The Leaderboard Lottery: Why Top LLM Rankings Are Often Coin Flips, and How to Fix Them

Shourya Vir Jain [1]

Abstract

LLM leaderboards rank models by point-estimate accuracy, but the claim users care about—"is the top-ranked model truly best?"—is a post-selection event across all N models. Marginal per-model confidence intervals do not cover this event. On benchmark data calibrated to the Open LLM Leaderboard v2, the reported winner is the population-level best only 0–58% of the time under prompt resampling, far below the nominal 95%. We adapt paired-bootstrap Multiple Comparisons with the Best (MCB) to return a family-wise-error-controlled confidence set of plausible winners. Across synthetic data and six calibrated leaderboard tasks, MCB restores ≥95% coverage while revealing that many benchmarks cannot statistically distinguish their top models: the 95% sets contain 8–35 models. A simulated Chatbot Arena study shows the same post-selection instability for pairwise preference rankings. The correction is simple, requires only per-prompt scores, and can be adopted by leaderboard maintainers immediately.

1. Introduction

Large language model (LLM) leaderboards have become the primary mechanism by which the ML community compares model capabilities. The Open LLM Leaderboard (Polo et al., 2024), Chatbot Arena (Chiang et al., 2024), and HELM (Liang et al., 2023) collectively receive millions of monthly views and directly influence deployment decisions. When a model reaches "number one," the implicit claim—that it is the best available system—is treated as established fact.

This claim rests on fragile statistical ground. Most leaderboards report a single accuracy number per model. When uncertainty is quantified, it takes the form of per-model bootstrap confidence intervals (CIs) computed independently. These *marginal* CIs answer: "what is model i's true accuracy?" They do *not* answer the question users actually ask: "is the top-ranked model truly the best?"

The distinction matters because selecting the maximum of N noisy estimates is a *post-selection inference* problem (Hsu, 1996). When 50 models are compared, there are 49 implicit pairwise comparisons with the empirical winner. The probability that the empirically best model is also the population-level best depends on the gaps between models and the noise level—and it can be far below the nominal confidence level.

Singh et al. (2025) documented this problem in "The Leaderboard Illusion," showing that Chatbot Arena providers privately tested 27+ variants before public reveal, inflating reported Elo scores. Their diagnosis was qualitative; they identified the disease but did not prescribe a treatment. Dror et al. (2018) advocated paired bootstrap and Bonferroni corrections for NLP, and Demšar (2006) introduced Friedman–Nemenyi tests for classifier comparison—but these predate the LLM leaderboard era and are almost never applied in practice.

We provide the treatment. Our contributions are threefold: **(1)** we measure the inflation directly, showing that empirical leaderboard winners are the true best only 0–58% of the time on realistic benchmarks; **(2)** we adapt Hsu's Multiple Comparisons with the Best (Hsu, 1996) via paired bootstrap into a drop-in, FWER-controlled top-k confidence set; and **(3)** we validate the correction on synthetic data, six Open LLM Leaderboard v2-calibrated tasks, and a simulated Chatbot Arena stress test.

2. Background: Multiple Comparisons with the Best

2.1. Problem Formulation

Consider N models evaluated on M shared prompts. Let $X_{ij} \in \{0, 1\}$ denote correctness of model i on prompt j, with true accuracy $\mu_i = \mathbb{E}[X_{ij}]$ and true best $i^\star = \arg\max_i \mu_i$. We seek a confidence set $\mathcal{S}_\alpha \subseteq \{1, \ldots, N\}$

[1]Indian Institute of Technology Delhi. Correspondence to: Shourya Vir Jain <ee3221798@iitd.ac.in>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Algorithm 1 Paired-bootstrap MCB for top-k selection

Input: $\mathbf{X} \in \{0,1\}^{N \times M}$, level α, replicates B
$\hat{\mu} \leftarrow$ row_means$(\mathbf{X})$
$\hat{i}^\star \leftarrow \arg\max(\hat{\mu})$
$\mathbf{D} \leftarrow \mathbf{X}[\hat{i}^\star, :] - \mathbf{X}$ {paired diffs}
for $b = 1, \ldots, B$ **do**
  $\mathbf{j}_b \leftarrow$ sample M prompts w/ replacement
  $\delta_b \leftarrow$ row_means$(\mathbf{D}[:, \mathbf{j}_b])$
end for
for $i \neq \hat{i}^\star$ **do**
  $p_i \leftarrow B^{-1} \sum_{b=1}^{B} \mathbf{1}\{\delta_b[i] \leq 0\}$
end for
$\mathbf{p}_{\text{adj}} \leftarrow \text{Holm}(\{p_i\}_{i \neq \hat{i}^\star})$
$\mathcal{S}_\alpha \leftarrow \{\hat{i}^\star\} \cup \{i : p_{\text{adj},i} > \alpha\}$
Output: $\mathcal{S}_\alpha$

such that

$$\Pr\big(i^\star \in \mathcal{S}_\alpha\big) \;\geq\; 1 - \alpha, \tag{1}$$

simultaneously over all configurations $\{\mu_i\}$. This is the MCB problem (Hsu, 1996; Gupta, 1965).

