WHAT IS THE TOP OF A LEADERBOARD WORTH? MEASURING AND CORRECTING SELECTION BIAS IN TABULAR AND TIME-SERIES BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A leaderboard's top row is a *selected* quantity: the winner minimizes average error over the very datasets on which its margin is then reported. We measure what this costs on four method×dataset matrices (TabArena, GIFT-Eval, fev-bench, and a long-horizon grid of our own) with a subsampling audit whose ground truth comes from disjoint held-out datasets. At $D'{=}8$, the suite size typical of long-horizon forecasting tables, TabArena's top entry keeps only $8.0 \pm 0.3\%$ of its reported advantage over a tuned gradient-boosting baseline out of sample on our primary scale (-8.5% to $+38\%$ across three aggregations, so 0.4 is the defensible ceiling); the reported number one is the true number one 38% of the time; on GIFT-Eval the leader's margin over the runner-up reverses sign; and drawing whole source *families* rather than i.i.d. units costs more still. Textbook remedies fail: leave-one-dataset-out returns *exactly* the naive value in 100% of folds on all four matrices, and the classical $\sqrt{2 \log M}$ envelope overpredicts measured optimism by 4.8–$6.0\times$, mostly because entries differ in true effect rather than because they correlate (4.8–$4.9\times$ against 1.8–$2.1\times$). We therefore introduce SCB, a selection-corrected bootstrap that resamples datasets jointly across methods around a variance-matched plug-in, needs only the published matrix, and has a cluster-blocked variant for families. It removes bias where there is bias and does not cry wolf where there is none: at full suite size TabArena's margin gives up about a third, while two champions and a published ICLR 2025 table keep 97.7% or more. Code, data and corrected leaderboards are in the anonymized supplement.

1 INTRODUCTION

Empirical machine learning runs on method×dataset result matrices: a tabular or forecasting paper evaluates M methods on D datasets, averages a per-dataset score into a column of means, bolds the smallest one, and reports the winner with its margin; living leaderboards (Erickson et al., 2025; Aksu et al., 2024; Shchur et al., 2025) do the same at larger M and D. The community has grown suspicious of the *stability* of these rankings: winners flip under alternative aggregation choices, top groups are statistically indistinguishable, and small changes to the setup shift conclusions (Oh, 2026; Brigato et al., 2026; Rubachev et al., 2025; Dehghani et al., 2021). What it does not provide is the quantity a reader needs: *conditioned on this leaderboard, how much of the advertised advantage of its top row would I realize on a dataset that is not on the leaderboard?*

This question is not answered by a significance test, since it is not a hypothesis about a fixed method: the top row is chosen *by* the matrix, so its reported mean inherits the optimism of a maximum over M correlated, noisy columns, a winner's curse (Cawley & Talbot, 2010; Andrews et al., 2024). Nor by cross-validation: leave-one-dataset-out selection returns *exactly* the published number in 100% of folds on all four matrices, since one deletion never changes the winner of a 45–119 entry leaderboard. Nor by classical multiplicity theory: the textbook $\sigma\sqrt{2 \log M}$ envelope assumes exchangeable candidates of equal true effect, and real leaderboards are neither, so a correction built on M overpredicts the measured optimism by 3.9–$5.0\times$ at the exact expected maximum (4.8–$6.0\times$ at the looser envelope) and the effective-count version still by 1.8–$2.5\times$ (§4).

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 treat the problem as estimation rather than testing. Let A_{NAIVE} be the advantage over a fixed reference that the leaderboard reports for its top row, and A_{true} the advantage the *selected* method delivers in expectation on new data. We develop and validate an estimator of A_{true} from the published matrix alone: a bootstrap resampling datasets *jointly across methods* (inheriting the empirical correlation structure that defeats the closed form) around a plug-in truth whose cross-method variance is matched to a method-of-moments estimate of the true effect spread. Both choices are load-bearing: an unshrunken plug-in under-corrects, James–Stein over-corrects, and the variance-matched plug-in lands within two Monte-Carlo SE on 18 of 19 cells of a synthetic grid with known truth (17 at 1.96).

Our protocol supplies ground truth without assumptions. From each full matrix we repeatedly draw D' datasets, let the mechanism pick its winner and report its margin there, then score that same method on the disjoint complement. The reported-minus-realized gap is the selection optimism at suite size D'; the realized number is what every estimator is judged against. We sweep D' from 4 to 48, with particular attention to $D' \approx 8$, the suite size of the tables that dominate long-horizon forecasting (Zeng et al., 2023; Chen et al., 2025); tabular suites are larger (Grinsztajn et al., 2022; McElfresh et al., 2023), which is why we also report every leaderboard at its own full D. We run it twice, drawing units i.i.d. and then whole *source families*, because three of our four matrices are configurations of a smaller set of underlying series, so an i.i.d. split leaves siblings of the held-out units in the selection set.

Findings. Three regimes emerge, and distinguishing them is the point of the paper. On *packed* leaderboards (TabArena, where seven entries are statistically indistinguishable from the top), selection dominates: at D'=8 the top row's advantage over tuned-and-ensembled CatBoost retains 8%, and even at the full D=51 our corrected estimate removes a third of it. On GIFT-Eval the champion's advantage over statistical baselines is real and large, but the *identity* race is illusory: the #1-versus-#2 margin reverses sign out of sample. On *champion* leaderboards (fev-bench, and a published ICLR 2025 forecasting table we re-analyze cell-for-cell), the correction is 0.3% and 0.05%: the estimator certifies transferable claims instead of shrinking every one, which is what a reporting standard has to do.

A replicated axis in two matrices additionally splits the driving noise into a method×dataset interaction σ_γ^2 and a replicate term σ_e^2. The interaction dominates (median replicate ICC, ICC_{rep}, of 0.75 on TabArena's 30 folds and 0.66 on our 8-block grid), giving a budget rule with measured constants: at fixed training runs the R-seeds-on-D-datasets convention pays a variance penalty of $(\sigma_\gamma^2 R + \sigma_e^2)/(\sigma_\gamma^2 + \sigma_e^2)$ against one run on RD datasets, rising to 7.8× at R=10. The same machinery prices dataset redundancy: the canonical LTSF table carries $D_{\text{eff}} \approx 5.2$ effective units out of 36, quantifying the ETT redundancy observed qualitatively by Brigato et al. (2026).

Contributions.

- **A decision-relevant audit** of four result matrices under a disjoint-holdout subsampling protocol, run i.i.d. and family-blocked, measuring how much of a top row's reported margin transfers, as a function of suite size, reference strength, unit definition and aggregation scale (§4).
- **Two negative results**: leave-one-dataset-out has zero corrective power on real leaderboards (winner unchanged in 100% of folds, all four matrices), and multiplicity corrections overshoot, mostly because candidates differ in true effect, not because they are correlated (§4).
- **SCB**, a selection-corrected bootstrap estimator of the top row's transferable advantage from the published matrix alone, with a variance-matched plug-in whose necessity we show by ablation and a cluster-blocked variant for family structure; within Monte-Carlo error on synthetic truth, and lower RMSE on a majority of real cells (§5).
- **Corrected leaderboards and design rules**: corrected top rows for the three public matrices; a cell-level re-analysis of a published ICLR 2025 table; breadth-over-replication budgets; and effective-suite-size estimates with cluster-bootstrap intervals (§6).

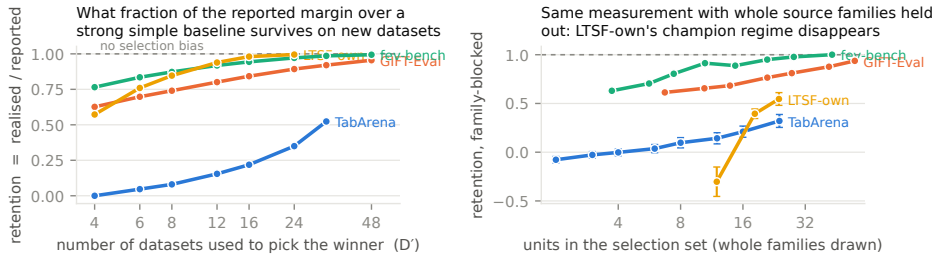

Figure 1: **The top row of a leaderboard overstates its transferable advantage, and the effect is governed by suite size and how packed the leaderboard is.** *Left:* the fraction of the reported margin over a strong reference (§4) surviving on disjoint held-out datasets, against the number of datasets D' used to select the winner. *Right:* the same with whole *source families* held out, so no sibling straddles the split (± 1.96 Monte-Carlo SE throughout). TabArena, whose 51 units are distinct datasets, is the control and barely moves; LTSF-own, whose 36 units are 9 series $\times$ 4 horizons, goes negative at two held-in families and $+0.39$ at three (App. A.11).

2 Related Work

Fragility of benchmark rankings. A 2025–2026 wave of work established that published rankings in our two domains are unstable: Oh (2026) finds that in over half of top-model comparisons across ten public leaderboards, TabArena and TSFM-Bench among them, at least one assumed property of superiority fails; Brigato et al. (2026) train $\sim$5,000 networks to show long-horizon rankings are setup artifacts, and that averaging the four ETT datasets lifted a win rate from 33.3% to 71.4%; Rubachev et al. (2025) show tabular rankings flip under temporal splits; Zhang & Zhao (2026) decompose an AutoML win into test-set selection and unenforced budgets; earlier work named the benchmark lottery (Dehghani et al., 2021) and contested the deep-versus-simple narratives (Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022; McElfresh et al., 2023; Zeng et al., 2023; Tan et al., 2024; Zhu et al., 2026). We take this fragility as *premise*: our object is the corrected *value* of the selected top row, which none of these estimate.

