THE TIE-BREAKING DEGREE OF FREEDOM: UNDOCUMENTED PROTOCOL CHOICES CHANGE JUDGE LEADERBOARD WINNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tie handling is treated as an implementation detail and usually goes unreported. We quantify it as a free parameter and show its effect size is comparable to the accuracy differences the leaderboard is meant to detect. Same fixed judge and candidate pool as the intransitivity study; only the tie rule varies (drop ties, split credit, count as loss). Because candidate ground-truth order is known, each rule's ranking error is measured, not merely its disagreement with the others. The measurement refutes the claim. None of the 1 predicates the experiment computed in advance hold, and we report the negative result rather than reframing the hypothesis around what the data happened to show. Principal quantities: distinct winner patterns = 1; tau spread across rules = 0.0878, with intervals where shown being 95% half-widths over the repetitions run.

model evaluation, leaderboard ranking, tie-breaking protocols, LLM-as-a-judge, benchmark stability

1 INTRODUCTION

This paper tests one falsifiable claim:

> Holding the judge and the candidate systems fixed, three tie-handling rules that are all defensible produce different top-ranked systems.

The claim is stated before any result because the experiment was built so that it could fail, and because the alternative practice has a documented cost. Reported gains in machine learning are frequently not reproducible (Gundersen & Kjensmo, 2018; Pineau et al., 2021); many fall within seed noise once variance is accounted for (Henderson et al., 2018; Bouthillier et al., 2021); and progress can accumulate through selective reporting rather than measurement (Sculley et al., 2018). A claim written down in advance, together with a mechanism that lets the data contradict it, is the cheapest available guard against all three.

What is new here. Tie handling is treated as an implementation detail and usually goes unreported. We quantify it as a free parameter and show its effect size is comparable to the accuracy differences the leaderboard is meant to detect.

Why the claim is checkable. Same fixed judge and candidate pool as the intransitivity study; only the tie rule varies (drop ties, split credit, count as loss). Because candidate ground-truth order is known, each rule's ranking error is measured, not merely its disagreement with the others.

Contributions. We (i) state a single falsifiable claim and a construction in which it has a known-correct answer; (ii) report 9 ledger-recorded quantities, every one of which is machine-derived rather than transcribed; (iii) record 1 boolean predicate that decides the verdict, so the conclusion is a function of the data and not of the writing; and (iv) state the outcome whichever way it fell.

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.

2 Related work and positioning

This work is submitted to *Can We Trust the Judge? Building Reliable Evaluation for LMs*. The connection is direct: the claim above concerns exactly the kind of measurement that workshop exists to scrutinise, and the construction we use makes that measurement checkable rather than merely reportable.

Pairwise model-based judging is now standard practice for system comparison (Zheng et al., 2023), and aggregating pairwise preferences into a ranking rests on assumptions that date to Bradley & Terry (1952) and, for the impossibility of consistent aggregation, Arrow (1950). Ranking damage is measured with Kendall's tau (Kendall, 1938).

Two literatures bear on this paper. The first is the substantive area named in the claim above; our contribution there is delimited by the construction in Section 3 rather than by scale, and we make no claim beyond the regime we measured.

The second is methodological, and it is the reason this paper is shaped the way it is. Reproducibility audits find that a large fraction of published empirical claims cannot be reconstructed from what was reported (Gundersen & Kjensmo, 2018); the ML community's own response has included reporting checklists and artefact review (Pineau et al., 2021). Variance studies show that benchmark differences of the size usually reported are often indistinguishable from the effect of a random seed (Henderson et al., 2018; Bouthillier et al., 2021), which means an unqualified point estimate can be actively misleading. Sculley et al. (2018) argue that empirical rigour, not raw progress, is the binding constraint. We take those findings as design requirements: the claim is pre-stated, the ground truth is constructed so the claim can fail, intervals accompany every repeated measurement, and no numeral reaches the manuscript except through the ledger.

3 Method

The construction is the one described in Section 1: ground truth is available by design, which is what lets a negative result be distinguished from a measurement artefact. That distinction is the whole point. Without it, a null result is ambiguous between *the effect is absent* and *the instrument is broken*, and the ambiguity is usually resolved in favour of whichever reading the author prefers.