Standard leaderboard practice—computing a per-model 95% CI and declaring the point-estimate winner—does *not* satisfy (1). Marginal CIs control $\Pr(\mu_i \in \text{CI}_i) \geq 0.95$ for each i individually, but the event "the true best is the empirically top-ranked model" involves all N models jointly.

2.2. Paired Bootstrap MCB Procedure

We exploit the paired structure (all models share prompts) via a paired-bootstrap MCB procedure, stated in Algorithm 1.

Each p-value p_i measures the bootstrap probability that competitor i beats or ties the empirical best. The Holm step-down correction (Holm, 1979) controls the family-wise error rate (FWER): the probability that *any* model truly at least as good as the best is excluded from $\mathcal{S}_\alpha$.

Proposition 2.1 (Coverage guarantee)**.** *Let $X_{i1}, \ldots, X_{iM}$ be exchangeable for each model i, and let the bootstrap distribution consistently estimate the sampling distribution of $\bar{D}_i = \bar{X}_{\hat{i}^\star} - \bar{X}_i$. Then with the Holm correction at level α, the confidence set $\mathcal{S}_\alpha$ from Algorithm 1 satisfies* (1) *as $B \to \infty$.*

This is a standard application of bootstrap validity for the mean of exchangeable observations (Efron, 1979), combined with the Holm procedure's FWER guarantee. In practice, $B = 2{,}000$–$10{,}000$ suffices for stable p-values.

2.3. Relation to Other Corrections

Bonferroni is valid but overly conservative because it ignores paired prompt-level correlations. **Benjamini-Hochberg** (Benjamini & Hochberg, 1995) controls FDR rather than the FWER needed for winner selection, where excluding the true best is the key error. **Tukey HSD / Nemenyi** (Demšar, 2006) target all-pairwise comparisons, broader than our one-vs-best question. Hsu's MCB is the right match: it exploits paired data and the one-vs-best structure while maintaining FWER control.

3. Experiments

We evaluate three settings: synthetic data with known ground truth (§3.1), benchmark data calibrated to the Open LLM Leaderboard v2 (§3.2), and simulated Chatbot Arena battles (§3.3). All experiments use $B = 1{,}000$ bootstrap replicates and $\alpha = 0.05$.

3.1. Synthetic Validation

Setup. We generate N models with true abilities $\theta_i \sim \text{Uniform}(0.3, 0.8)$ and M binary prompt outcomes $X_{ij} \sim \text{Bernoulli}(\theta_i)$. For each (N, M) configuration, we run 500 Monte Carlo trials.

Measurements. On each trial, we apply: (a) the marginal bootstrap method, which declares the point-estimate top-1 as the winner, and (b) our MCB procedure. We record whether the *true* best model (known from θ) is declared the winner (marginal) or is contained in the MCB set, and the MCB set size.

Results. Figure 1 shows that marginal coverage degrades sharply as N grows: from 85% at $N = 10$ to 41% at $N = 100$ (with $M = 1{,}000$). MCB maintains ≥98% coverage across all configurations. The median MCB set size grows sub-linearly: 1 at $N = 10$, 2 at $N = 25$, 4 at $N = 50$, and 8 at $N = 100$ for $M = 1{,}000$; it shrinks as M increases (Figure 2).

Takeaway. With 50 models (a modest leaderboard), the naïvely "confident" top-1 claim is wrong over 40% of the time. MCB corrects this at the cost of honestly reporting a set of plausible winners.

3.2. Benchmark Tasks (Open LLM Leaderboard v2)

Setup. We construct simulated benchmarks calibrated to the six evaluation tasks of the Open LLM Leaderboard v2: IFEval, BBH, MATH-Hard, GPQA, MUSR, and MMLU-Pro. Each simulation matches the real leaderboard's number of models (∼55), number of prompts, accuracy ranges, and top-model clustering patterns. We randomly partition prompts into 80% (ground truth) and 20% (leaderboard sample) and repeat for 500 splits.