Statistics of cross-dataset comparison. The classical toolchain of Dietterich (1998); Nadeau & Bengio (1999); Demšar (2006), its Bayesian variants (Benavoli et al., 2017; Corani et al., 2017) and autorank (Herbold, 2020) tests hypotheses about *prespecified* methods and does not address post-selection reporting; Bouthillier et al. (2021) and Neuhof & Benjamini (2026) quantify within-task variance and hierarchical rank uncertainty, the latter flagging post-selection inference as open. For LLM evaluation Miller (2024); Kotawala (2026); Mustahsan et al. (2025) give clustered errors, resolution diagnostics and an ICC decomposition with the more-items-fewer-trials rule; our budget identity imports that classical result (Cochran, 1977; Raudenbush, 1997) and supplies the constants for domains where both D and R cost a training run, while the RL line (Colas et al., 2018; Agarwal et al., 2021; Patterson et al., 2024) allocates seeds within environments, not datasets versus replicates.

Winner's curse and post-selection inference. Estimating a parameter after selecting it is the subject of the selective-inference literature (Berk et al., 2013; Fithian et al., 2014), with conditional and empirical-Bayes treatments of winners in econometrics and statistics (Andrews et al., 2024; Efron, 2011). Two older lines supply the pieces SCB recombines, and we claim neither one. Our variance-matched plug-in is, in the homoscedastic case, algebraically Louis's constrained empirical-Bayes estimator (Louis, 1984; Ghosh, 1992; Shen & Louis, 1998): weight $\tau^2/(\tau^2+v)$ times inflation $\sqrt{(\tau^2+v)/\tau^2}$ is exactly our c, and so is the motivation—posterior means are right for per-element loss and wrong when the *ensemble spread* is the target. Resampling-and-reselecting to subtract an average selection gap is the standard winner's-curse correction in genome-wide association studies (Sun & Bull, 2005; Zöllner & Pritchard, 2007; Faye et al., 2011). SCB adds the joint column resampling that inherits the empirical cross-method correlation rather than assuming independence, the estimand (a transferable margin, not an effect size at a locus), and the protocol validating it. Closest in construction is the forecast-evaluation line of White's Reality Check (White, 2000), Hansen's SPA test (Hansen, 2005) and the Model Confidence Set (Hansen et al., 2011), which resample evaluation units jointly across models. Beside them sit Hung & Fithian (2019), whose conditional

rank-verification test asks whether the observed winner is truly best, an identity question rather than a transferable margin, and Tibshirani & Tibshirani (2009), who correct the bias of the minimum of correlated cross-validated means. Those target the best model's loss on a fixed evaluation sample; we target the transferable margin over a dataset super-population, and implement both Tibshirani & Tibshirani's correction and Efron & Morris-style shrinkage as comparators (§5). In ML, Cawley & Talbot (2010) diagnosed selection-induced optimism and McLatchie & Vehtari (2023); Blum & Hardt (2015); Dwork et al. (2015); Recht et al. (2019); Roelofs et al. (2019); Xu et al. (2026); Schneider et al. (2025); Balcan et al. (2021) correct or bound it *within* one dataset, one test set, one benchmark, or a portfolio. Three lines come closest, and each stops short of this estimand. Hothorn et al. (2005) formalize datasets as draws from a data-generating process with variance components, which is the framing of §3, but analyze a fixed method set. Dodge et al. (2019) correct selection over hyperparameters *within* one dataset. Nießl et al. (2022) resample to quantify over-optimism arising from analyst degrees of freedom, a different target from the out-of-suite value of one published margin. The methods×datasets matrix (selection over correlated method columns, units that are entire datasets, evaluation at $D \approx 8$) is covered by none: to our knowledge no prior work estimates the transferable value of a leaderboard's top row from the matrix alone, nor reports the failure of leave-one-dataset-out or the M_{eff} gap. TabRepo (Salinas & Erickson, 2024) uses leave-one-dataset-out to build *portfolios*, presuming transfer rather than estimating it.

3 Setting: matrices, scales, and the estimand

Result matrices. A benchmark is a pair of a method set $\{1, \dots, M\}$ and a suite of evaluation units $\{1, \dots, D\}$, with $E[m, d] > 0$ the error of method m on unit d. We analyze four matrices spanning both domains (Table 3, appendix): **TabArena** (Erickson et al., 2025), 45 methods × 51 datasets with a replicate axis; **GIFT-Eval** (Aksu et al., 2024), 119 forecasting models × 97 configurations (MASE), statistical baselines through foundation models; **fev-bench** (Shchur et al., 2025), 22 models × 91 tasks (MASE); and **LTSF-own**, a grid we run in the canonical long-horizon protocol (Zeng et al., 2023): 9 datasets × 4 horizons, 13 methods from naive through trained linear variants to MLPs, 8 rolling-origin test blocks per cell. At full channel width it reproduces the literature's anchors on all three datasets we check ($\text{MSE} = 0.370/0.140/0.410$ at L=336, H=96 on ETTh1/Electricity/Traffic against $0.375/0.140/0.410$ in Zeng et al., 2023), and the minimal-scale channel caps leave the within-cell method ordering intact (Spearman ≥ 0.96, App. A.7). The zoo is deliberately simple: it is the one matrix whose replicates, splits and family labels we control end to end—what §6's variance-component and design-effect analyses require. Each matrix is reduced to its complete-case rectangle, an outcome-dependent filter worth stating: fev-bench loses 9 of 100 tasks on which some model returned nothing, GIFT-Eval 2 of 121 models. Trading models for tasks instead moves retention at D'=8 by at most $+0.015$ and never changes the winner (App. A.4). The public matrices carry the modern architectures it omits, so no conclusion rests on our zoo alone.

Comparison scale. Averaging raw errors across heterogeneous datasets is meaningless, so all analysis is on a paired, scale-free loss: fixing a reference method m_0 named a priori per suite, never by rank,

$$Z[m, d] \;=\; \text{clip}\big(\log E[m, d] - \log E[m_0, d];\; [-\log 10,\, \log 10]\big), \tag{1}$$

the (clipped) log skill score, identical in spirit to fev-bench's clipped geometric-mean skill and to the tabular literature's shifted geometric means (Shchur et al., 2025; Ye et al., 2024). Lower is better and $Z[m_0, \cdot] \equiv 0$. Because §4 subsamples the *method* axis, per-dataset normalization must not depend on the current method set: we also report a *studentized* scale (Z over the per-dataset cross-method SD computed once on the full set) and a *frozen-rank* scale (per-dataset ranks on the full set, rescaled to $[0, 1]$), both frozen under subsampling. Every headline result is reproduced on all three (§6).

Units and families. A benchmark's units are not always distinct sources. TabArena's 51 units are 51 OpenML datasets, but GIFT-Eval's 97 configurations come from 28 base datasets (four of them contributing 8 apiece), fev-bench's 91 tasks include ETT at three frequencies, and our LTSF grid's 36 cells are 9 series × 4 horizons. We assign every unit an *a priori* family label (source dataset; source series with the frequency suffix stripped; identity) and report the audit both ways, with TabArena as the control.

Leaderboard mechanism and estimand. Write $\hat{\mu}_m = \frac{1}{D} \sum_d Z[m, d]$ for the leaderboard column, $m^\star = \arg\min_m \hat{\mu}_m$ for the top row, and

$$A_{\text{NAIVE}} = -\hat{\mu}_{m^\star}, \qquad A_{\text{true}} = -\mu_{m^\star}, \qquad \rho = A_{\text{true}}/A_{\text{NAIVE}}, \tag{2}$$

Table 1: Retention of the top row's reported advantage over the strong reference (log-ratio scale): realized-on-complement over reported-on-subsample at selection suite size D'; below zero, the selected method was *worse* than the reference out of sample. Retention and $P(\#1)$ use 40,000 draws per cell (Monte-Carlo SE ≤ 0.005, printed on the D'=8 column); Table 7's 600-draw pass sits 0.02–0.03 higher at D'=8, and the estimator comparison of §5 runs at 600 draws. Right block, at $D' = 8$: the #1-over-#2 margin as reported and as realized; $P(\#1)$, the chance the chosen #1 is also the complement's #1, with, in brackets, what *independently drawn* selection sets of 8 and of 21 units achieve on a common 22-unit scoring set (8 and 14 on 14 units for LTSF-own, D=36), i.e. what enlarging the suite buys (App. A.3); and $\bar{r}$, the mean number of methods that beat the chosen one out of sample.