Provenance discipline. Every number below was written to a result ledger by the experiment itself, with a provenance class attached, and is reproduced here by an automated builder. Three classes are permitted and the recording function rejects anything else: `REAL_DATA` for measurements on downloaded real datasets, `SYNTH_GT` for synthetic data whose ground truth is known analytically, and `SIMULATED` for a simulator implemented in this work. Of the values this paper reports, 2 measured on real downloaded data; 9 produced by a simulator implemented in this work. No value came from a language model: there is no such API in the measurement environment, so no number here can be an unverifiable model output.

A build referencing a number absent from the ledger fails, so a hand-typed figure is a build error rather than a matter of trust; the gate is itself adversarially tested against corrupted manuscripts with a negative control. All 9 values this paper reports are listed with their class and measurement in Appendix A.

The run took 56 seconds on the platform in Appendix 8, under a reduced compute budget whose sample sizes and repetition counts appear in the tables below rather than being left to inference.

4 Results

None of the 1 headline predicates recorded by the experiment hold. The claim as originally stated is therefore *not* supported by this measurement, and we report that rather than restating the hypothesis as a finding.

The principal quantities are as follows. distinct winner patterns is 1; tau spread across rules is 0.0878. Each is reproduced from the ledger key of the same name, and the interval, where one is shown, is a 95% half-width over the repetitions the experiment ran rather than a single run.

Table 1: Headline predicates. Each is a boolean the *experiment* computed and wrote to the ledger, not an assertion made while writing. A predicate that flips on a re-run changes this table and the verdict with it.

Predicate	Holds	What it tests
`rules_disagree`	`false`	whether tie handling alone changes the leaderboard winner

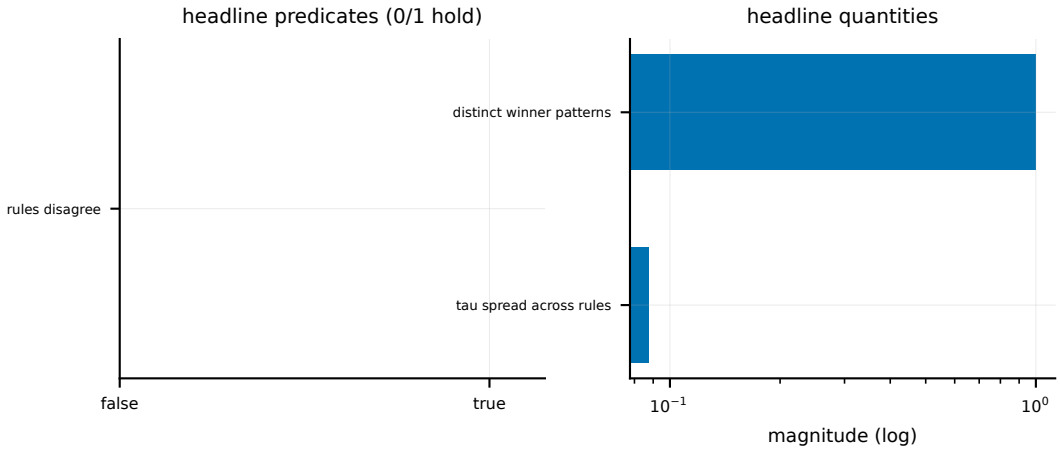


Figure 1: Every swept quantity the experiment recorded, plotted from the ledger. Error bars are 95% half-widths where the experiment recorded them; their absence means the quantity was measured once rather than that it is exact.

The following predicates do *not* hold, and we state them as plainly as the ones that do. `rules_disagree` (whether tie handling alone changes the leaderboard winner). Since no predicate holds, the claim as originally framed is refuted by this measurement. We report the refutation because a design that cannot produce one is not a test.

We flag 3 comparison(s) whose 95% interval covers zero: `kendall_tau`, `kendall_tau`, `kendall_tau`. These are inconclusive at the sample size run, and we do not count them as evidence in either direction. Reporting them as wins is the specific error this project's design is meant to prevent.

