development, Office of Community Planning and Development. 2017. Notice Establishing Additional Requirements for a Continuum of Care Centralized or Coordinated Assessment System. Technical Report Notice CPD-17-01.
- Wan, F.; Wang, K.; Wang, T.; Qin, H.; Fondrevelle, J.; and Duclos, A. 2025. Enhancing healthcare resource allocation through large language models. *Swarm and Evolutionary Computation*, 94: 101859.
- Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; and Sui, Z. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. Bangkok, Thailand: Association for Computational Linguistics.
- Wang, V.; Zhang, M. J.; and Choi, E. 2025. Improving LLM-as-a-Judge Inference with the Judgment Distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 23173–23199. Suzhou, China: Association for Computational Linguistics.
- Wang, Y.; Song, Y.; Zhu, T.; Zhang, X.; Yu, Z.; Chen, H.; Song, C.; Wang, Q.; Wu, Z.; Dai, X.; Zhang, Y.; Wang, C.; Ye, W.; and Zhang, S. 2026. TrustJudge: Inconsistencies of LLM-as-a-Judge and How to Alleviate Them. In *The Fourteenth International Conference on Learning Representations*.
- Wilkey, C.; Donegan, R.; Yampolskaya, S.; and Cannon, R. 2019. Coordinated Entry Systems: Racial Equity Analysis of Assessment Data. Technical report, C4 Innovations.

- Williams, C. Y.; Zack, T.; Miao, B. Y.; Sushil, M.; Wang, M.; Kornblith, A. E.; and Butte, A. J. 2024. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA network open*, 7(5): e248895.
- Wolf, L.; Ceci, K.; McCallum, D.; and Brecher, D. 2023. *Emergency Severity Index Handbook*. Schaumburg, IL: Emergency Nurses Association, 5th edition.
- Xu, Y.; Ruis, L.; Rocktäschel, T.; and Kirk, R. 2025. Investigating Non-Transitivity in LLM-as-a-Judge. In *Forty-second International Conference on Machine Learning*.
- Yu, Y.; Liu, Y.; He, M.; Tao, S.; Meng, W.; Yang, X.; Zhang, L.; Ma, H.; Li, D.; Wei, D.; Chen, B.; and Li, F. 2026. ELSPR: Evaluator LLM Training Data Self-Purification on Non-Transitive Preferences via Tournament Graph Reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21): 17975–17983.
- Zeng, Y.; Tendolkar, O.; Baartmans, R.; Wu, Q.; Chen, L.; and Wang, H. 2026. LLM-RankFusion: Mitigating Intrinsic Inconsistency in LLM-based Ranking. *Transactions on Machine Learning Research*.