	Retention vs. strong reference						Top-1 vs. top-2 at $D' = 8$		
Matrix	D'=4	6	8	12	16	24	margin rep. → real.	P(#1 holds) [8→21u]	$\bar{r}$
TabArena	−.000	.047	$\mathbf{.080}^{\pm.003}$	.155	.219	.350	+0.015 → +0.006	0.38 [.27→.33]	5.6
GIFT-Eval	.627	.698	$\mathbf{.740}^{\pm.001}$	.801	.843	.893	+0.077 → −0.014	0.34 [.25→.34]	22.0
fev-bench	.765	.835	$\mathbf{.873}^{\pm.002}$	.918	.944	.971	+0.024 → +0.018	0.58 [.49→.66]	1.1
LTSF-own	.573	.760	$\mathbf{.847}^{\pm.003}$	.939	.981	.995	+0.049 → +0.056	0.89 [.82→.90]	0.2

where μ_m is method m's population mean contrast. A_{NAIVE} is what the leaderboard advertises, A_{true} what the practitioner who follows it obtains, and *retention* ρ the exchange rate. Note $m^\star$ is random: A_{true} is the value of the *selected* method, not the oracle-best one. Andrews et al. (2024) distinguish the estimand conditional on the winner's identity from the procedure average; SCB corrects the latter, and §5 reports a conditional variant indistinguishable from it here.

Two readings of μ deserve separation, because dataset suites are not i.i.d. samples from anything. Under a *superpopulation* reading units are draws from a task distribution and $\mu_m = \mathbb{E}_d Z[m, d]$, licensing "on a new dataset" language at the price of an exchangeability assumption false for, say, the ETT family. Under a *design-based* reading the suite is fixed and the estimand is a protocol: draw A, $|A| = D'$, uniformly; select $m^\star(A)$; score on the complement. All randomness is the analyst's own, and every number in §4 is of this kind. It is defined only for $D' < D$ while the number people want is at D'=D: the gap §5 fills, and why it must be validated where assumption-free truth exists.

Decomposition. For fixed m, $\mathbb{E}[\hat{\mu}_m] = \mu_m$, but Jensen's inequality applied to the (concave) minimum gives $\mathbb{E}[\min_m \hat{\mu}_m] \leq \min_m \mu_m$, and $\min_m \mu_m \leq \mu_{m^\star}$ holds pointwise, hence

$$\underbrace{\mathbb{E}[A_{\text{NAIVE}}]}_{\text{advertised}} \;\geq\; \underbrace{-\min_m \mu_m}_{\text{oracle value}} \;\geq\; \underbrace{\mathbb{E}[A_{\text{true}}]}_{\text{delivered}}, \tag{3}$$

so the advertised-minus-delivered gap splits into *optimism* plus *selection regret*, both vanishing as $D \to \infty$. Both inequalities hold in expectation only, so retention is *not* confined to $[0, 1]$: we observe $\rho \leq 0$ on TabArena at D'=4, and any estimator must be able to report such values.

4 The audit: what selection costs on real matrices

Protocol. For each matrix, scale and suite size $D' \in \{4, 6, 8, 12, 16, 24, 32, 48\}$ we repeat the draw: take D' units uniformly without replacement, let the mechanism select $m^\star$ on that sub-matrix and record its reported advantage over a fixed reference, then score $m^\star$ on the disjoint complement (40k draws for retention, 600 for the estimator comparison). References are named *a priori*, never by rank: a weak one (default random forest; seasonal naive) and a strong practitioner default (tuned+ensembled CatBoost; AutoTheta; a statistical ensemble; Linear-M, the medium-lookback linear model). The *family-blocked* variant holds out whole families instead.

Retention is a property of the leaderboard, not a universal constant. Table 1 and Figure 1 show the three regimes. On TabArena, whose top is packed (seven methods tied with the leader: AutoGluon, REALMLP, TabM, TabPFNv2 and CatBoost variants within noise of one another), the margin over tuned CatBoost is almost entirely selection: retention is $8.0 \pm 0.3\%$ at D'=8 and only 35% at D'=24; at D'=4 the selected method is on average *no better* than the reference out of sample. On GIFT-Eval the absolute claim survives (74.0% at D'=8: pretrained forecasters really do beat AutoTheta) while the identity race collapses: the #1-versus-#2 margin of +0.077 log units becomes

-0.014 out of sample, and 22.0 of the other 118 methods beat the selected leader on average. On fev-bench and LTSF-own, each led by a genuine champion (Chronos-2; NLinear over our deliberately weak zoo), retention approaches one, and the audit does not manufacture pessimism where none exists. The bracket says what a larger, *independently drawn* selection set buys (partitions of a fixed suite penalise the larger slice mechanically, App. A.3): 21 TabArena datasets identify the 22-dataset winner 33% of the time against 8 datasets' 27%, so the top resolves, but a suite $2.6\times$ larger is still wrong two times in three; the same enlargement buys $0.25 \to 0.34$ on GIFT-Eval and $0.49 \to 0.66$ on fev-bench, so the packedness belongs to TabArena, not the protocol. Retention is finally a property of the (leaderboard, reference) *pair*: against the weak reference that same top row keeps 0.81 rather than 0.08 at $D'=8$ (GIFT-Eval 0.70, fev-bench 0.93, LTSF-own 0.97), a comparison against a default random forest being no maximum over near-ties. The strong reference itself sits in the tied top group, so TabArena's 0.08 says the crown is contested, not that its top row is weak. We lead with the strong reference because it is the comparison a practitioner faces.

Blocking by source family removes an inflation that runs one way. Table 3 (right block, appendix) repeats the audit drawing whole families. Because retention is steep in the unit count and a blocked draw of G' families holds a random number of units, each blocked cell is compared against i.i.d. draws with sizes resampled from that cell's own unit-count distribution, removing the confound exactly. On TabArena, the control, where each unit is its own family, nothing systematic moves (cost -0.06 to $+0.05$), the check that blocking is not merely pessimistic. On the two public forecasting matrices the median cost below 15 units is 0.09 (range -0.02 to $+0.14$) and shrinks as the suite grows; our own grid, excluded there as the outlier under discussion, costs $+1.23$, hitting hardest where the i.i.d. analysis was most reassuring: our LTSF grid, a champion at 0.85, goes negative once a series' four horizons cannot straddle the split. That cell has exactly 15 realizations, so a Monte-Carlo interval is the wrong uncertainty: enumerated it is -0.35, dropping any one family moves it anywhere in $[-0.89, +0.28]$, and three held-in families give $+0.39$ (App. A.11); the direction is consistent, the magnitude is not identified. Nor does it turn on where the family lines fall (App. A.16): a size-matched random partition costs nothing anywhere, and splitting ETT into separate families still costs $+0.84$ against $+1.29$ pooled, though the retention *level* does not survive that split. The ladder does not show provenance is the only mechanism, though: on our grid, blocking by *horizon* costs $+0.88$ while leaving every series on both sides, so a systematic split can be expensive without shared provenance. *Counting configurations as datasets can manufacture champions.*

Multiplicity corrections overshoot, and not for the reason one expects. Subsampling the method axis (Figure 4, left, App. A.2) grows the top row's optimism monotonically on the packed matrices ($+0.004$ at $M=2$ to $+0.023$ at $M=45$ on TabArena) and non-monotonically on the champion matrices, returning to zero once the methods that separate the champion are present: the winner's curse needs a contested crown. Against the exact expected maximum of n Gaussians (the $\sqrt{2\log n}$ envelope Embrechts et al., 1997 is 20–36% looser still) the closed form predicts 0.088 and 0.130 against measured 0.023 and 0.026: overshoots of $3.9\times$ and $5.0\times$, halved by substituting the participation ratio $M_{\text{eff}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ (Gao et al., 2017; Li & Ji, 2005). *Which* idealisation fails is testable, but the two interact, so each is read at both corners: destroying the correlation costs 1.2–$1.4\times$ with true effects nulled and 2.6–$3.1\times$ at the real effects; destroying the effect differences costs 3.2–$3.4\times$ at independence and $\approx 7\times$ at the real correlation. The symmetric attribution, the geometric mean of the two orderings, multiplies back to the 8.6–$10.2\times$ total and gives correlation 1.8–$2.1\times$ against heterogeneity 4.8–$4.9\times$ (App. A.10). The envelope overshoots more because leaderboard entries have genuinely different means than because they are correlated, by a factor of about two and a half, not the five a single-corner reading suggests; M_{eff}, though a real diagnostic, must be read against its own floor (23.9 on TabArena, 53.4 on GIFT-Eval for independent methods), so 6.3 of 119 is an $8\times$ narrowing, not $19\times$. A procedure that inherits the correlation structure needs none of these inputs—the estimator we build next.

Why the obvious fix fails. Leave-one-dataset-out select-and-score is unbiased for the procedure value in expectation over suites. Applied to one observed leaderboard, however, it returns *exactly* A_{NAIVE} whenever removing a single dataset does not change $\arg\min$, and on all four matrices, across all three scales, the winner survives every fold: the top row of a 51–97 dataset leaderboard is far too entrenched for one deletion to matter. LOO is thus not conservative but *vacuous* at deployment, which removes the objection "just cross-validate over datasets." (At small D' the winner does move and LOO becomes informative, so we keep it as a competitor.)

5 SCB: CORRECTING THE TOP ROW FROM THE MATRIX ALONE

The estimator. SCB (selection-corrected bootstrap) estimates A_{true} from the published matrix in four steps.