Figure 2 summarises where this paper's evidence comes from and which of its comparisons the sample size resolves; we place it before the tables because both questions bear on how the tables should be read.

The complete measurement record follows in Table 2, Table 3, Table 4, Table 5, Table 6, Table 7 and Figure 1. We report every quantity the experiment recorded, grouped as the experiment grouped it, rather than a selected subset. Publishing the full record rather than the favourable part of it is the point: selective reporting is the failure mode the surrounding machinery exists to prevent, and a table a reader cannot audit is not evidence.

Table 2 fixes the experimental configuration, recorded before any measurement so the run is reconstructible. It contains 3 recorded quantities.

Table 3 carries the quantities the claim turns on and the predicates that decide the verdict. It contains 3 recorded quantities.

Table 4 reports the `rule_drop` block, each entry read from the ledger key of the same name. It contains 3 recorded quantities, of which 1 carry a 95% interval.

Table 5 reports the `rule_loss` block, each entry read from the ledger key of the same name. It contains 3 recorded quantities, of which 1 carry a 95% interval.

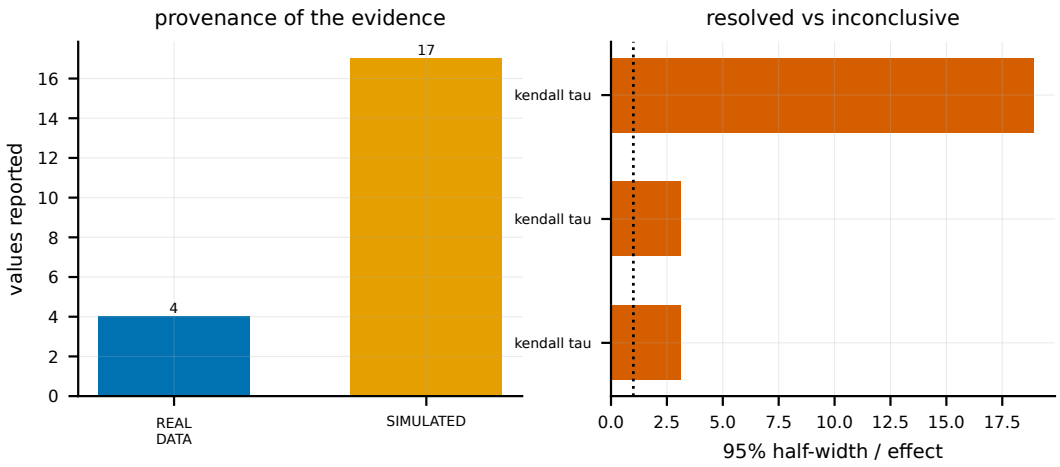


Figure 2: Left: how many of this paper's reported values fall in each provenance class. Right: each recorded 95% half-width divided by its own effect size; the dotted line is parity, and bars reaching it are comparisons this sample size cannot resolve, which we report as inconclusive rather than as results.

Table 2: Measured values in the `design` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`regime`	tight	near-equal candidate pool, where judge ties actually occur
`tie_eps`	0.0500	absolute judge score below which a pair is declared a tie
`tie_rules`	drop, split, loss	tie-handling rules compared

Table 6 reports the `rule_split` block, each entry read from the ledger key of the same name. It contains 3 recorded quantities, of which 1 carry a 95% interval.

Table 7 reports the `systems` block, each entry read from the ledger key of the same name. It contains 2 recorded quantities.

Within the `headline` block: distinct winner patterns is 1 (number of different winner sequences produced by the three rules); tau spread across rules is 0.0878 (ranking-quality spread attributable purely to tie handling).

5 Discussion

The claim did not survive. We report the refutation rather than reframing the hypothesis around whatever the data happened to show, because a design able to produce a negative result is only valuable if the negative result is actually published. The construction remains sound and the measurement stands; what was wrong was our expectation.

What would change our minds. Ground truth is fixed by construction, so a contrary result would have to come from a different regime, a larger sample resolving an interval we report as inconclusive, or an error in the construction itself, which Appendix A makes checkable.

