LEADERBOARDS DISPLAY RANKS, BENCHMARKS CERTIFY TIERS: SIMULTANEOUS INFERENCE AND A CAPACITY LAW FOR LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Public LLM leaderboards display a total order over models, but the finite item sample behind each leaderboard supports only a partial order. We make the certified objects explicit: simultaneous rank sets, the certified pairwise partial order, and *certified tiers*: ordered groups whose every cross-tier comparison holds jointly with familywise confidence $1 - \alpha$. Auditing five public leaderboards across three data modalities shows that the data support a great deal (57–77% of pairwise orderings certified) while the displayed *total order* is almost entirely unsupported: AlpacaEval 2.0 shows 65 ranks but certifies 2 tiers; LiveBench 94 and 1; a 57k-battle Chatbot Arena release 63 and 2; SWE-bench Verified 134 and 1; the Open LLM Leaderboard's 4,576 ranks reduce to at most 2 tiers per benchmark (5 on its top-100 slice); and top-1 plausible sets range from a unique champion to 24 models (87 under task-clustered inference). A *tier capacity law*, an approximate scaling relation in the sense of empirical scaling laws, predicts how certifiable tiers grow with items as $\exp(-\kappa/\sqrt{n})$ and collapse with model density (slope constants correct to within $\sim 10\%$; the crowding decline reproduced within 35–46% at the smallest subsets), and inverts into item budgets for target resolution. On two refreshing benchmarks, certified tiers did not invert on fresh items in the one refresh where the check is non-vacuous, and rank-flip rates match the closed-form prediction over 130 pooled adjacent pairs. We release `tiercert`, a CPU-only toolkit that turns any score matrix or battle log into rank sets, certified tiers, and item-budget prescriptions; code, data, and result files are provided as supplementary material.

1 INTRODUCTION

Every public LLM leaderboard answers the question *which model is better?* with a total order: m models, m distinct ranks. The data behind the order do not support that answer. On the AlpacaEval 2.0 leaderboard, 65 models occupy 65 displayed ranks, yet the judged instructions behind the board (796 with complete coverage) can simultaneously certify only two ordered groups: the top model, and everyone else. On LiveBench, no displayed distinction among 94 models survives simultaneous inference at all. Leaderboards render precision their benchmarks do not possess, and that precision is consumed as if real: models are chosen by position, progress is claimed by rank, and labs compete (Singh et al., 2025) for slots the data leave unordered.

The community has assembled most of the ingredients for diagnosing this: paired standard errors and pairwise power analysis (Miller, 2024; Kotawala, 2026; Zhuang et al., 2026), rank intervals on individual benchmarks (Neuhof & Benjamini, 2026b;a), measured saturation (Akhtar et al., 2026), and rankings known to flip under perturbations, seeds, and data removal (Alzahrani et al., 2024; Madaan et al., 2024; Huang et al., 2025). Three gaps remain. No inference target matches the object people actually read, the displayed ranking as a whole rather than one pair or one rank at a time. No theory says how much certifiable structure a benchmark of n items and m models *can* support, so neither maintainers nor designers can reason about the resolution they are buying. And the public leaderboards the field consults daily have not been audited.

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.

We replace the displayed total order with the coarsest object the data warrant: a partition into *certified tiers*, ordered groups in which every cross-tier comparison holds simultaneously at familywise level $1-\alpha$, and prove the displayed partition valid as a whole (Proposition 1). We derive and empirically validate a *capacity law* for how the number of certifiable tiers scales with items, models, and score dispersion. And we audit five public leaderboards spanning three data modalities under one framework, validating certification out of sample on refreshing benchmarks. Concretely:

C1: Certified tiers (§4). We define a partition of a displayed leaderboard into ordered tiers such that *every* cross-tier comparison is simultaneously certified, and prove (Proposition 1) that the entire displayed tier structure is correct with probability at least $1-\alpha$. The partition is summarized by the number of certified tiers T and by the Rank Support Ratio $\mathrm{RSR} = T/m$.

C2: A tier capacity law (§5). Under a spacing model for true scores we derive $\mathbb{E}[T] \approx 1 + (m-1)\exp\{-(m-1)\,\Delta^*(n,m,\alpha)/R\}$ with $\Delta^*(n,m,\alpha) = c(m,\alpha)\sqrt{\bar{\delta}/n}$ ($\bar{\delta}$: per-item variance of near-adjacent score differences), yielding three falsifiable predictions: $\sqrt{n}$ item scaling, a *crowding* effect in the number of models, and an item-budget inversion $n^*(T_0)$ (items required for T_0 tiers). All three hold on real leaderboard matrices: the fitted item-scaling constant is within 10% of the closed form on AlpacaEval (linear in $n^{-1/2}$ with $R^2 = 0.96$), and the law reproduces LiveBench's crowding-curve decline (within 35–46% at the smallest subsets).

C3: A cross-platform audit (§6). We audit five public leaderboards spanning three data modalities under one inferential framework: objective accuracy (LiveBench; SWE-bench Verified in an agentic regime; the aggregate-only Open LLM Leaderboard v2), LLM-judge preference (AlpacaEval 2.0), and human battles (Chatbot Arena). Displayed ranks vastly outnumber certified tiers everywhere (65→2, 94→1, 63→2, 134→1, 4,576→ $\leq$5), and natural experiments on two refreshing benchmarks (LiveBench monthly refreshes; AIME exam pairs on MathArena) are consistent with certification out of sample.

C4: Consequences and guidance (§6, §7). Both community defaults fail at leaderboard scale. Uncorrected pairwise testing produces false certifications at familywise rate 0.91–1.00 under the null, and the marginal-CI-overlap heuristic used by leaderboard frontends, roughly per-pair calibrated at m=20 (0.20–0.22), reaches 0.90 at m=100. We also quantify the pairing dividend of shared-item evaluation (1.5–4.5× effective sample size) and convert it into item budgets for maintainers. Finally, *local* certificates over k user-chosen candidates, the modal practical query, fully order 43–64% of random 3-candidate shortlists even where global boards certify almost nothing, though less among capability-matched candidates.

C5: Toolkit. `tiercert` converts a score matrix or battle log into rank sets, certified tiers, capacity curves, and design prescriptions on a laptop CPU.

Rank confidence sets and paired minimum-detectable effects exist in prior work (Mogstad et al., 2024; Neuhof & Benjamini, 2026b; Kotawala, 2026; Miller, 2024); our claims are the tier object with its familywise guarantee, the capacity law with its validation, the unified audit, and the toolkit.

2 Related work

Statistics of LLM evaluation. Miller (2024) frames evals as samples from a question superpopulation and supplies paired and clustered standard errors with pairwise power analysis; Kotawala (2026) inverts paired tests into a minimum-detectable-effect diagnostic on two benchmarks; Zhuang et al. (2026) publishes the same closed form for quantization deltas. Madaan et al. (2024) and Heineman et al. (2025) quantify benchmark noise empirically; Card et al. (2020) did so for pre-LLM NLP. Subset-selection methods (Maia Polo et al., 2024; Vivek et al., 2024; Zhou et al., 2025; Hofmann et al., 2025) compress benchmarks for point estimation; Yauney et al. (2026) shows micro-benchmarks cannot rank close pairs below ∼250 items. None targets the displayed ranking of a deployed leaderboard or provides a leaderboard-level guarantee.

Rank inference. Simultaneous confidence sets for ranks originate in the league-table literature (Goldstein & Spiegelhalter, 1996; Xie et al., 2009; Klein et al., 2020; Mogstad et al., 2024). Neuhof & Benjamini (2026b;a) bring rank intervals to MMLU-style benchmarks (Holm-controlled rejection counting within a task, conformal prediction intervals across tasks), without a partition object or a joint guarantee on the displayed structure; Chatzi et al. (2024) constructs rank-sets under judge-vs-

Table 1: Positioning against the closest prior work.

	inference target	guarantee	data	capacity
Miller (2024)	one model pair	per-pair	—	—
Kotawala (2026)	one pair (MDE)	per-pair	2 benchmarks	—
Neuhof & Benjamini (2026b)	per-model rank CI	simultaneous (per task)	MMLU, TabArena	—
Ameli et al. (2025)	heuristic BT tiers	none for tiers	Arena	—
This work	partition + rank sets + DAG	simultaneous FWER	5 boards, 3 modalities	law + $n^*(T_0)$

human uncertainty, Boyeau et al. (2025) and Divekar (2026) de-bias judge-based evaluation metrics by prediction-powered inference, and Li et al. (2026) obtains rank sets and certified top-K membership under sparse pairwise comparisons via low-rank structure. We adopt this machinery as infrastructure; Table 1 makes the differences precise.