1. Compute $\hat{\mu}_m = \frac{1}{D}\sum_d Z[m,d]$ and the per-column sampling variances $\hat{v}_m = \frac{1}{D(D-1)}\sum_d (Z[m,d]-\hat{\mu}_m)^2$.
2. Estimate the spread of *true* effects by method of moments, $\hat{\tau}^2 = \max\{0, \mathrm{Var}_m(\hat{\mu}) - \bar{\hat{v}}\}$, where $\bar{\hat{\mu}} = \frac{1}{M}\sum_m \hat{\mu}_m$, $\bar{\hat{v}} = \frac{1}{M}\sum_m \hat{v}_m$ and $\mathrm{Var}_m(\hat{\mu}) = \frac{1}{M-1}\sum_m (\hat{\mu}_m - \bar{\hat{\mu}})^2$ run over all M rows, the reference's (identically 0) included; form the *variance-matched plug-in* $\tilde{\mu} = \bar{\hat{\mu}} + c\,(\hat{\mu} - \bar{\hat{\mu}})$, $c = \sqrt{\hat{\tau}^2 / \mathrm{Var}_m(\hat{\mu})}$: a deconvolution preserving the ordering while shrinking the spread to $\hat{\tau}^2$.
3. For $b = 1, \ldots, B$: draw counts $w \sim \mathrm{Mult}\big(D; \frac{1}{D}, \ldots, \frac{1}{D}\big)$, form $\hat{\mu}^*_b[m] = \tilde{\mu}_m + \frac{1}{D}\sum_d w_d \,\mathrm{resid}[m,d]$ with $\mathrm{resid}[m,d] = Z[m,d] - \hat{\mu}_m$, and record $\beta_b = \hat{\mu}^*_b[k_b] - \tilde{\mu}[k_b]$ at $k_b = \arg\min_m \hat{\mu}^*_b$.
4. Report $A_{\text{SCB}} = -(\hat{\mu}_{m^\star} - \bar{\beta})$, $\bar{\beta} = \frac{1}{B}\sum_b \beta_b$.

Resampling whole columns makes every draw reuse the same weights across methods, so the empirical cross-method covariance of sampling errors, the object that defeats the closed form, is reproduced exactly, with no Gaussian or independence assumption. It does treat columns as independent, which §3 says they are not when units are configurations; the **cluster-blocked variant** therefore resamples *families* with replacement and estimates $\hat{v}_m$ from the between-family variance of family means (exact update in App. A.5). That correction runs in the conservative direction, because under within-family correlation the i.i.d. version understates $\mathrm{Var}(\hat{\mu})$, inflates $\hat{\tau}^2$, shrinks too little and under-corrects. It matters most where family structure is. At eight held-in GIFT-Eval families the residual bias falls from $+0.055$ to $+0.013$ and on our LTSF grid from $+0.072$ to -0.017, while on TabArena, where families are units, the two variants agree to 0.0002. What the plug-in does to the second moment is the crux. Raw $\hat{\mu}$ leaves a spread inflated by sampling noise, so the correction comes out too small; James–Stein shrinkage collapses it to $\tau^4/(\tau^2+\sigma^2)$, so the correction comes out too large. Matching to $\hat{\tau}^2$ is first-order correct, with mean bias $+0.054/-0.027/-0.012$ (none/JS/matched) at $M{=}30, D{=}8$ and $+0.017/-0.012/-0.000$ at $M{=}100, D{=}24$.

Synthetic ground truth. We validate on a generative model with known truth ($Z[m,d] = \mu_m + \gamma_{md}$, cross-method correlation ρ_c controlling M_{eff}): 19 cells sweeping D, M and ρ_c plus two degenerate regimes, 400 simulations each (App. A.6). The naive report is significantly optimistic ($z > 3$) in $17/19$ cells (the exceptions have nothing to correct), while SCB's residual is within two Monte-Carlo SE in $18/19$ ($17/19$ at 1.96, the marginal cell being $|z|{=}1.99$) and its RMSE beats naive in $17/19$; where it errs, at $\rho_c{=}0.8$, it over-corrects, the safe direction. The sharpest cell is the null: with *no true differences* ($\tau{=}0$, $M{=}30$, $D{=}24$) the naive top row advertises $+0.174$ log units of pure artifact; SCB reports $+0.006$.

Real-matrix validation. The subsampling protocol of §4 gives assumption-free truth on real data, so we score six estimators against it in one run (Table 5, App. A.2; every number here is rendered from `results/claims.json` by `code/paper_numbers.py`). At $D'{=}8$ the naive bias is $+.044$ (TabArena) and $+.167$ (GIFT-Eval); SCB leaves $+.015$ and $-.013$, at RMSE .066 and .217 against LOO's .066 and .273. Tibshirani & Tibshirani's correction is competitive on TabArena (RMSE .067) but collapses on GIFT-Eval, where its per-unit minimum over 119 methods sits far below the winner (bias $-.199$): a within-dataset correction does not transfer to a heterogeneous suite. A *conditional* SCB, averaging only over draws whose winner is the observed one, targets Andrews et al.'s estimand but is indistinguishable here (.064/.066, .221/.217). The comparator that matters is plain Efron & Morris-style empirical-Bayes shrinkage of the column (Efron & Morris, 1975), which ignores the selection event and *beats* SCB on RMSE ($59/87$ cells against 54): shrinking everything toward the mean is a variance bargain, paid for in bias where it matters—$-.049$ on fev-bench and $-.054$ on GIFT-Eval against SCB's $-.006$ and $-.013$, five points off a champion's genuine margin. For a number that is *reported* rather than averaged over hypothetical redraws that is the wrong trade, and the loss decides: report A_{SCB} when the published number is the selected winner's margin, where bias in that coordinate is the loss; use shrinkage when estimating all M columns at once under average squared error, where its champion bias is priced in. The two disagree materially only where a genuine

champion holds a large margin, which the tied count and the printed c flag in advance (Table 5). Across all 87 (matrix $\times$ scale $\times$ D') cells, SCB's RMSE beats the naive report's in $54/87$ and LOO's in $78/87$, a median relative change of -2.1% against naive (best -26.4%, worst $+8.8\%$) and -9.0% against LOO. That 62% is not dominance and we do not claim it: the cells are nested (three scales, eight sizes), and the aggregate hides what matters. Per matrix, SCB beats naive in $21/21$ TabArena and $17/24$ GIFT-Eval cells against $7/24$ on fev-bench and $9/18$ on LTSF-own: it wins wherever there is bias to remove and is a wash where there is none, leaving neither a packed leaderboard's optimism standing nor hallucinating a champion's.

Limits. The plug-in is an *affine* deconvolution: it matches the effect distribution's second moment and nothing else, so conditional on a fixed configuration it is not unbiased. The failure mode is a champion one to two standard errors clear of a tied pack, where the shrink factor $c = \sqrt{\hat{\tau}^2 / \operatorname{Var}_m(\hat{\mu})}$ collapses and SCB over-corrects, by up to 0.11 log units in a sweep of the lead (App. A.12). The diagnostic is c, free to print: it certifies the shrinkage applied, which bounds the room any shape error has, and on the four real matrices $c \in [0.970, 0.995]$, within 3% of the identity. Below $c \approx 0.8$ the correction is shape-sensitive and c alone cannot tell a champion from a calibrated spread, so A_{NAIVE} belongs beside A_{SCB} there. SCB also corrects selection over methods *given* the matrix: it cannot see tuning asymmetries (Shwartz-Ziv & Armon, 2022; Zhang & Zhao, 2026), curation of which methods appear at all (which can only make true optimism larger), or cross-paper test-set reuse (Blum & Hardt, 2015). It needs $D \gtrsim 6$ for $\hat{\tau}^2$, and at $\rho_c \geq 0.8$ it leaves a small conservative residue. Its TabArena residual is significantly positive at every audited $D' \geq 6$ (z up to 6.7); App. A.14 attributes it to the plug-in's *shape* rather than its noise (an oracle τ^2 leaves it untouched, an oracle μ removes it) and finds the prize uncollectable for now: a mixture and an NPMLE deconvolution both overshoot on all four matrices (z of -6 to -10), for the reason James–Stein does—a deconvolution fitted for per-element accuracy returns a G too narrow, and too narrow a plug-in leaves the bootstrap winner too much room to be lucky. Of the four plug-ins, variance matching is the only one whose residual is within Monte-Carlo error on three of the four matrices, so it stays the default and TabArena stays an upper bound—though the oracle is both unbiased and lowest-RMSE on TabArena, so the headroom is real.

6 APPLICATIONS

Corrected leaderboards. At full suite size SCB turns each headline into an estimate of what it is worth (Table 2). TabArena's top row (AutoGluon 1.3 at a 4-hour budget) advertises $+0.028$ over tuned+ensembled CatBoost; corrected, $+0.019$—a third of the margin over the strongest default is selection, while against a weak default (RF, $0.214 \to 0.203$) it barely matters. GIFT-Eval's champion keeps 97.7% of its (large) margin over AutoTheta and fev-bench's Chronos-2 keeps 99.7%: corrected leaderboards *certify* these. Family blocking moves them (94.9%, 100.9%; our grid $99.8\% \to 87.6\%$), so quote the clustered column when units are configurations. Because SCB's residual on TabArena is positive at every audited size (§5), the 67% is an extrapolation past the largest audited D' and an upper bound, and the audit says by how much. On $6{,}000$ shared draws the gap between what SCB implies and what is measured closes as the suite grows (0.476 against 0.196 at D'=16, 0.554 against 0.517 at D'=32), so carrying the last measured gap, 0.037, forward unchanged and subtracting it from the full-suite 0.669 puts the honest figure near 63% rather than 67%: still about a third of the margin given up. The asymmetry is the message: a procedure that cuts a packed leaderboard's margin by a third leaves a champion's intact, so publishing A_{SCB} beside A_{NAIVE} separates the regimes as no significance star does.