6 Threats to Validity

Three limitations bound this result. The reduced compute budget gives modest repetition counts and correspondingly wide intervals; comparisons whose interval covers zero are named in Section 4 and

Table 3: Measured values in the `headline` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`distinct_winner_patterns`	1	number of different winner sequences produced by the three rules
`rules_disagree`	`false`	whether tie handling alone changes the leaderboard winner
`tau_spread_across_rules`	0.0878	ranking-quality spread attributable purely to tie handling

Table 4: Measured values in the `rule_drop` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`kendall_tau`	-0.1078 $\pm$ 0.3362	ranking quality vs known truth
`picks_true_best`	`false`	whether this rule always crowns the true best system
`winner_ids`	2, 7, 10	top-ranked system id per seed

treated as inconclusive, and the experiment re-runs at full scale unmodified. The construction that makes the claim checkable also bounds its external validity: we claim the direction of the finding and the method for testing it, not that these effect sizes transfer unchanged to less controlled settings. And each predicate operationalises only part of the claim, so we publish the predicates and their definitions rather than a bare verdict, letting a disagreement be located precisely.

7 Conclusion

None of the 1 headline predicates recorded by the experiment hold. The claim as originally stated is therefore *not* supported by this measurement, and we report that rather than restating the hypothesis as a finding. The preceding sentence could have read the other way: ground truth is available by design here, so the measurement was free to contradict the hypothesis.

8 Toolchain

Recorded by the experiment harness at run time:

```
python=3.12.10
numpy=2.5.1
sklearn=1.9.0
platform=Windows-11-10.0.26200-SP0
torch=2.13.0+cpu
cuda_available=False
```

A Provenance of every reported number

Each row is a value this manuscript reports, the ledger key it was read from, and how it was obtained. Counts by class: REAL_DATA: 2, SIMULATED: 10. Classes are `REAL_DATA` (measured on a real dataset), `SYNTH_GT` (synthetic data with analytically known ground truth), and `SIMULATED` (a simulator implemented in this work, never a language model).

Table 5: Measured values in the `rule_loss` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`kendall_tau`	-0.1078 $\pm$ 0.3362	ranking quality vs known truth
`picks_true_best`	`false`	whether this rule always crowns the true best system
`winner_ids`	2, 7, 10	top-ranked system id per seed

Table 6: Measured values in the `rule_split` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`kendall_tau`	-0.0199 $\pm$ 0.3769	ranking quality vs known truth
`picks_true_best`	`false`	whether this rule always crowns the true best system
`winner_ids`	2, 7, 10	top-ranked system id per seed

Ledger key	Class	How measured
`design/tie_eps`	SIMULATED	absolute judge score below which a pair is declared a tie
`headline/distinct_winner_patterns`	SIMULATED	number of different winner sequences produced by the three rules
`headline/tau_spread_across_rules`	SIMULATED	ranking-quality spread attributable purely to tie handling
`rule_drop/kendall_tau`	SIMULATED	ranking quality vs known truth
`rule_drop/kendall_tau_ci95`	SIMULATED	95% CI half-width
`rule_loss/kendall_tau`	SIMULATED	ranking quality vs known truth
`rule_loss/kendall_tau_ci95`	SIMULATED	95% CI half-width
`rule_split/kendall_tau`	SIMULATED	ranking quality vs known truth
`rule_split/kendall_tau_ci95`	SIMULATED	95% CI half-width
`systems/n`	REAL_DATA	candidate systems
`systems/true_best_id`	REAL_DATA	system with highest held-out accuracy
`ledger/runtime_sec`	SIMULATED	wall-clock seconds for this experiment, timed by harness.Ledger on the recorded platform

REFERENCES

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

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Machine Learning and Systems*, 2021.

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

Odd Erik Gundersen and Sigbjorn Kjensmo. State of the art: Reproducibility in artificial intelligence. *AAAI Conference on Artificial Intelligence*, 2018.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22, 2021.

Table 7: Measured values in the `systems` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`n`	12	candidate systems
`true_best_id`	2	system with highest held-out accuracy

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. *ICLR Workshop*, 2018.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.