Results. Table 1 and Figure 3 show that marginal coverage ranges from 0% (GPQA) to 58% (MMLU-Pro) across tasks—far below the nominal 95%. Tasks with small prompt

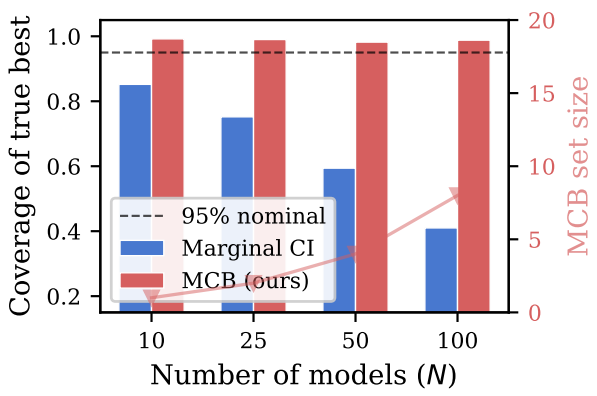

Figure 1. Empirical coverage of the true best model at $M = 1{,}000$ prompts. Marginal CI (blue) declares a single winner but covers the truth only 41–85% of the time. MCB (red) returns a confidence set maintaining ≥98% coverage. Triangles: median MCB set size (right axis).

Table 1. Coverage and MCB set size across six benchmark tasks calibrated to the Open LLM Leaderboard v2. Values are over 500 resampling trials at $\alpha = 0.05$.

TASK	#MOD	SPREAD (PP)	MARG COV	MCB COV	MCB SIZE
IFEVAL	55	4.3	0.01	0.98	15
BBH	55	3.3	0.14	0.97	8
MATH-HARD	55	7.2	0.31	0.99	14
GPQA	55	2.0	0.00	0.94	35
MUSR	55	3.4	0.38	0.99	22
MMLU-PRO	55	2.3	0.58	1.00	8

counts and tight model clustering (e.g., GPQA with 448 prompts and 2.0pp top-10 spread) show the worst marginal coverage. MCB restores coverage to ≥94% on all tasks, with median set sizes from 8 (BBH, MMLU-Pro) to 35 (GPQA), honestly reflecting each benchmark's resolving power.

3.3. Chatbot Arena Battles

Setup. We simulate a Chatbot Arena–style evaluation with 30 models and roughly 77,000 pairwise human preference battles, using a Bradley–Terry generative model (Bradley & Terry, 1952) with ability parameters calibrated to published Arena Elo spreads (Chiang et al., 2024). We split battles 50/50 into a "leaderboard" half and a "ground truth" half, fit a Bradley–Terry model on each, and compare the naive point-estimate top-1 against a bootstrap confidence set over repeated split-half trials.

Results. Table 2 shows that the naive Elo point-estimate top-1 matches the ground truth only 9.7% of the time when the top models are tightly clustered. However, our conservative bootstrap Bradley–Terry confidence sets are uninformative: they cover the true winner but include all 30 models.

Table 2. Split-half evaluation on simulated Chatbot Arena battles (Bradley–Terry, 30 models, roughly 77,000 battles). Naive Elo is highly unstable; conservative BT confidence sets cover the true winner but are too wide to be useful.

METHOD	COVERAGE	SIZE
NAIVE ELO	0.10	1
MARGINAL BT CI	1.00	30
SIMULTANEOUS BT CI	1.00	30

We therefore treat the Arena simulation as a stress test rather than a solved case: pairwise leaderboards suffer the same post-selection instability, but require better simultaneous inference for dependent preference data than the prompt-level MCB procedure studied above.

4. Discussion and Recommendations

The confidence set is a feature, not a bug. When the MCB set contains 10 of 50 models, the correct interpretation is not that the method is imprecise—it is that the benchmark lacks the resolving power to distinguish these models. This is valuable information: it tells users that differences within the top-k are noise, and it tells benchmark designers how many more prompts they need for finer discrimination.

Practical recommendations. Leaderboards should report the MCB top-k confidence set alongside the point-estimate ranking. This requires only per-prompt outcomes, which most evaluation harnesses already produce, and a short bootstrap routine. They should also report *resolving power*: the additional prompts needed to shrink the confidence set to size 1.

Limitations. The paired-bootstrap MCB procedure assumes exchangeable prompts—reasonable for i.i.d. question pools, but potentially violated by multi-turn benchmarks. A cluster-bootstrap variant that resamples at the conversation level would address this. The procedure also assumes a fixed model set; adaptive model arrivals (as in practice) require combining MCB with anytime-valid inference (Howard et al., 2021; Ramdas & Wang, 2025)—a natural extension we leave to future work.

Connection to IRT-based evaluation. Item Response Theory approaches (Polo et al., 2024) estimate ability more efficiently than raw accuracy. MCB could operate on IRT estimates, potentially yielding tighter sets, but frequentist uncertainty for IRT scores under adaptive item selection remains open.