A published table, cell for cell. On the artifact the estimator is meant for, Table 1 of SimpleTM (Chen et al., 2025) with 11 methods and 8 datasets, the verdict is an acquittal with a caveat: the winner's 23.0% margin over the table's simple baseline survives correction almost intact (22.998%, a shave of 0.05%), on MAE too, but its 11 methods span $M_{\text{eff}} = 2.8$ axes and its 8 datasets carry $D_{\text{eff}} \approx 6.9$ units, so it discriminates less than its dimensions suggest even where its top claim is sound. We do *not* correct that table's #1-versus-#2 margin, though it is the number such tables emphasise: the runner-up is chosen by rank, so the contrast is a difference of two selected quantities that our bootstrap, which re-selects only the winner, is not validated on—and Table 1's margin column shows how badly it transfers (App. A.13).

Table 2: Selection-corrected top rows at full suite size (log-ratio scale). Each leaderboard's #1 over its *strong* reference (§4), reported and corrected, with the cluster-blocked correction beside it—the number to read when units are configurations of a shared source. Right: SimpleTM's ICLR 2025 Table 1 (M=11, D=8; MSE, horizons averaged).

Matrix	Reference	A_{NAIVE}	A_{SCB}	kept	clust.	SimpleTM Tab. 1	value
TabArena	CAT (t+e)	$+.028$	$+.019$	67%	67%	winner / rows won	SimpleTM, $7/8$
GIFT-Eval	AutoTheta	$+.581$	$+.567$	98%	95%	margin vs. DLinear	$23.01\% \to 23.00\%$
fev-bench	Stat. Ens.	$+.249$	$+.248$	100%	101%	M_{eff} (of 11)	2.8
LTSF-own	Linear-M	$+.061$	$+.061$	100%	88%	D_{eff} (of 8)	6.9 (ICC_{fam} CI $[0, 1]$)

Variance components and the breadth-versus-replication budget. Decomposing each paired contrast $\Delta_{dr} = \delta + \gamma_d + e_{dr}$, the method×dataset interaction dominates replicate noise in both domains (median $\text{ICC}_{\text{rep}} = \sigma^2_\gamma/(\sigma^2_\gamma + \sigma^2_e)$ of 0.75 on TabArena, 0.66 on LTSF-own), so under a fixed budget of $C = D \cdot R$ training runs the penalty for R replicates over $R\times$ more datasets is the classical $1 + \text{ICC}_{\text{rep}}(R-1)$ (Cochran, 1977; Raudenbush, 1997; Mustahsan et al., 2025); the constants are ours: $2.5\times$ at R=3 and $7.8\times$ at R=10 on TabArena, $2.0\times$/$5.4\times$ on LTSF-own (pooled components, implied ICC_{rep} 0.76/0.49; per-method medians give the *larger* $2.3\times$/$7.0\times$; App. A.9). The three-to-ten-seeds convention thus buys noise reduction at several times the price of breadth, which would also starve the winner's curse.

Effective suite size. Kish's design effect (Kish, 1965) gives $D_{\text{eff}} = D/(1 + (\bar{m} - 1)\,\text{ICC}_{\text{fam}})$, with ICC_{fam} the within-family correlation of unit scores, not the replicate ICC_{rep} above, and $\bar{m} = \sum_f m_f^2 / \sum_f m_f = 9.33$ is the size of the family a *randomly drawn cell* sits in, not the family average $36/6 = 6$, since ETT holds 16 of 36 cells (a separate constant $m_0 = 5.33$ debiases the ANOVA estimate of ICC_{fam}; App. A.8). Our LTSF table has $\text{ICC}_{\text{fam}} = 0.70$, so $D_{\text{eff}} \approx 5.2$: a $6.9\times$ design effect, 90% interval $[4.3, 12.0]$, while the 8-dataset SimpleTM table's ICC_{fam} interval covers $[0, 1]$. D_{eff} is a property of (suite, method set, scale): dropping our untrained baselines moves it to 6.9, the linear family alone to 13.5, and the studentized scale to 12.6 (App. A.8), so those coordinates should be reported with it. With few families D_{eff} is an order-of-magnitude diagnostic to be reported with its interval, but a 36-entry table is closer to five measurements than thirty-six, and $\sqrt{D}$ in any *analytic* error bar should be $\sqrt{D_{\text{eff}}}$ (a family-blocked bootstrap interval already carries the design effect and must not be scaled again).

Scale robustness. The three regimes survive on the studentized and frozen-rank scales, but magnitudes move and on TabArena a sign does too (0.080 log-ratio, -0.085 studentized, 0.384 frozen-rank at D'=8), so ≤ 0.4 rather than 0.08 is that matrix's defensible ceiling, and on two matrices the leader itself changes with the scale (Table 4).

7 Discussion

What we claim and what we do not. Not that leaderboards are useless: two of our matrices carry champions whose margins survive correction intact. We claim that the advertised margin is upward-biased, that on packed leaderboards it is most of the margin at the suite sizes papers publish at, and that a correlation-preserving bootstrap around a variance-matched plug-in corrects it from the matrix alone.

What to report. The audit reduces to five numbers, all computable from the published matrix: beside the naive margin, print A_{SCB} with its symmetric percentile-t interval, (D_{eff}, D), the tied count (entries whose paired-bootstrap 95% interval of the gap to the leader covers zero), and the shrink diagnostic c (below 0.8, print the naive margin too; §5). That interval covers 0.905 at a nominal 0.90 on a packed synthetic leaderboard, against 0.607 for the t interval papers print today (App. A.15).

Limitations. Our own zoo is linear-and-MLP only, fev-bench is missing 8 of 30 models, six families cannot identify a blocked magnitude (App. A.11, A.16), and leaderboards count configurations rather than learners (App. A.17). SCB corrects the margin over an a priori reference, not the #1-versus-#2 contrast (§6), and sees only in-matrix selection (Zhang & Zhao, 2026; Dwork et al., 2015), so A_{SCB} bounds from above what survives upstream tuning or curation.

REPRODUCIBILITY STATEMENT

The supplement contains the four result matrices with their provenance, the estimator library with unit tests, one script per experiment, the corrected leaderboards, and the resumable LTSF grid with its exact splits. App. A.7 gives the grid protocol and the anchor cells that reproduce published values, App. A.6 the generative model behind the synthetic validation, App. A.8 the design-effect constants, and §4 the subsampling protocol every number is measured under. Everything runs on two CPU cores with no GPU: the pipeline behind every number in the paper takes under an hour end to end, and the four Monte-Carlo diagnostics of App. A.14–A.16 and App. A.12 add about 50 min; the release README itemises the per-stage times and gives the run order.

Verification. Because this is a measurement paper, we made verification mechanical rather than editorial. The bodies of four of the six tables and the numbers quoted in §5 are rendered from one released results file by `code/paper_numbers.py`, and `code/consistency_scan.py` runs 718 assertions over the whole paper, 243 of which match a string as rendered in the compiled PDF while the other 475 recompute the quantity from the released files; the script prints that split when it runs. A value that is stale in the results or mis-transcribed into a printed table therefore fails the scan rather than reaching a reader. Derivations that are easy to get wrong were settled by simulation rather than by argument: the cluster-size constant in D_{eff} by `code/deff_sim_check.py`, and the ceiling design of App. A.3 by the partition-versus-independent comparison reported there. The pipeline reproduces two of three published anchors to three decimals and the third to within 0.005 (App. A.7); `code/repro_check.py` compares every numeric value in an earlier run against the released one: of $7{,}579$ numeric leaves, $1{,}299$ moved, and every one of them is attributable to a declared definition change carrying a written reason—the design-effect constant of App. A.8, the anchor-free definition of M_{eff} and the M_{eff}-based prediction downstream of it, and the a priori clip anchor of §3—while the rest are bit-identical; the script prints both counts; the estimators carry unit tests; and the novelty of the contribution was checked against the literature by hand. Every bibliography entry carrying an arXiv identifier or DOI was re-pulled from that record and diffed against our entry (author list, title and venue) by `code/check_bib.py`; the remaining entries, which carry neither an arXiv identifier nor a DOI, were checked by hand.

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: in particular, we did not use them 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. The simulated data analyzed in App. A.6 and App. A.12 is produced by the released simulation code, drawn from the generative models stated there; no data content was authored by a generative AI tool. 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.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024.

Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Generalization in portfolio-based algorithm selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 2015.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, 2021.

Lorenzo Brigato, Rafael Morand, Knut Strømmen, Maria Panagiotou, Markus Schmidt, and Stavroula Mougiakakou. There are no champions in supervised long-term time series forecasting. *Transactions on Machine Learning Research*, 2026. arXiv:2502.14045.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *International Conference on Learning Representations*, 2025.

William G. Cochran. *Sampling Techniques*. Wiley, 3rd edition, 1977.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018.

Giorgio Corani, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon. Statistical comparison of classifiers through Bayesian hierarchical modelling. *Machine Learning*, 106: 1817–1837, 2017.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

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

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. doi: 10.1080/01621459.1987.10478410.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025. arXiv:2506.16791.

Laura L. Faye, Lei Sun, Apostolos Dimitromanolakis, and Shelley B. Bull. A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design. *Statistics in Medicine*, 30(15):1898–1912, 2011.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Peiran Gao, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv 214262*, 2017.