Leaderboards and arenas. Chatbot Arena fits Bradley–Terry models with marginal confidence intervals (Chiang et al., 2024); Ameli et al. (2025) adds tie modeling and covariance structure with heuristic tier groupings; Singh et al. (2025) documents governance distortions; Huang et al. (2025) shows tiny preference subsets flip top ranks. AlpacaEval 2.0 (Dubois et al., 2024) and LiveBench (White et al., 2025) publish the per-instance data our audit consumes but report no separation guarantees; Bayesian treatments (Longjohn et al., 2025) quantify comparison uncertainty differently. None constructs a leaderboard-level object with a joint guarantee. Saturation (Akhtar et al., 2026), protocol sensitivity (Alzahrani et al., 2024; Lunardi et al., 2025), label noise (Gema et al., 2025), judge noise (Feng et al., 2026), and construct validity (Bean et al., 2025) are orthogonal reliability axes we hold fixed: we certify what each leaderboard *claims to measure*.

3 Setup and certification machinery

Score matrices. A benchmark is a set of n items drawn i.i.d. from a question population; m models produce a score matrix $X \in [0,1]^{m\times n}$ with true scores $\mu_a = \mathbb{E}[X_{a,i}]$ and observed scores $\hat{p}_a = \frac{1}{n}\sum_i X_{a,i}$. The leaderboard displays models sorted by $\hat{p}$. (For a fixed item bank scored by stochastic attempts, e.g. SWE-bench or AIME, i indexes an attempt-level record and the certificate covers attempt noise and item composition jointly, §8.) For a model pair (a,b) the paired difference $\hat{D}_{ab} = \hat{p}_a - \hat{p}_b$ has item-level variance $\sigma^2_{ab} = \mathrm{Var}(X_{a,i} - X_{b,i})$; for binary scores $\sigma^2_{ab} = \delta_{ab} - \Delta^2_{ab}$, where $\Delta_{ab} = \mu_a - \mu_b$ and δ_{ab} is the McNemar disagreement rate (McNemar, 1947). When items belong to tasks or subjects we use cluster-robust variants (clusters = tasks) (Miller, 2024).

Pairing dividend. Define $\gamma_{ab} = \sigma^2_{ab}/(v_a + v_b)$ with v the marginal item variances; $\gamma = 1$ corresponds to independent errors and $1/\gamma$ is the effective-sample-size multiplier bought by evaluating all models on *shared* items; empirically $1/\gamma$ is 1.5–4.5× (§6).

Simultaneous certification. We certify the ordered claim $a \succ b$ iff $T_{ab} = \hat{D}_{ab}/\widehat{\mathrm{SE}}_{ab} > c_\alpha$ where c_α is simultaneous over all $M = \binom{m}{2}$ pairs. Each pair carries two directed claims ($2M = m(m-1)$ one-sided hypotheses), so the closed-form Bonferroni value is $c_\alpha = z_{1-\alpha/(2M)}$ (Dunn, 1961), the paired, heteroskedastic counterpart of Tukey's studentized range (Tukey, 1949); the default is the studentized max-t value, the $(1-\alpha)$ quantile of $\max_{(a,b)} |Z^*_{ab}|$ under an item or task-cluster bootstrap (Westfall & Young, 1993; Romano & Wolf, 2005). We keep the single-step rule for its transparent inversion into rank sets; Romano–Wolf step-down adds certified pairs but no tier boundary on any board (App. D). Familywise rather than FDR control is forced by the object: a tier boundary is a *conjunction* of cross-pair claims (§4), and a criterion that tolerates a fraction of false claims cannot certify a conjunction.

Lemma 1 (Simultaneous validity)**.** *Under the super-population model with finite second moments, as $n \to \infty$ (for cluster resampling, as the number of clusters $G \to \infty$),* $\Pr\{\exists\,(a,b) : a \succ b$ *certified and* $\mu_a \le \mu_b\} \le \alpha + o(1)$.

Finite-sample behavior is verified by simulation (Table 2; binary one-parameter-logistic IRT (1PL-IRT) and graded Beta-noise DGPs, 200–2,000 replications; App. B): the max-t rule holds familywise error at or within Monte-Carlo error of target (max 0.053) in every null configuration up to $m{=}100$, never certifies a false boundary, and keeps $\geq$99.0% joint rank-set coverage; the closed-form rule is valid and slightly conservative (0.025–0.040). The two community defaults fail in opposite ways:

Table 2: Validity simulations. "Boundary err": probability that any certified tier boundary is wrong (Proposition 1); coverage: joint rank-set coverage; naive: uncorrected two-sided pairwise z-tests at α; CI: marginal-CI overlap.

Scenario (truth)	FWER	boundary err	coverage	naive FWER	CI FWER
null, m=20, n=500	0.049	0.000	0.995	0.920	0.215
null, m=20, n=2000	0.042	0.000	0.994	0.920	0.200
null, m=100, n=500	0.035	0.000	1.000	1.000	0.890
null, m=100, n=2000	0.040	0.000	1.000	1.000	0.900
graded null, m=20, n=500	0.053	0.000	0.999	0.914	0.008
4 tied groups, m=20, n=1000	0.018	0.000	0.995	0.734	0.082
graded 4 groups, m=20, n=1000	0.020	0.000	0.990	0.755	0.010

uncorrected pairwise testing (two-sided at α per pair) false-certifies at familywise rate 0.91–1.00 under the null and still 0.16 under realistic spread at n=250 (Table 5), while the marginal-CI-overlap heuristic, roughly per-pair calibrated at m=20, degrades catastrophically with scale (familywise error 0.89–0.90 at m=100). From certifications we obtain simultaneous *rank sets* $[r_a^{\mathrm{lo}}, r_a^{\mathrm{hi}}]$ by rejection counting, as in Mogstad et al. (2024); Neuhof & Benjamini (2026b).

Battle data. For pairwise-preference logs (Chatbot Arena) we fit a Bradley–Terry model by maximum likelihood with ties counted as half-wins, compute a sandwich covariance, and certify contrasts with a simulated max-contrast critical value; rank sets and tiers then follow identically.

4 CERTIFIED TIERS

Definition 1 (Certified boundary)**.** *Sort models by displayed score. Position $k \in \{1, \dots, m-1\}$ is a* certified boundary *iff every pair (a, b) with a in the displayed top-k and b in the displayed bottom-$(m-k)$ is certified $a \succ b$.*

Definition 2 (Certified tier partition)**.** *Tiers are the maximal segments between consecutive certified boundaries; T denotes the number of tiers and $\mathrm{RSR} = T/m$ the Rank Support Ratio.*

Proposition 1 (Tier validity)**.** *On the event that no certification is false (probability at least $1-\alpha$ by Lemma 1), every certified boundary is correct: for each certified k, every top-k model's true score strictly exceeds every bottom model's true score. Consequently the displayed tier partition, read as "each tier outperforms every lower tier," holds* jointly *over all boundaries with probability at least $1-\alpha$.*

Proof. Each boundary claim is a conjunction of certified pairwise claims; the union bound over all of them is already paid inside c_α. □

We state this as a proposition, not a theorem: the proof is immediate by construction. The contribution is the *object* and its reporting semantics: a leaderboard-level summary whose displayed structure carries a joint guarantee.

Remarks. (i) The guarantee attaches to the *displayed* object (the partition a reader consumes), not to an abstract parameter. (ii) The partition asserts no order *within* a tier; within-tier pairs may still be certified, and those orderings live in the certified DAG and rank sets below (SWE-bench certifies 74% of its pairs yet has T=1). (iii) A cut checks all $k(m-k)$ cross pairs; with heteroskedastic standard errors the binding pair need not be the adjacent one, so we check the cut exactly. We also report the *adjacent certification rate* (fraction of displayed neighbors $(k, k+1)$ certified), which connects to pairwise audits (Kotawala, 2026), and the size of the top-1 plausible set $\{a : r_a^{\mathrm{lo}} = 1\}$.