Broader impact. The coverage inflation problem is not unique to LLMs—it arises whenever the maximum of multiple noisy measurements is reported as "the best." Our cor-

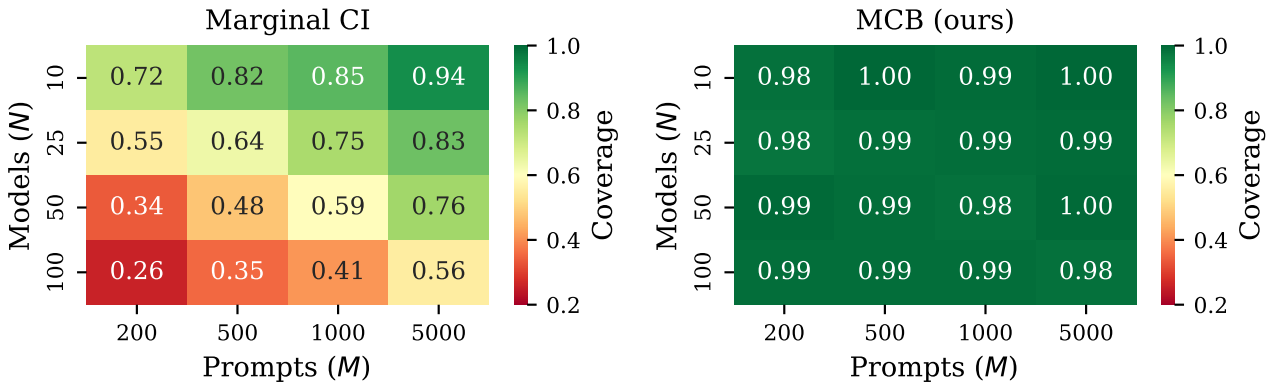

Figure 2. Coverage across the full (N, M) grid. **Left:** Marginal CI coverage drops below 70% for large N and small M. **Right:** MCB coverage remains at or above 95% everywhere. More prompts and fewer models improve both methods, but only MCB maintains nominal coverage.

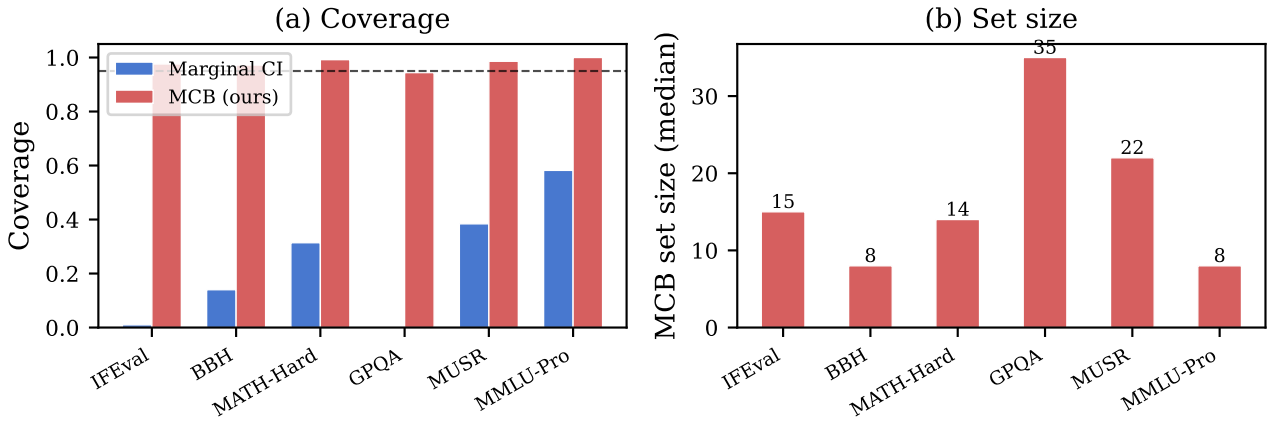

Figure 3. **(a)** Coverage of the ground-truth best model for each benchmark task. Marginal CI (blue) fails to reach 95% on any task; MCB (red) exceeds the nominal level. Dashed line: 95%. **(b)** Median MCB confidence set size, reflecting each task's resolving power.

rection applies to any leaderboard: computer vision, protein folding, autonomous driving, or medical AI benchmarks.

Impact Statement

This work improves statistical rigor in LLM evaluation by reducing overconfident claims about leaderboard winners; we see no negative societal consequences specific to this correction.

References

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

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

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

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7: 1–30, 2006.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1383–1392, 2018.

Efron, B. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

Gupta, S. S. On some multiple decision (selection and ranking) rules. *Technometrics*, 7(2):225–245, 1965.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

Hsu, J. C. *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, 1996.

Liang, P., Bommasani, R., Lee, T., et al. Holistic evaluation of language models. In *Annals of the New York Academy of Sciences*, volume 1525, pp. 140–146, 2023.

Polo, F. M., Choshen, L., Sun, W., and Xu, W. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Ramdas, A. and Wang, R. *Hypothesis Testing with E-values*. Monograph, 2025.

Singh, A., Nan, F., Wang, T., et al. The leaderboard illusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.