Malay Ghosh. Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87(418):533–540, 1992.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2022.

Peter Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 16(3): 927–953, 1988.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79 (2):453–497, 2011.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

Steffen Herbold. Autorank: A Python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173, 2020.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47(2):758–782, 2019.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009. doi: 10.1214/08-AOS638.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956. doi: 10.1214/aoms/1177728066.

Leslie Kish. *Survey Sampling*. Wiley, 1965.

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

Jian Li and Lei Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–227, 2005.

Thomas A. Louis. Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79(386):393–398, 1984.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

Yann McLatchie and Aki Vehtari. Efficient estimation and correction of selection-induced bias with order statistics. *arXiv preprint arXiv:2309.03742*, 2023.

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

Zairah Mustahsan, Abel Lim, Megna Anand, Saahil Jain, and Bryan McCann. Stochasticity in agentic evaluations: Quantifying inconsistency with intraclass correlation. *arXiv preprint arXiv:2512.06710*, 2025.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

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

Christina Nießl, Moritz Herrmann, Chiara Wiedemann, Giuseppe Casalicchio, and Anne-Laure Boulesteix. Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *WIREs Data Mining and Knowledge Discovery*, 12(2): e1441, 2022.

YongKyung Oh. Position: State-of-the-art claims require state-of-the-art evidence. *arXiv preprint arXiv:2605.17273*, 2026.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25, 2024.

Stephen W. Raudenbush. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173–185, 1997.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks. In *International Conference on Learning Representations*, 2025.

David Salinas and Nick Erickson. TabRepo: A large scale repository of tabular model evaluations and its AutoML applications. In *International Conference on Automated Machine Learning*, 2024. arXiv:2311.02971.

Lennart Schneider, Bernd Bischl, and Matthias Feurer. Overtuning in hyperparameter optimization. In *International Conference on Automated Machine Learning*, 2025. arXiv:2506.19540.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Türkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Wei Shen and Thomas A. Louis. Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B*, 60(2):455–471, 1998.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Lei Sun and Shelley B. Bull. Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, 28(4):352–367, 2005.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ryan J. Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3(2):822–829, 2009.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking. *arXiv preprint arXiv:2605.05973*, 2026.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets. *arXiv preprint arXiv:2407.00956*, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Guilin Zhang and Kai Zhao. Winning by peeking: Unenforced budgets and test-set selection inflate short-budget AutoML comparisons. *arXiv preprint arXiv:2608.07303*, 2026.

Xiaokun Zhu, Louis Carpentier, and Mathias Verbeke. When foundation models are one-liners: Limitations and future directions for time series anomaly detection. In *International Conference on Learning Representations*, 2026.

Sebastian Zöllner and Jonathan K. Pritchard. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615, 2007.

A Appendix

Table 3: The four result matrices, and what family-blocking does to them. M_{eff} is the participation ratio of the cross-method correlation spectrum of per-dataset scores (log-ratio scale); "tied" counts methods whose paired-bootstrap 95% interval of the gap to the observed leader covers zero; "LOO same" is the share of leave-one-dataset-out folds leaving the winner unchanged, the degeneracy that voids LOO as a correction. The right block is retention of the top row's margin over the strong reference: i.i.d. draws at $D'=8$, then whole-family draws at two family counts with their mean unit count in parentheses (± 1.96 Monte-Carlo SE, 600 draws; §4 matches each blocked cell against i.i.d. draws at *its* unit counts). TabArena is the control.

							retention vs. strong reference		
Matrix	M	D	fam.	M_{eff}	tied	LOO same	i.i.d., 8	blocked	blocked
TabArena	45	51	51	4.0	7	100%	$+0.080$	$+0.10 \pm .05$ (8.0u)	$+0.32 \pm .07$ (24u)
GIFT-Eval	119	97	28	6.3	3	100%	$+0.740$	$+0.61 \pm .01$ (6.7u)	$+0.81 \pm .01$ (28u)
fev-bench	22	91	51	5.7	1	100%	$+0.873$	$+0.80 \pm .01$ (7.4u)	$+0.95 \pm .01$ (21u)
LTSF-own	13	36	6	2.7	1	100%	$+0.847$	$\mathbf{-0.30 \pm .15}$ (12u)	$+0.39 \pm .05$ (18u)

A.1 Full audit grids

Complete retention, bias, and RMSE tables for all $D' \in \{4, \ldots, 48\}$, all four matrices, and all three scales (log-ratio, studentized, frozen-rank), together with weak-reference and self-reference variants, are provided in the supplementary results files (`results/claims.json`, `results/claims_summary.md`). Table 4 summarizes retention at $D' = 8$.

The three regimes (packed / mixed / champion) are the same on all three scales, but retention is not sign-stable: TabArena is $+0.080$, -0.085, $+0.384$, so on the studentized scale its top row is worse than tuned CatBoost out of sample. The remaining eleven matrix–scale cells keep their sign.

Table 4: Scale robustness: retention at $D' = 8$ (40,000 draws, SE ≤ 0.005) and the full-suite leader, per aggregation scale.

Matrix	log-ratio	studentized	rank	leader (log-ratio)	leader (rank)
TabArena	+0.080	−0.085	+0.384	AutoGluon 1.3 (4h)	AutoGluon 1.3 (4h)
GIFT-Eval	+0.740	+0.756	+0.937	EXAONE-Forecast-Agent	STRIDE w/ Synapse
fev-bench	+0.873	+0.902	+0.885	Chronos-2	Chronos-2
LTSF-own	+0.847	+0.871	+0.522	NLinear	Ridge

A.2 Additional figures

Figure 2 gives the estimator comparison of Table 5 across the full range of suite sizes, Figure 3 the nominal against effective method counts with the noise floor of App. A.10, and Figure 4 the method-axis sweep and the replication-budget penalty of §6.

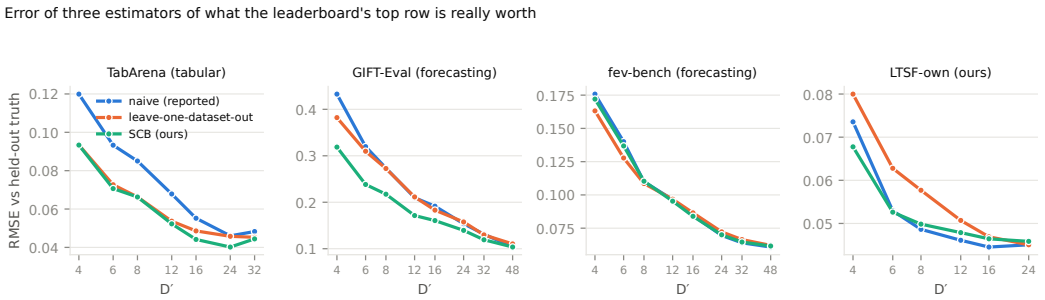

Figure 2: Estimator RMSE against disjoint-holdout truth across the full range of selection suite sizes D' (strong reference, log-ratio scale), complementing Table 5. SCB has the lowest RMSE where selection bias is material (GIFT-Eval clearly, TabArena marginally) and tracks the naive estimator where it is not (fev-bench, LTSF-own).

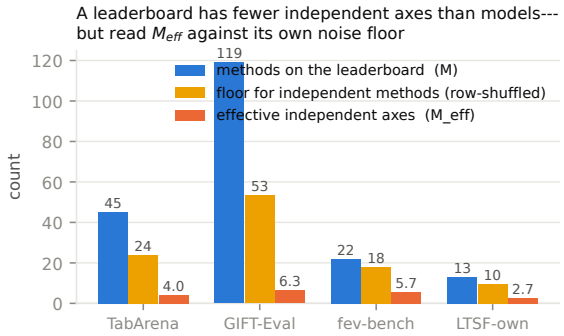

Figure 3: Nominal number of leaderboard entries M versus the effective number of independent axes M_{eff} (participation ratio of the method-correlation spectrum, log-ratio scale) for the four matrices.

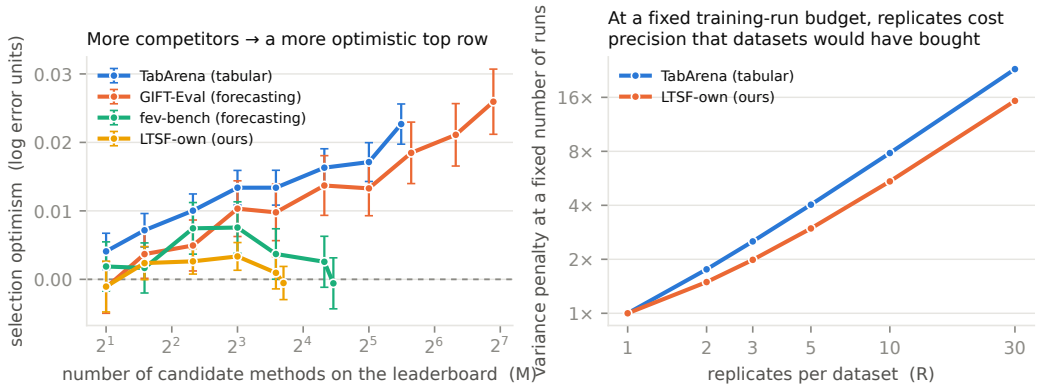

Figure 4: **Left: the winner's curse grows with the number of candidates—on packed leaderboards.** Measured optimism of the top row's own score (select on half the datasets, score on the other half) as the method axis is subsampled to M candidates. On fev-bench the curse is non-monotone, being significantly positive at $M = 5$–8 ($+0.008$, $z \approx 3.9$) and back to zero at the full $M = 22$, because the last methods added are the ones that separate the champion, rather than absent at every M. Error bars: ± 1.96 Monte-Carlo SE. **Right: replicates are expensive noise reduction.** Measured variance-component ratios imply that spending a fixed training-run budget on R replicates per dataset instead of $R\times$ more datasets inflates the variance of any cross-dataset mean contrast by $(\sigma_\gamma^2 R + \sigma_e^2)/(\sigma_\gamma^2 + \sigma_e^2)$.