Beyond tiers: the certified partial order. The certification relation is acyclic (certified pairs have $\hat{D}_{ab} > 0$, a difference of means), so it is a DAG whose transitive closure is a *partial order*; certification itself need not be transitive ($a \succ b$ and $b \succ c$ do not force $T_{ac} > c_\alpha$). The tier partition is its coarsest display-compatible summary: one uncertifiable pair straddling a cut dissolves that boundary, so T can be small while the DAG is dense, and adding a near-clone of a top model legitimately destroys a boundary. We therefore report *certified-pair density* alongside T: 0.77 on AlpacaEval,

0.65 on the Arena release (simulated max-contrast), 0.57 on LiveBench, 0.71 on MMLU-Pro's top-100, and 0.74 on SWE-bench Verified: most pairwise orderings *are* known even when almost no total-order prefix is. T answers whether the displayed ranking is jointly correct; density, rank sets, and the local certificates of §7 answer which individual comparisons are resolved. Neither subsumes the other: rank sets count certified wins per model, a boundary needs *every* cross pair.

5 THE TIER CAPACITY LAW

How much certified structure *can* a benchmark support? Model the m true scores as a point process with mean spacing $R/(m-1)$ over range R; adjacent spacings are then approximately $\mathrm{Exp}(\text{mean } R/(m-1))$. A boundary certifies when its binding (adjacent) gap survives the certification threshold

$$\Delta^*(n, m, \alpha) \;=\; c(m, \alpha)\,\sqrt{\bar{\delta}/n}, \qquad c(m, \alpha) = z_{1-\alpha/(m(m-1))} \approx \sqrt{2\log\!\left(\tfrac{m^2}{\alpha}\right)}, \tag{1}$$

where $\bar{\delta}$ is the typical near-adjacent per-item difference variance ($\sigma^2 \approx \delta$ in the small-gap regime) and the radical is the leading-order scaling of the directional Bonferroni constant of §3 (all numerics use the exact quantile; the radical overstates it by $\sim$12% at m=65). Hence

$$\mathbb{E}[T] \;\approx\; 1 + (m-1)\,\exp\!\Big\{-\tfrac{(m-1)}{R}\,\Delta^*(n, m, \alpha)\Big\}. \tag{2}$$

To derive it, note that under the spacing model the gap at any cut is approximately exponential with mean $R/(m-1)$. A cut certifies essentially when its binding pair (generically the adjacent one) clears Δ^*, an event of probability $\approx \exp\{-(m-1)\Delta^*/R\}$, and summing over the $m-1$ candidate cuts gives equation 2. Each of these steps is an approximation. Step (ii) is an *upper envelope*: a cut needs all $k(m-k)$ cross pairs to clear, so the adjacent-gap event over-predicts interior boundaries (most at mid-board cuts and moderate m), which is why the law's small-m' interior peak is not seen (Fig. 1); spacings are exchangeable, not independent, and noise smooths the threshold. Equation (2) is therefore a *law* in the sense of empirical scaling laws, an approximate relation with theoretical motivation, not a theorem: we state its assumptions and let its value rest on three falsifiable predictions:

P1 (item scaling). $\log[(\bar{T}-1)/(m-1)]$, with $\bar{T}$ the mean tier count over item draws, is linear in $n^{-1/2}$ with slope $-(m-1)c\sqrt{\bar{\delta}}/R$. We test the *form* by the linearized fit and the *constant* by nonlinear least squares of $\bar{T}$ on equation 2 itself (the log transform explodes as $\bar{T} \to 1$, so its slope is dominated by near-saturated points; App. D reports both). On AlpacaEval, a dense 8×8 subsampling grid at full m is linear in $n^{-1/2}$ with $R^2 = 0.96$ and gives fitted constant -125.7 against the parameter-free closed form -139.6 (ratio 0.90); fixed-subset dense curves at m'=16 give ratios 0.97 (AlpacaEval, form $R^2 = 0.84$) and 1.04 (LiveBench, $R^2 = 0.42$). The law's constant is right to within $\sim$10% and its functional form holds to R^2 of 0.42–0.96 across settings. These fits inherit two caveats from their inputs. Because $(R, \bar{\delta})$ are measured on the same matrices, P1 and P2 test the law's form with measured inputs, and the parameter-free AlpacaEval comparison is the strictest instance. Observed adjacent gaps are also inflated by ranking on noisy scores, a winner's-curse effect strongest at small n and large m, which predicts the mild upward deviations in T at small n and in top-bin flip rates.

P2 (crowding). At fixed n, adding models destroys certified structure: density crushes spacings while multiplicity inflates $c(m, \alpha)$. Because random m'-subsets also shrink the observed range R (confounding density with range contraction), we test P2 with *range-anchored* subsets (every draw contains the global top and bottom model, with the middle models stratified across score quantiles), which holds $R(m')$ at 0.56 throughout. The crowding decline survives: on AlpacaEval, anchored $T(m') = [5.2, 5.0, 3.5, 2.0, 2.0]$ across $m' = [6, 16, 32, 48, 65]$ (random-subset variant and measured-R law curves in App. D). On LiveBench the law reproduces the decreasing branch's shape but runs 35–46% high at the two smallest subsets (observed $[2.6, 1.3, 1.0, 1.0, 1.0]$ vs. predicted $[3.8, 1.8, 1.0, 1.0, 1.0]$), the upper-envelope slack of step (ii). The spacing model is checked directly: adjacent-gap exponential-QQ correlations are 0.92–0.96 on AlpacaEval, LiveBench, SWE-bench, and AIME, but 0.68 on MMLU-Pro, the one platform where the law fails (frontier near-clones; App. D). The law is consistent with the certification machinery: its Bonferroni-form constant sits within 1% of the max-t value actually used (max-t 4.18 vs. Bonferroni 4.22 on AlpacaEval), and

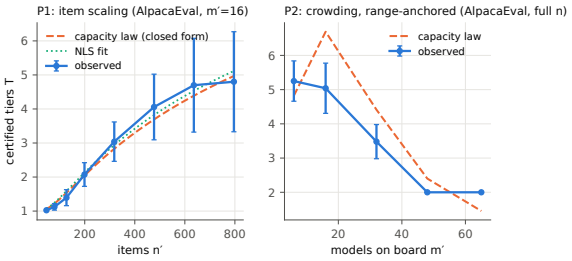


Figure 1: The capacity law on real data. Left: item scaling (P1) on AlpacaEval at m'=16, observed mean $\pm$ sd vs. the parameter-free closed form and the one-parameter NLS fit (constant ratio 0.97; form $R^2 = 0.84$; the $R^2 = 0.96$ form test of §5 is the 8×8 grid at full m). Right: crowding (P2) at full n, range-anchored; the law tracks the observed decline, while its predicted small-m' interior peak is not resolved (App. D).

a zero-parameter check, which measures $(n, m, R, \bar{\delta})$ and predicts $\mathbb{E}[T]$, lands within half a tier on four of five platforms, MMLU-Pro being the under-predicted exception (App. D), exactly where the QQ check says the spacing model is weakest.

P3 (design inversion). Solving equation 2 for n gives the item budget for a target tier count $1 < T_0 < m$, $n^*(T_0) = \bar{\delta}\left[c(m,\alpha)(m-1)/\left(R \log \frac{m-1}{T_0-1}\right)\right]^2$; a full total order ($T_0$=$m$) is unattainable ($n^* \to \infty$) and at fixed T_0 the budget grows as $m^2/\log m$ with the roster (App. A). At current model density, AlpacaEval (n=796) would need $\sim$2.1k items to certify 4 tiers and $\sim$4.0k for 8; LiveBench ($n = 318$ shared items, $m = 94$) would need $\sim$4.2k items to certify even 2 tiers. Moving from 4 to 8 certified tiers costs $\approx$1.9$\times$ the items at these densities.

6 Auditing Five Public Leaderboards

6.1 Data and Modalities

LiveBench (White et al., 2025): per-question ground-truth scores from the official judgment release; the largest complete block is $m = 94$ models on $n = 318$ shared questions (6 tasks, 3 categories).[1] **AlpacaEval 2.0** (Dubois et al., 2024): per-instance weighted GPT-4-Turbo judge preferences vs. the shared baseline ($m = 65$ displayed-board models with complete coverage of $n = 796$ instructions; $m = 221$ for the community board). **Chatbot Arena**: the public 55k human-preference release (57,377 battles, $m = 63$ models with $\geq$200 battles), Bradley–Terry with sandwich covariance. **SWE-bench Verified**: per-instance resolution records for all 134 official-leaderboard submissions (2023 retrieval baselines through late-2025 agents) on the 500 verified instances, a currently-active board in an agentic task regime. **Open LLM Leaderboard v2** (OLL): 4,576 models' aggregate scores on six benchmarks; per-item files are gated, so the full-board audit brackets $\sigma^2_{ab} = \gamma\,(v_a + v_b)$ with γ calibrated where we hold per-item data, and we recompute the top-100 slice *exactly* from the gated per-item details (100 models $\times$ 6 benchmarks). We audit the most recent release with public per-item data in each case: SWE-bench (submissions through December 2025) and LiveBench (2025-era models) are current; AlpacaEval, the Arena release, and the archived OLL are 2024-era. The protocol applies unchanged to any future release.

6.2 Main Results

Table 3 is the audit headline. Every leaderboard displays far more ranks than its data certify; sub-leaderboards (LiveBench categories, $n = 50$–140) certify no tier boundary. Across Table 3's boards the Rank Support Ratio T/m runs from 2×10^{-4} (OLL full boards) to 0.05 (MMLU-Pro top-100): the data support at most a twentieth of the displayed resolution. Two features of the audit matter beyond the headline counts. First, *task heterogeneity dominates LiveBench*: under task-clustered

[1] The snapshot lists two models twice under case-variant names (distinct runs); merging them (m=92) leaves the iid audit unchanged and moves the task-clustered top-1 set from 87/94 to 62/92 (App. D).

Table 3: Cross-platform audit at $\alpha = 0.05$ (simultaneous max-t wherever per-item data are held, App. C; the task-clustered row resamples only G=6 clusters, $c_\alpha \approx 18$, a conservative bound). Displayed ranks vastly outnumber certified tiers.

Leaderboard	ranks m	items/battles	tiers T	adj. cert.	top-1 set
AlpacaEval 2.0 (displayed)	65	796	2	1.6%	1
AlpacaEval 2.0 (community)	221	787	1	0%	5
LiveBench (overall, iid)	94	318	1	0%	6
LiveBench (task-clustered)	94	318	1	0%	87
Chatbot Arena (full public release)	63	57,377	2	1.6%	2
SWE-bench Verified (all submissions)	134	500	1	0%	19
Open LLM Lb. v2 (per benchmark, γ=1)	4,442–4,576	541–12,032	1–2	≈0%	—
top-100 slice, exact per-item pairing:					
IFEval / BBH / GPQA / MuSR	100	541–5,761	1	0%	4–24
MATH Lvl 5	100	1,324	2	1.0%	20
MMLU-Pro	100	12,032	**5**	4.0%	**1**

inference 87 of 94 models are plausibly #1. Part of that width is mechanical (six resampled clusters inflate c_α from 4.3 to ≈18, App. D), but the direction is substantive: model ordering there is largely a claim about task composition, echoing subject-level heterogeneity on MMLU (Neuhof & Benjamini, 2026a). Second, certification is not vacuous where data suffice: MMLU-Pro ($n = 12{,}032$; median $\gamma = 0.22$, a $4.5\times$ pairing dividend) certifies 5 tiers with a uniquely certified top model, the regime the capacity law says buys structure; subject-clustered inference (14 clusters, conservative as on LiveBench) compresses this to T=2 with a four-model top tier (App. D).

Reading individual boards. The certified structure that does survive is informative. On the Arena release, the certified tier 1 is {gpt-4-1106-preview, gpt-4-0125-preview}, two GPT-4-Turbo variants the data cannot order, with rank sets widening fast below (the displayed #5 is only locatable to ranks 3–22); this is sharper than the platform's own marginal-CI rank ranges because paired contrasts and simultaneous control are built in. On AlpacaEval, tier 1 is a single model yet rank sets stay informative near the top (displayed #2 certified to ranks 2–3). On GPQA's top-100 slice, 24 models are plausibly #1 and no boundary certifies: at n=1,192 items GPQA cannot separate frontier models. SWE-bench Verified separates certified knowledge from certified tiers most sharply: 74% of its 8,911 pairwise orderings are certified (49% repository-clustered, App. D; 2025 agents provably beat 2023 baselines), yet no tier boundary survives and 19 submissions are plausibly #1, because the frontier is dense even though the historical spread is huge. And LiveBench's category subboards (n=50–140) certify no tier boundary: coding and language retain 38% and 41% of pairwise certificates (top-1 sets of 11 and 13), instruction following (n=50) certifies not a single pair, yet the site renders all three with the same visual authority as the overall board.

The pairing dividend on real data. Across every per-item matrix we hold, model errors are strongly positively correlated on shared items: median γ is 0.66 on AlpacaEval, 0.54 on LiveBench, and, per OLL benchmark on the top-100 slice, 0.51 (IFEval), 0.45 (BBH), 0.58 (GPQA), 0.37 (MuSR), 0.57 (MATH), and 0.22 (MMLU-Pro): effective sample-size multipliers of 1.5–4.5× relative to evaluating models on disjoint item draws. The γ brackets used for the gated full board are thereby validated in direction: the exact top-100 recomputation agrees with the bracket on five of six benchmarks, and on MMLU-Pro the bracket under-predicts ($T = 5$ vs. bracketed ≤ 3), erring conservatively as intended (App. D). Pairing is also a design lever: a leaderboard that evaluates each model on its own item sample (private reruns, rotating subsets) forfeits this multiplier and needs $1/\gamma$ times more items for the same resolution.

6.3 Out-of-sample validation: natural experiments on refreshing benchmarks

LiveBench across months. LiveBench refreshes questions monthly, so distinct months are independent item samples from (approximately) the same measurement target. Using the one task fielded in both the 2024-06 and 2024-11 releases (`connections`, language category; $m = 38$ models, $n_A = n_B = 50$), certification transfers exactly as Proposition 1 predicts: **zero** of 37 cross-certified-tier pairs invert on the fresh month (a low-power check: at n_B=50 even correct certifications could flip by chance, so we read zero inversions as consistency, not confirmation), while

Table 4: Distinguishable tiers under three error criteria ($\alpha = 0.05$; Definition 1 applied to each criterion's certification matrix: two-sided per-pair z-tests; directional BH on two-sided p-values (Benjamini & Yekutieli, 2005); the max-t/max-contrast rule).

Platform	ranks m	uncorrected	BH–FDR	FWER
AlpacaEval 2.0 (displayed)	65	7	6	2
LiveBench (overall)	94	2	2	1
Chatbot Arena (full release)	63	4	4	2
MMLU-Pro (top-100, per-item)	100	9	9	5

adjacent pairs (36 of 37 uncertified) flip at rate 0.43 against a closed-form prediction of 0.37 (within binomial error over N_{pairs}=37; no-task-filter variant 0.44 vs. 0.40). The prediction is the plug-in $\Phi(-|\hat{D}^{(A)}_{ab}|\sqrt{n_B}/\hat{\sigma}^{(A)}_{ab})$; since adjacent pairs are selected *by* batch-A order, $|\hat{D}^{(A)}|$ is winner's-curse inflated and the plug-in leans low.

AIME exam pairs on MathArena. AIME exams I and II are distinct 15-problem papers set to one difficulty template a week apart and evaluated on an identical roster: an item refresh with the roster held fixed. On AIME 2025 ($m = 66$, 4 attempts per problem), exam I→II adjacent-rank flips occur at rate 0.354 against a predicted 0.396 (65 pairs; the smaller 2026 edition, $m = 29$: 0.250 vs. 0.400, within 1.7 binomial SEs of its 28 pairs). The capacity law also passes a third-platform check: at $n = 15$ it predicts $\mathbb{E}[T] \approx 1$, and $T = 1$ is what certification returns; a 15-problem exam supports *no* simultaneous ordering claim over 29–66 models.

Certified structure survives resampling while displayed adjacent ranks churn at the predicted rate (pooled calibration over 130 adjacent pairs, Fig. 3; bins are small, so this shows consistency rather than sharp calibration). The LiveBench overlap is a single task and the AIME tier test is vacuous by design ($T = 1$), so these are proof-of-concept validations; the design generalizes to any refreshing benchmark.