Table 5: Estimator error against disjoint-holdout truth on real matrices (strong reference, log-ratio scale). Bias ($\pm$ Monte-Carlo SE) and RMSE of the reported value (NAIVE), leave-one-dataset-out (LOO), SCB, conditional SCB (CND), Tibshirani & Tibshirani (2009) (TT), and empirical-Bayes shrinkage of the column (EB, Efron, 2011) over 600 subsample draws at each D' (Monte-Carlo SE on every bias is $\leq .011$, $\leq .005$ outside GIFT-Eval; per-cell SEs are in `results/claims.json`); all six are scored on the *same* draws in one run; bold marks the lowest RMSE in the row, which EB often takes at the bias cost §5 prices, and every value here and in §5 is rendered from `results/claims.json`. Where selection bias is large (GIFT-Eval) SCB removes most of it at materially lower RMSE than LOO; on TabArena a significant residual remains ($z = +5.8$ at D'=8), so our corrected TabArena margin in Table 2 should be read as an *upper* bound on what transfers. TT is the informative failure: its within-unit minimum over M methods overshoots badly once the suite is heterogeneous. On champion matrices (bottom rows) every estimator is close because there is little to correct.

		Bias						RMSE					
Matrix	D'	NAIVE	LOO	SCB	CND	TT	EB	NAIVE	LOO	SCB	CND	TT	EB
TabArena	8	+.044	−.001	+.015	+.018	−.003	+.028	.085	.066	.066	.064	.067	**.056**
	24	+.020	−.002	+.004	+.009	−.045	+.016	.046	.046	**.040**	.044	.061	.042
GIFT-Eval	8	+.167	−.017	−.013	−.007	−.199	−.054	.274	.273	.217	.221	.311	**.154**
	24	+.069	+.004	+.002	+.014	−.359	−.006	.154	.158	.139	.141	.387	**.114**
fev-bench	8	+.030	−.005	−.006	−.003	−.032	−.049	.110	**.109**	.110	.110	.111	.111
	24	+.007	−.001	−.003	+.001	−.067	−.021	**.070**	.072	.070	.071	.097	.071
LTSF-own	8	+.007	−.005	−.004	−.001	−.034	+.004	.049	.058	.050	.050	.062	**.045**
	24	−.002	−.002	−.003	−.003	−.043	−.003	.045	.045	.046	.046	.063	**.045**

A.3 $P(\#1$ HOLDS$)$ AT EVERY SUITE SIZE, AGAINST THE OMNISCIENT CEILING

Table 7 gives the probability that the method selected on D' units is also the best of the disjoint complement, at all eight sizes and for all four matrices, each with the ceiling a μ-omniscient selector attains on the same complement. Two things are visible. (i) The measure is not monotone in D' on TabArena $(0.34, 0.37, 0.40, 0.35, 0.29, 0.26, 0.28)$: its complement shrinks from 47 to 19 units as D' grows, so the target it is scored against gets noisier, and the ceiling drops in step ($0.98 \to 0.58$).

Dividing through, the ceiling-normalized rate is $0.35, 0.38, 0.44, 0.40, 0.40, 0.39, 0.48$, showing no trend beyond Monte-Carlo noise, i.e. TabArena's packed top never resolves at any suite size we can audit. (ii) On the other three matrices the ceiling stays near 1 over most of the range and the raw rate rises with D', as one would expect (the $0.59 \to 0.55$ step on GIFT-Eval is within Monte-Carlo error). Reporting $P(\#1)$ without its ceiling would confound the selector with the yardstick.

The ceiling in Table 7 is built from the *full-suite* winner, which overlaps the scoring complement (at D'=8 on TabArena the complement is 43 of the 51 units the oracle sees), so it is an upper bound on the ceiling rather than the ceiling. A cleaner version splits the units three ways: A (selection, D' units), C (proxy), B (scoring), with C and B equal halves of the complement, so the proxy never sees the set it is scored on. Both arms are then scored on the same B, which is half the size of Table 7's complement, so both numbers are lower there, and only the comparison within a row is meaningful.

The three-way split introduces an artifact of its own. When A, C and B *partition* a fixed suite, whatever the proxy takes into C it removes from B: a method that looks good on C has been depleted on the very set it is scored against, and the depletion grows with $|C|$. At 40,000 draws on TabArena the 21-unit proxy scores 0.219 against the 8-unit selector's 0.228: the proxy is *worse*, which is impossible for an unbiased comparison. Drawing A, C and B independently with replacement (the superpopulation reading that licenses "on a new dataset" language) removes it: 0.275 at 8 units against 0.334 at 21. We run the same comparison on all four matrices at sizes their D allows, and release it. We therefore report the independent-draw version in Table 1 and in §4. Under it TabArena's top does resolve with more data, but slowly: enlarging the selection set $2.6\times$ raises the hit rate from 0.27 to 0.33, so a suite $2.6\times$ larger than TabArena's is still wrong about the identity of the best method two times in three. The same enlargement buys $0.25 \to 0.34$ on GIFT-Eval and $0.49 \to 0.66$ on fev-bench, and $0.82 \to 0.90$ on our own grid, whose D=36 allows only $8 \to 14$.

Table 6: Three-way split versus independent draws, at two selection-set sizes. $P(\#1)$ is the selector's hit rate on B; "proxy" is a selector given the larger slice C and scored on the same B; "full-suite" is the oracle of Table 7, which overlaps B and so overstates the ceiling.

	$D' = 8$			$D' = 24$		
Matrix	$P(\#1)$	proxy	full-suite	$P(\#1)$	proxy	full-suite
TabArena	0.23	0.21	0.57	0.25	0.21	0.50
GIFT-Eval	0.27	0.57	0.79	0.39	0.45	0.71
fev-bench	0.53	0.93	0.96	0.79	0.85	0.93
LTSF-own	0.87	0.96	0.98	0.84	0.69	0.84

The full-suite oracle overstates the ceiling by up to a factor of three on the packed matrix and by little or nothing on the champion matrices, so a ceiling read off the full suite is not a ceiling where it matters. The three-way split is the better instrument, but it carries an artifact of its own, described above; the numbers reported in §4 come from the independent-draw version, which has neither problem.

Table 7: $P(\#1 \text{ holds})$ with, in brackets, the ceiling attained by a selector that knows the true means (strong reference, log-ratio scale, 600 draws per cell; "–" where D' leaves fewer than eight complement units).

Matrix	D'=4	6	8	12	16	24	32	48
TabArena	0.34 [0.98]	0.37 [0.97]	0.40 [0.92]	0.35 [0.88]	0.29 [0.73]	0.26 [0.66]	0.28 [0.58]	–
GIFT-Eval	0.24 [1.00]	0.30 [1.00]	0.36 [1.00]	0.41 [1.00]	0.44 [0.99]	0.59 [0.97]	0.55 [0.91]	0.65 [0.80]
fev-bench	0.42 [1.00]	0.51 [1.00]	0.60 [1.00]	0.67 [1.00]	0.81 [1.00]	0.84 [1.00]	0.93 [1.00]	0.95 [0.97]
LTSF-own	0.74 [1.00]	0.83 [1.00]	0.92 [1.00]	0.96 [1.00]	0.99 [1.00]	0.98 [0.98]	–	–

A.4 What the complete-case filter costs

Every matrix is reduced to its complete-case rectangle before anything is measured, and that filter is outcome-dependent in the same way the paper's subject is: a model leaves because it failed to return

somewhere, a task leaves because some model failed on it. Only fev-bench is materially affected, losing 9 of 100 tasks, so Table 8 cuts that rectangle the other way. For $k = 0, \dots, 4$ it drops the k non-reference models whose removal recovers the most tasks (ties broken alphabetically) and re-runs the retention audit on what remains. The reference column is never a candidate: removing it would change the estimand rather than the filter, which is the defect that made an earlier version of this check uninterpretable. Retention rises monotonically with k, from 0.872 on the shipped 22×91 rectangle to 0.887 once three models are gone, and Chronos-2 is the full-suite winner at every rung. The filter therefore costs at most 0.015 of retention, in the direction that makes fev-bench look *less* like a champion matrix, and no claim in the paper turns on it. The script is `code/casewise_sensitivity.py`.