6.4 What current practice certifies instead

Replacing simultaneous certification with common practice is not a small error (Table 2; visual comparison in App. B): uncorrected pairwise testing falsely certifies at familywise rate 0.92, and the marginal-CI-overlap heuristic of leaderboard frontends at 0.200–0.215, versus 0.042–0.049 for max-t (target 0.05). On tied groups, naive testing reports 4.0 tiers on average but places a false boundary inside a tied group in 3.5% of runs, where max-t recovers 3.6 of the true 4 without ever crossing a true boundary (boundary FWER 0.000 everywhere). Selection effects compound the problem: the displayed top score is a maximum over noisy estimates and inherits winner's-curse inflation whenever the top is contested. Split-half checks make this concrete: on AlpacaEval, where the top model leads by six points, rank 1 is stable (split-half top-1 flip probability 0.00, mean inflation 0.003); on LiveBench the split-half top-1 is likewise stable (0.00), yet six models are plausibly #1 under iid inference and 87 under task-clustered inference. Stability of the point estimate and resolvability of the top are different properties; top-1 plausible sets (Table 3) are the summary to consult before quoting "the best model." Heuristic tier renderings fare no better: a 1-D strength-clustering baseline (in the spirit of the tier displays of Ameli et al., 2025) reports >1 tier in *100%* of null replications on the real Arena battle structure (it cannot report one tier: the silhouette is undefined at k=1, App. D), marginal-CI chaining in 31%, vs. 0% for certification; on the real release the cluster heuristic happens to agree (T=2) while CI chaining reports 9 tiers.

The choice of error criterion is itself part of what a leaderboard should disclose (cf. replicability analysis in NLP; Dror et al., 2017). Our headline does not depend on it, however (Table 4). Relaxing familywise control to Benjamini–Hochberg FDR, or abandoning correction entirely, certifies more structure with weaker guarantees, yet displayed ranks still exceed distinguishable tiers by an order of magnitude under *every* criterion: even uncorrected per-pair testing collapses AlpacaEval's 65 ranks to 7 tiers and LiveBench's 94 to 2. Only the familywise column carries Proposition 1's joint guarantee on the displayed partition.

7 GUIDANCE AND THE TIERCERT TOOLKIT

For **leaderboard maintainers**: (i) render rank sets and certified tiers, not bare ranks (the machinery is $O(m^2 n)$ on a laptop); (ii) report T and RSR alongside the board; (iii) treat onboarding as a statistical decision: by P2, every added model erodes certified resolution at fixed n. For **benchmark designers**: size items to claims via $n^*(T_0)$, with $(R, \bar{\delta})$ from a pilot on a few anchor models or a precursor benchmark of the same format (the law needs a shared-item matrix; a capacity law for sparse battle graphs is open); LiveBench-scale boards need thousands, not hundreds, of shared items before even two tiers certify; maximize the pairing dividend by evaluating all models on identical items (1.5–$4.5\times$ effective sample); and if cluster-robust inference is intended, field task clusters in the tens, not six. For **consumers**: the practical question is rarely the full ranking but "which of my k candidates is best", where multiplicity shrinks from $\binom{m}{2}$ to $\binom{k}{2}$. This *local certification* is far more optimistic: where the global boards certify 1–5 tiers, a random triple of models is *fully ordered* 64% of the time on AlpacaEval, 56% on MMLU-Pro, and 43% on LiveBench (k=5: 17%/12%/2%). Capability-matched shortlists resolve less, since frontier gaps are compressed: triples from the displayed top quartile are fully ordered 61%/28%/4% of the time, and from the top ten 52%/62%/0% (MMLU-Pro's boundaries sit inside its top ten). The $\binom{k}{2}$ multiplicity is valid for shortlists formed independently of the scores (cost, latency, availability); a shortlist selected *by* score is a post-selection comparison, for which the global-α certificate is the right one. `tiercert` implements all analyses (accuracy matrices, judge preferences, battle logs; max-t and Bonferroni; capacity fits and n^*) on CPU, with `audit_matrix(X, alpha=0.05)` and `n_star_for_tiers(T0, m, R, delta_bar)` as entry points.

8 LIMITATIONS

We certify each leaderboard's ranking *with respect to its own measurement*. Judge bias, contamination, label noise, and construct validity are orthogonal reliability axes (Feng et al., 2026; Gema et al., 2025; Bean et al., 2025), so a tier certified on AlpacaEval is a claim about the official judge's preferences rather than about quality. Guarantees are asymptotic (in items, or in clusters for cluster resampling, where $G \leq 14$ makes the clustered critical values conservative), with simulation-verified finite-sample behavior. The capacity law is an approximation, not a theorem about any board: its constant is accurate to $\sim 10\%$ where the item curve is resolved, while the linearized fit and the sparsest crowding subsets run up to $\sim 40\%$ off (App. D). The full Open LLM Leaderboard audit brackets variances where per-item data are gated, and those brackets proved conservative on the exact top-100 slice. For fixed-item boards (SWE-bench, AIME) the certificates cover attempt noise and item composition jointly, not generalization beyond the item bank. For Arena, non-i.i.d. routing and Bradley–Terry misspecification are only partially absorbed by the sandwich covariance, and public per-item availability selects for older, smaller snapshots, so live-scale data would certify more structure; our numbers audit the public record. FWER over all pairs is a deliberately strict standard, but the conclusion is criterion-robust (Table 4). The non-vacuous tier-transfer test rests on one LiveBench task (T=1 on AIME by construction); battle ties are halved, not modeled (Ameli et al., 2025). Every certificate is a snapshot claim about the item set and roster at audit time; a refreshing board needs the standing protocol of §6.3.

9 CONCLUSION

We asked how much of a displayed leaderboard ranking the underlying data support. Across five public boards and three modalities the answer is a small fraction: 65 displayed ranks reduce to 2 certified tiers, 94 to 1, and 134 to 1. What fails is the total order rather than the measurement, since 57–77% of the pairwise orderings on those boards are certified. The capacity law accounts for the gap and inverts into the item budget a target resolution requires.

For maintainers this makes reliability measurable and improvable. Rendering the tier partition and rank sets beside the ranking is one function call on data already collected. Reporting γ and $\bar{\delta}$ lets a reader compute the resolution floor Δ^*, while $n^*(T_0)$ sizes a new item pool before launch, and a refreshing board can publish its cross-version flip rate against the previously certified structure. Natural extensions are anytime-valid certification and composition with judge-reliability corrections (Feng et al., 2026; Chatzi et al., 2024).

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: the synthetic scores used in the validity study (§3 and App. B) are drawn by the released code from the statistical models specified there and were not generated by a language model, and in particular we did not use such tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes publicly released evaluation data of AI systems; it involves no human subjects, no personal data, and no new data collection. Model names appearing in audits are public leaderboard identifiers. Findings concern the statistical reliability of evaluation infrastructure and are intended to improve reporting practice; we see no foreseeable ethical risks beyond the general considerations of benchmark governance discussed in the paper.

REPRODUCIBILITY STATEMENT

All experiments run on CPU with zero API cost. Data are public: the AlpacaEval repository, LiveBench, Chatbot Arena, and MathArena releases on HuggingFace, the SWE-bench `experiments` repository, and Open LLM Leaderboard `contents` plus gated per-item details (available after accepting the dataset terms). The supplementary archive contains the complete analysis library, per-experiment scripts, one-command data acquisition and orchestration (`download_data_minimal.sh`, `run_all_minimal.sh`), the result JSONs behind every table and figure, and fixed seeds throughout; App. E details the mapping from scripts to results.

REFERENCES

Mubashara Akhtar, Anka Reuel, Prajna Soni, et al. When AI benchmarks plateau: A systematic study of benchmark saturation. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.16763.

Norah Alzahrani et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of ACL*, 2024. arXiv:2402.01781.

Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. A statistical framework for ranking LLM-based chatbots. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.18407.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets & Benchmarks Track*, 2025. arXiv:2511.04703.

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

Pierre Boyeau, Anastasios N. Angelopoulos, Nir Yosef, Jitendra Malik, and Michael I. Jordan. Autoeval done right: Using synthetic data for model evaluation. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2403.07008.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of EMNLP*, 2020. arXiv:2010.06595.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Prediction-powered ranking of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.17826.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.04132.

Abhishek Divekar. Statistically reliable LLM-based ranking evaluation via prediction-powered inference. *arXiv preprint arXiv:2606.05308*, 2026.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 2017.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.20913.

Aryo Pradipta Gema et al. Are we done with MMLU? In *Proceedings of NAACL*, 2025. arXiv:2406.04127.

Harvey Goldstein and David J. Spiegelhalter. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A*, 159(3):385–409, 1996.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. *arXiv preprint arXiv:2508.13144*, 2025.

Valentin Hofmann, David Heineman, Ian Magnusson, et al. Fluid language model benchmarking. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2509.11106.

Jenny Y. Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. Dropping just a handful of preferences can change top large language model rankings. *arXiv preprint arXiv:2508.11847*, 2025.

Martin Klein, Tommy Wright, and Jerzy Wieczorek. A joint confidence region for an overall ranking of populations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):589–606, 2020.

Anany Kotawala. Resolution diagnostics for paired LLM evaluation. *arXiv preprint arXiv:2605.30315*, 2026.

Jiachun Li, David Simchi-Levi, and Will Wei Sun. Low rank for rank: Uncertainty-aware task-specific LLM ranking under sparse pairwise comparisons. *arXiv preprint arXiv:2605.29395*, 2026.

Rachel Longjohn, Shang Wu, Saatvik Kher, Catarina Belém, and Padhraic Smyth. Bayesian evaluation of large language model behavior. *arXiv preprint arXiv:2511.10661*, 2025.

Riccardo Lunardi, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. On robustness and reliability of benchmark-based evaluation of LLMs. In *European Conference on Artificial Intelligence (ECAI)*, 2025. arXiv:2509.04013.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.14992.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Magne Mogstad, Joseph P. Romano, Azeem M. Shaikh, and Daniel Wilhelm. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies*, 91(1):476–518, 2024.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. *arXiv preprint arXiv:2607.16259*, 2026a.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. *arXiv preprint arXiv:2606.08679*, 2026b.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

John W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of EACL*, 2024. arXiv:2309.08638.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, 1993.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. Spotlight. arXiv:2406.19314.

Minge Xie, Kesar Singh, and Cun-Hui Zhang. Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association*, 104(486):775–788, 2009.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026. Oral. arXiv:2510.08730.

Jin Peng Zhou, Christian K. Belardi, Ruihan Wu, Travis Zhang, Carla P. Gomes, Wen Sun, and Kilian Q. Weinberger. On speeding up language model evaluation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.06172.

Zexin Zhuang, Yanhang Li, and Zhichao Fan. Pre-registering the detectable effect: A paired-mde budget for 4-bit quantization benchmarks, with a pilot audit. *arXiv preprint arXiv:2605.28873*, 2026.

A PROOFS AND DERIVATIONS

Proof of Lemma 1. Write $Z_{ab} = \sqrt{n}(\hat{D}_{ab} - \Delta_{ab})/\hat{\sigma}_{ab}$ for the studentized paired statistics. Under the super-population model with i.i.d. items and finite second moments, the multivariate CLT gives $(Z_{ab})_{(a,b)} \Rightarrow \mathcal{N}(0, \Sigma)$ jointly over the $M = \binom{m}{2}$ pairs, where Σ is the correlation structure induced by shared items, and $\hat{\sigma}_{ab} \to \sigma_{ab}$ a.s. A false certification of $a \succ b$ requires $Z_{ab} > c_\alpha - \sqrt{n}\,\Delta_{ab}/\hat{\sigma}_{ab} \geq c_\alpha$ whenever $\Delta_{ab} \leq 0$; under an exact tie both directions are false, so the relevant event for an unordered pair is $|Z_{ab}| > c_\alpha$. Hence $\Pr\{\text{any false certification}\} \leq \Pr\{\max_{(a,b)} |Z_{ab}| > c_\alpha\}$. For the Bonferroni constant $c_\alpha = z_{1-\alpha/(2M)}$ the union bound over the M unordered pairs gives $M \cdot 2(1-\Phi(c_\alpha)) = \alpha$ up to the Gaussian approximation error (with $z_{1-\alpha/M}$ the same bound is only 2α). For the max-t constant, the nonparametric bootstrap over items consistently estimates the law of $\max_{(a,b)} |Z_{ab}|$ (Westfall & Young, 1993; the studentized max statistic is a Hadamard-differentiable functional of the empirical process), so its $(1-\alpha)$ quantile converges to the population quantile and the bound holds with $\alpha + o(1)$. Cluster resampling replaces items by task blocks and the same argument applies with clusters as the resampling unit, with the asymptotics in the number of clusters G; at small G (six on LiveBench, twelve on SWE-bench) the resampled covariance is noisy and the resulting critical values are conservative (App. D). □

Proof of Proposition 1. Let E be the event that no certification is false; $\Pr(E) \geq 1 - \alpha$ by Lemma 1. Fix any certified boundary k: by Definition 1 every pair (a, b) with a in the displayed top-k and b below is certified, so on E every such pair satisfies $\mu_a > \mu_b$. This holds simultaneously for all certified boundaries on the same event E, with no further multiplicity cost. □

Derivation of the capacity law. Model true scores as m points with ordered spacings $g_1, \ldots, g_{m-1}$ over range R; under the uniform-order-statistics (Poisson) approximation each spacing is $\mathrm{Exp}(\text{mean } R/(m-1))$ marginally. Treat boundary k as certified when its binding cross-pair (generically the adjacent pair, whose gap is g_k) clears the threshold $\Delta^*(n, m, \alpha) = c(m, \alpha)\sqrt{\bar{\delta}/n}$:

$$\Pr\{\text{boundary } k \text{ certified}\} \;\approx\; \Pr\{g_k > \Delta^*\} \;=\; \exp\{-(m-1)\Delta^*/R\}.$$

Linearity of expectation over the $m-1$ candidate boundaries gives $\mathbb{E}[T] \approx 1 + (m-1)\exp\{-(m-1)\Delta^*/R\}$, i.e. equation (2). The three approximations (binding-adjacent pair; marginal exponentiality with exchangeable-not-independent spacings; a sharp threshold in place of the smooth power curve) are the slack observed empirically. The first is one-directional: Definition 1 requires all $k(m-k)$ cross pairs to clear, and the joint survival of that many noisy statistics is below the adjacent pair's alone, so $\Pr\{g_k > \Delta^*\}$ is an upper envelope on the certification probability and equation (2) over-predicts interior boundaries, most at mid-board cuts where $k(m-k)$ is largest. This is the mechanism behind the unresolved small-m' interior peak of Fig. 1 and the over-prediction on the densest LiveBench subsets.

Design inversion. Setting $\mathbb{E}[T] = T_0$ in equation (2) and solving for n: $(m-1)\Delta^*/R = \log\frac{m-1}{T_0-1}$, hence $\sqrt{\bar{\delta}/n}\,c(m,\alpha) = \frac{R}{m-1}\log\frac{m-1}{T_0-1}$ and $n^*(T_0) = \bar{\delta}\left[\frac{c(m,\alpha)(m-1)}{R\log\frac{m-1}{T_0-1}}\right]^2$, defined for $1 < T_0 < m$: at $T_0 = m$ the logarithm vanishes ($n^* \to \infty$; a total order is unattainable under continuous spacings) and for $T_0 > m$ the expression is meaningless, so `tiercert` rejects both rather than returning a finite budget. For fixed T_0 and growing m, $c(m,\alpha) \sim \sqrt{4\log m}$ and the logarithm grows like $\log m$, so $n^*(T_0) = O(m^2/\log m)$: holding resolution fixed while onboarding models costs near-quadratically many items.

B VALIDITY SIMULATIONS IN FULL

Data-generating processes. Binary scenarios are 1PL-IRT: item difficulties $d_i \sim \mathcal{N}(0,1)$ i.i.d., abilities θ_a, and $X_{a,i} \sim \mathrm{Bernoulli}(\sigma(\theta_a - d_i))$ with σ the logistic function; true scores $\mu_a = \mathbb{E}_d[\sigma(\theta_a - d)]$ are computed by Monte Carlo (2×10^5 draws). Ability configurations: *null* $\theta_a \equiv 0$; *spread* θ_a equally spaced on $[-1, 1]$; *4 groups* five models at each of $\theta \in \{-0.9, -0.3, 0.3, 0.9\}$. m=20 throughout except the two m=100 null rows. Graded scenarios keep the same θ and d and draw $X_{a,i} \sim \mathrm{Beta}(\kappa\mu_{a,i}, \kappa(1-\mu_{a,i}))$ with $\mu_{a,i} = \sigma(\theta_a - d_i)$ and concentration $\kappa = 4$

(heavy per-item noise on $[0, 1]$, as in judge-score data). All closed-form rules use the directional accounting of §3: uncorrected means $|T_{ab}| > z_{1-\alpha/2}$ per pair, CI-overlap uses marginal $(1-\alpha)$ intervals, Bonferroni uses $z_{1-\alpha/(m(m-1))}$.

Table 5 reports every scenario of the validity study: binary 1PL-IRT at 2,000 replications (max-t bootstrap B=2,000) and the extended scenarios (leaderboard-scale m=100; graded Beta-noise scores) at 200–1,000 replications, B=400. Cells are FWER / boundary-error / rank-coverage. The graded-null row reports the 1,000-replication re-run; an initial 200-replication pass (also in the released JSONs) gave max-t 0.065.

Table 5: Full validity results. Targets: FWER ≤ 0.05, boundary error ≤ 0.05, coverage ≥ 0.95.

Scenario	max-t	Bonferroni	naive	CI-overlap
null m=20, n=500	0.049/0.000/0.995	0.040/0.000/0.995	0.920/0.001/0.421	0.215/0.000/0.962
null m=20, n=2000	0.042/0.000/0.994	0.026/0.000/0.997	0.920/0.001/0.448	0.200/0.000/0.963
spread n=250	0.000/0.000/1.000	0.000/0.000/1.000	0.155/0.014/0.878	0.003/0.000/1.000
spread n=1000	0.000/0.000/1.000	0.000/0.000/1.000	0.025/0.013/0.976	0.001/0.000/1.000
spread n=4000	0.000/0.000/1.000	0.000/0.000/1.000	0.000/0.000/1.000	0.000/0.000/1.000
4 groups n=1000	0.018/0.000/0.995	0.014/0.000/0.997	0.734/0.035/0.535	0.082/0.001/0.969
null m=100, n=500	0.035/0.000/1.000	0.025/0.000/1.000	1.000/0.000/0.020	0.890/0.000/0.815
null m=100, n=2000	0.040/0.000/1.000	0.035/0.000/1.000	1.000/0.000/0.020	0.900/0.000/0.805
graded null (1000 reps)	0.053/0.000/0.999	0.039/0.000/0.999	0.914/0.000/0.438	0.008/0.000/1.000
graded spread n=1000	0.000/0.000/1.000	0.000/0.000/1.000	0.000/0.000/1.000	0.000/0.000/1.000
graded 4 groups	0.020/0.000/0.990	0.015/0.000/0.990	0.755/0.040/0.545	0.010/0.000/0.990

The closed-form rule. With the directional constant $z_{1-\alpha/(m(m-1))}$ the Bonferroni rule is valid and, as the union bound predicts for positively correlated contrasts, conservative: null familywise error 0.025–0.040 across the five null configurations (Table 5) against 0.035–0.053 for max-t, at the cost of a critical value 1–7% larger on the real matrices (App. C). The bootstrap max-t rule, which learns the joint tail, is the default everywhere in the paper; the closed form is the convenience variant for aggregate-only boards (§6).

Bootstrap-depth sensitivity. Reported quantities are stable in the bootstrap depth B with one borderline exception: the community-board (AlpacaEval, m=221) top-1 plausible set is 6 at $B \leq 800$ and 5 at $B \geq 2{,}000$; the paper reports the converged B=5,000 value.

Power calibration. The closed-form certification probability $\Phi(\Delta\sqrt{n}/\sigma - c)$ tracks the empirical certification rate across all twenty prediction bins in the synthetic study (maximum deviation 0.03–0.05) and across the full range on real matrices (random model pairs, item-subsampled replicas). Out-of-sample flip calibration pooled over the natural experiments is shown in Fig. 3.

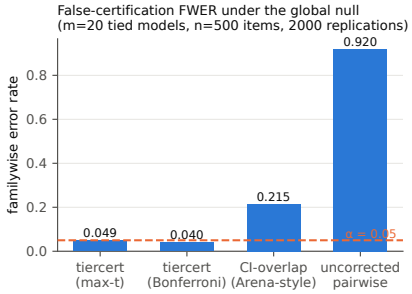


Figure 2: Null-scenario familywise error, visual companion to Table 5 (first row).

C Complete audited leaderboards

Tier compositions at $\alpha = 0.05$ (max-t unless noted):

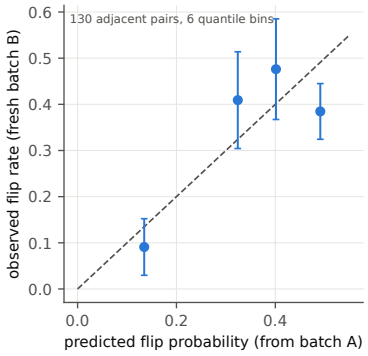


Figure 3: Out-of-sample flip calibration pooled over the natural experiments: closed-form flip probabilities predicted from batch A vs. realized flip rates on fresh batch B (quantile bins, binomial error bars). Bins are small; this shows consistency rather than sharp calibration.

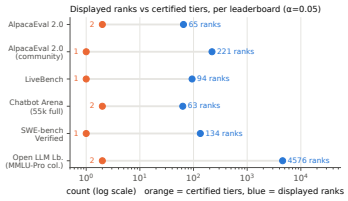


Figure 4: Displayed ranks (blue) vs. certified tiers (orange) per leaderboard, log scale, $\alpha = 0.05$ (Table 3 in one picture).

AlpacaEval 2.0 (displayed, m=65): T=2 with tier 1 = {yi-large-preview} and tier 2 the remaining 64; displayed #2 (gpt-4o-2024-05-13) is certified to ranks 2–3. **LiveBench (m=94):** T=1; top-1 plausible set has 6 members (87 under task clustering). **Chatbot Arena (m=63):** T=2 with tier 1 = {gpt-4-1106-preview, gpt-4-0125-preview}; displayed #5 has rank set 3–22. **SWE-bench Verified (m=134):** T=1; certified-pair density 0.74; 19 submissions plausibly #1. **MATH Lvl 5 top-100 (per-item):** T=2 with tier sizes $[76, 24]$: the one certified boundary separates the bottom 24 models from the top 76, and 20 models are plausibly #1 (identical under Bonferroni and max-t, c_α 4.41 vs. 4.24). **MMLU-Pro top-100 (per-item; T, tier sizes, and top-1 identical under Bonferroni and max-t, c_α 4.41 vs. 4.12):** T=5 with tier sizes $[1, 1, 2, 1, 95]$; the unique certified champion is MaziyarPanahi/calme-3.2-instruct-78b. **IFEval / BBH / GPQA / MuSR top-100 (per-item):** T=1; top-1 plausible sets of 14, 4, 24, and 7. **Open LLM Leaderboard v2 (full boards, bracketed):** $T \leq 2$ on every benchmark at every bracket setting.

All six top-100 rows are computed from the gated per-item matrices under the max-t rule (B=2,000; 5,000 for MMLU-Pro); the closed-form Bonferroni rule, which sits 4–7% above the max-t critical values on these matrices, returns the same T and tier sizes on every benchmark, widening top-1 sets by at most four models (IFEval 14 → 18, GPQA 24 → 27).

Full rank-set renderings for the AlpacaEval and LiveBench boards are shown in Figs. 5 and 6; machine-readable audits for every platform (per-model rank sets, tier memberships, certification matrices) ship with the code release.

D Robustness

γ brackets vs. exact per-item recomputation (OLL top-100). Exact paired certification vs. the bracket predictions: IFEval T=1 (γ_{med}=0.51), BBH T=1 (0.45), GPQA T=1 (0.58), MuSR T=1 (0.37), MATH Lvl 5 T=2 (0.57), MMLU-Pro T=5 (0.22). On five of six benchmarks the exact result equals the bracket range [γ=1 : 1–2, γ=0.4 : 1–3]; on MMLU-Pro the bracket under-predicts ($T = 5$), i.e. the brackets are conservative.

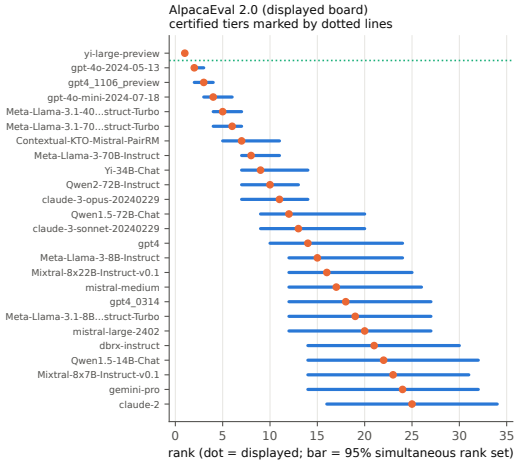


Figure 5: The AlpacaEval 2.0 displayed board, top 25: displayed rank (dot), 95% simultaneous rank set (bar), certified tier boundary (dotted line).

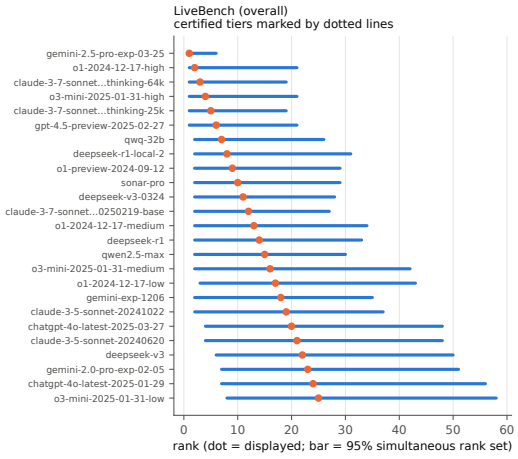


Figure 6: The LiveBench overall board, top 25: displayed rank (dot), 95% simultaneous rank set (bar). No tier boundary certifies.

Crowding without the range confound. Random m'-subsets shrink the observed range; range-anchored subsets hold it fixed (Table 6). The decline in $T(m')$ survives anchoring, and the law evaluated with each subset's measured R tracks both schemes.

Table 6: P2 on AlpacaEval: random vs. range-anchored model subsampling (100 draws per point).

scheme	m'	6	16	32	48	65
random	T (obs / law-$R(m')$)	3.9/3.9	4.0/5.2	3.2/3.8	2.5/2.2	2.0/1.5
random	$R(m')$	0.31	0.43	0.51	0.54	0.56
anchored	T (obs / law-R)	5.2/4.8	5.0/6.7	3.5/4.4	2.0/2.4	2.0/1.5
anchored	$R(m')$	0.56	0.56	0.56	0.56	0.56

Zero-parameter cross-platform prediction. Measuring $(n, m, R, \bar{\delta})$ per platform and predicting $\mathbb{E}[T]$ with no fitting: AlpacaEval 1.5 vs. observed 2; LiveBench 1.0 vs. 1; SWE-bench 1.0 vs. 1; AIME 2025 1.0 vs. 1; MMLU-Pro 1.4 vs. 5. This admits two readings. First, four of the five full boards sit in the near-saturated $T \approx 1$ regime, where a correct prediction is undemanding; the informative tests of the law are the subsampled regimes where T ranges over 1–5.2: the anchored crowd-

ing curve (predictions 1.5–6.7 vs. observations 2.0–5.2, Table 6) and the dense item-scaling curves, which the law tracks within its stated tolerance. Second, the one full-board miss is *diagnosable in advance*: MMLU-Pro's four boundaries sit exactly where the QQ check flags non-exponential (clustered) frontier spacings, so the spacing assumption fails precisely where it is used (MMLU-Pro's QQ correlation is 0.68 vs. 0.92–0.96 elsewhere). The QQ statistic thus doubles as an applicability check for the law, and its failure mode (under-prediction) is conservative.

Heuristic tier grouping (head-to-head). On the real Arena battle structure with all strengths equal, silhouette-selected 1-D clustering of BT strengths reports >1 tier in 100% of 100 null replications and marginal-CI chaining in 31%, vs. 0% for certification; on the real release: clustering $T{=}2$ (agrees), CI chaining $T{=}9$. The silhouette is undefined at $k{=}1$, so the plain rule (search $k{=}2, \ldots, 8$) cannot report a single tier at all: it has no null. Adding the customary acceptance threshold (split only if the best mean silhouette is ≥ 0.5) does not help, because null BT strengths are roughly Gaussian and a two-way split of Gaussian noise scores 0.54–0.82 (mean 0.59) in every replication: the thresholded rule still reports >1 tier in 100% of nulls.

Duplicate entries and clustered-bootstrap sensitivity (LiveBench). The public judgment snapshot lists two models twice under case-variant names, as distinct runs with distinct per-question scores. Merging each pair (keeping the higher-scoring run, $m{=}92$) leaves the iid audit unchanged ($T{=}1$, top-1 set 6, c_α 4.30 vs. 4.31) and moves the task-clustered top-1 set from $87/94$ to $62/92$: in both cases a large majority of the board is plausibly #1 once task heterogeneity is acknowledged, which is the claim the number supports. The clustered critical value itself rests on only 6 resampled task clusters ($c_\alpha \approx 18$ vs. 4.3 iid) and should be read as a conservative bound, not a sharp quantile. Implementation note: the max-t bootstrap excludes pairs whose resampled SE is exactly zero (models tying on every resampled item), which otherwise yield degenerate studentized values; this matters only for the small-n category sub-boards and the 6-cluster bootstrap, where near-duplicate runs make exact resampled ties likely, and changes no headline result.

Cluster-robust variants of the positive results. The two headline certifications survive few-cluster inference, with the expected conservative compression. MMLU-Pro top-100 under subject-clustered max-t (14 subject clusters, $c_\alpha{=}11.5$ vs. 4.1 iid): $T{=}2$ with tier sizes $[4, 96]$, top-1 plausible set of 2, and certified-pair density 0.30 (vs. 5 tiers, unique champion, 0.71 under iid items). SWE-bench Verified under repository-clustered max-t (12 repository clusters, $c_\alpha{=}8.9$): $T{=}1$ unchanged, top-1 set $19 \to 51$, density $0.74 \to 0.49$—half the pairwise orderings remain certified even under this coarse clustering. As with LiveBench's 6-cluster bound, critical values from so few resampled clusters are conservative; the reported iid audits and these clustered variants bracket the truth.

Step-down refinement. The Romano–Wolf step-down procedure recomputes the max-t critical value over the not-yet-certified pairs and iterates (Romano & Wolf, 2005); it controls the same familywise error asymptotically and can only add certifications. On AlpacaEval it converges in three steps (c_α $4.18 \to 3.78$) and adds 34 certified pairs (density $0.765 \to 0.782$); on LiveBench, three steps ($4.28 \to 4.12$), 63 pairs ($0.575 \to 0.589$); on MMLU-Pro's top-100, three steps ($4.12 \to 3.90$), 80 pairs ($0.713 \to 0.729$). No tier boundary and no top-1 plausible set changes on any board: the pairs a boundary is waiting on are the near-adjacent ones, which no step-down reaches. We therefore keep the single-step rule as the default for its transparent inversion into rank sets.

P1 estimators. The linearized fit of $\log[(\bar{T}-1)/(m-1)]$ on $n^{-1/2}$ tests the form; its slope is dominated by the near-saturated points ($\bar{T}$ within 0.02 of 1, where one or two certifying draws in a hundred move the log by units), so the constant is estimated by nonlinear least squares of $\bar{T}$ on equation (2). Fitted-over-law ratios: AlpacaEval 8×8 grid 0.90 (linearized 1.39, form $R^2 = 0.96$); AlpacaEval $m'{=}16$ 0.97 (linearized 1.24, $R^2 = 0.84$); LiveBench $m'{=}16$ 1.04 (linearized 0.95, $R^2 = 0.42$). The two estimators bracket the closed form; the linearized one is the less reliable where the curve saturates.

Natural-experiment sensitivity. LiveBench no-task-filter variant: flips 0.44 observed vs. 0.40 predicted ($m{=}37$, $n_A{=}253$, $n_B{=}100$, $T_A{=}1$). AIME cross-year pairs (roster overlap only $m{=}9$): AIME 0.125 vs. 0.254; HMMT (Harvard–MIT Mathematics Tournament, February papers) 0.375 vs. 0.316 (small samples, same direction). The 8×8 subsampling grids share items across points, so

their R^2 overstates independent-sample fit quality; the slope comparison, not the R^2, is the meaningful statistic.

E REPRODUCIBILITY

All experiments run CPU-only with zero API cost. Data: the AlpacaEval repository (per-instance annotations), LiveBench and Arena public releases and MathArena competition outputs on HuggingFace, the SWE-bench `experiments` repository, and Open LLM Leaderboard `contents` plus gated per-item details (accessible after accepting the terms). `run_all_minimal.sh` re-runs the audit suite from the data files at reduced replication counts and needs no network access or credentials; scripts whose inputs are gated (the OLL top-100 per-item recomputation, `exp6b_*`, which needs a Hub token with the details terms accepted) skip with a printed warning and the shipped result JSONs are used instead. Every table and figure regenerates from the committed result JSONs; the full-run commands behind the shipped JSONs are listed in the code README; seeds are fixed throughout. Every reference was verified against arXiv listings and publisher records.