Table 8: Trading models for tasks on fev-bench. At each k, the drop set is the one leaving the most tasks standing; retention is at D'=8 over $40{,}000$ draws against the strong reference (Stat. Ensemble), which is never dropped. The five incomplete columns are AutoARIMA (4 missing cells), Stat. Ensemble (4), AutoETS (3), DeepAR (3) and TabPFN-TS (1).

k	models dropped	rectangle	retention at D'=8	full-suite winner
0	—	22×91	0.8723 (± 0.0006)	Chronos-2
1	AutoETS	21×93	0.8793 (± 0.0006)	Chronos-2
2	AutoETS, DeepAR	20×95	0.8854 (± 0.0006)	Chronos-2
3	AutoETS, DeepAR, TabPFN-TS	19×96	0.8871 (± 0.0006)	Chronos-2
4	AutoARIMA, AutoETS, DeepAR, TabPFN-TS	18×96	0.8871 (± 0.0006)	Chronos-2

A.5 The cluster-blocked variant, exactly

Let the D units be partitioned into families $f = 1, \dots, G$ with index sets I_f of sizes m_f ($\sum_f m_f = D$), and write $\bar{Z}[m, f] = \frac{1}{m_f} \sum_{d \in I_f} Z[m, d]$ for the family means. The variant changes two lines of the i.i.d. estimator of §5 and nothing else. (i) The per-method sampling variance becomes the *unweighted* between-family variance of the family means, each family counting once whatever its size, $\hat{v}_m = \frac{1}{G(G-1)} \sum_f \left(\bar{Z}[m, f] - \frac{1}{G} \sum_{f'} \bar{Z}[m, f'] \right)^2$. The quantities $\hat{\tau}^2$, c and $\tilde{\mu}$ then follow the i.i.d. formulas with this $\hat{v}_m$, while $\hat{\mu}_m$ stays the unit-weighted column mean. (ii) Each bootstrap draw picks G families with replacement, $f_1, \dots, f_G$, and concatenates their columns, so the draw holds a variable number $D_b^* = \sum_j m_{f_j}$ of units. The resampled column is the unit-weighted mean over that concatenation, $\hat{\mu}_b^*[m] = \tilde{\mu}_m + \frac{1}{D_b^*} \sum_j \sum_{d \in I_{f_j}} \mathrm{resid}[m, d]$, so a large family weighs in proportion to its size, as it does in the leaderboard's own column mean, and no per-draw rescaling is needed. The winner k_b, β_b and A_{SCB} are recorded exactly as before. With every family a singleton the two lines reduce to the i.i.d. estimator, which is why TabArena's two columns agree to 0.0002 in §5. The released implementation (`code/blocked_audit.py`, `scb_cluster`) uses $B = 8{,}000$ draws for the full-suite corrections of Table 2 and $B = 400$ inside the 600-draw blocked audit. The interval of §7 is validated for the i.i.d. estimator (App. A.15); a cluster-blocked interval would carry the design effect through the same family resampling, and must then not be scaled by $\sqrt{D/D_{\mathrm{eff}}}$ a second time.

A.6 Synthetic validation details

The generative model draws $\mu_m \sim \mathcal{N}(0, \tau^2)$ ($\mu_{m_0} = 0$), a shared dataset factor $f_d \sim \mathcal{N}(0, 1)$, and idiosyncratic noise, with $\gamma_{md} = \sigma_g(\sqrt{\rho_c} f_d + \sqrt{1 - \rho_c}\, \varepsilon_{md})$. The 19 cells sweep $D \in \{8, 16, 32, 64, 128\}$ at $(M, \rho_c) = (30, 0.3)$; $M \in \{3, 6, 12, 25, 50, 100\}$ at $(D, \rho_c) = (24, 0.3)$; $\rho_c \in \{0, 0.2, 0.4, 0.6, 0.8, 0.95\}$ at $(M, D) = (30, 24)$; and the degenerate regimes $\tau = 0$ (no true differences) and $(\tau, \sigma_g) = (0.6, 0.1)$ (well separated). Each cell runs 400 simulations with $B = 800$ bootstrap draws. Verdicts: naive optimistic at $z > 3$ in $17/19$; SCB $|z| \leq 2$ in $18/19$ and $|z| \leq 4$ in $19/19$, the two largest residuals being $|z| = 1.99$ and $|z| = 3.18$; RMSE(SCB) $<$ RMSE(naive) in $17/19$, where the two losses are ρ_c=0.8 ($+4.2\%$) and ρ_c=0.95 ($+7.0\%$), the extreme-correlation cells where SCB over-corrects, not ties; EVT-with-M_{eff} closer to the measured optimism than EVT-with-M in $19/19$ (median absolute error 0.092 vs 0.155).

A.7 LTSF GRID PROTOCOL AND WHAT THE MINIMAL-SCALE REDUCTIONS COST

Standard splits (ETT: 12/4/4 months; others: 70/10/20), channel-independent windows, train-statistics z-normalization, MSE on the normalized scale. Methods: Naive, Mean, Drift, SeasonalNaive, Linear at three lookbacks, NLinear, RLinear, DLinear, ridge with validation-selected penalty, and two-layer MLPs at two lookbacks (Adam, 256 hidden units). Because ILI is only 966 steps long it cannot carry the standard grid, so methods are named by *role* rather than by absolute lookback: Linear-S/M/L and MLP-S/M are that architecture at the short, medium and long lookback of the dataset's own grid, which is $(96, 336, 512)$ everywhere except ILI, where it is $(24, 36, 60)$ with horizons $(24, 36, 48, 60)$ instead of $(96, 192, 336, 720)$. Linear-M is thus L=336 on 32 of the 36 cells and L=36 on ILI's four; the canonicalization is in `code/ltsf_matrix.py`. Replicates are 8 contiguous groups of rolling-origin forecast starts partitioned over the test region; adjacent groups' target windows overlap by up to H steps, understating σ_e (conservative for the budget identity's constants). Minimal-scale reductions (channel caps of 24 on Electricity and Traffic, window strides, a 120k training-window cap) are documented in the repository together with the flag that removes them.

Because the capped datasets supply 2 of the 6 families behind the blocked result, we certify them directly rather than by analogy with the 7-channel ETTh1 anchor (`code/anchor_check.py`, `code/cap_sensitivity.py`). At *full* channel width (321 and 862 channels, stride 1, 5.8M and 10.2M training windows) our L=336 linear map at H=96 reaches MSE 0.140 on Electricity and 0.410 on Traffic, against the published 0.140 and 0.410 of Zeng et al. (2023); ETTh1 gives 0.370 against their 0.375. The cap moves the *level* (capped: 0.206 and 0.321, since the first 24 channels are not a representative subset). What the audit consumes is the log-ratio contrast, so the question is how much *that* moves: re-running the whole 13-method zoo at H=96 under an identical window budget, contrasts among the trained methods shift by at most 0.13 log units on Electricity and 0.015 on Traffic, while the untrained baselines move up to 0.54; orderings correlate at Spearman 0.98 and 0.96 with the same winner in both. The top-of-table contrast this paper audits is thus stable under the cap, but the weak-baseline entries, which drive the $\mathrm{ICC}_{\mathrm{fam}}$ behind D_{eff}, are not, so this grid's D_{eff} belongs to the capped version of it. The residual risk is that a cap could still perturb a near-tie inside a single cell; the family-blocked result's instability, which we report rather than smooth over, is the larger caveat.

A.8 THE TWO CLUSTER-SIZE CONSTANTS IN D_{eff}

Unequal family sizes $m_1, \ldots, m_k$ ($\sum_f m_f = D$) require *two different* averages, and conflating them is an easy error to make in either direction. (i) Kish's design effect asks how much the variance of the suite mean inflates under clustering. Writing $\mathrm{Var}(\bar{z}) = \frac{\sigma^2}{D}\big(1 + (\bar{m}_{\mathrm{K}} - 1)\rho\big)$ with $\bar{m}_{\mathrm{K}} = \sum_f m_f^2 / \sum_f m_f$, the relevant average is size-biased: it is the expected size of the family containing a randomly drawn *cell*, not a randomly drawn family. Since $\bar{m}_{\mathrm{K}} \geq D/k$ with equality only for equal families, unequal sizes can only *lower* D_{eff} relative to the balanced case, a useful sanity check. (ii) The one-way ANOVA estimator of ρ itself uses a different constant, $m_0 = \big(D^2 - \sum_f m_f^2\big)/\big(D(k-1)\big)$, which makes $\mathbb{E}[\mathrm{MSB}] = \sigma_w^2 + m_0\sigma_b^2$ hold exactly under unequal sizes; we use $\hat{\rho} = (\mathrm{MSB} - \mathrm{MSW})/(\mathrm{MSB} + (m_0 - 1)\mathrm{MSW})$ as the estimate of $\mathrm{ICC}_{\mathrm{fam}}$ and then plug $\hat{\rho}$ into (i). For our 36-cell grid ($k = 6$ families, sizes $16, 4, 4, 4, 4, 4$) these are $\bar{m}_{\mathrm{K}} = 336/36 = 9.33$ and $m_0 = 5.33$; the balanced-formula value would be 6. We verified the choice by direct simulation: drawing 10^5 exchangeable clustered tables with $\rho = 0.704$ at the observed family sizes gives an empirical variance ratio of 6.886 against Var under independence, versus $1 + (\bar{m}_{\mathrm{K}} - 1)\rho = 6.867$, $1 + (m_0 - 1)\rho = 4.051$, and $1 + (D/k - 1)\rho = 4.520$. The reproduction script is `code/deff_sim_check.py`. Family assignment is by source series (all four ETT variants $\times$ four horizons form one family); horizons of the same series are treated as siblings because they share the same underlying data, which is the conservative choice.

A.9 THE BREADTH-VERSUS-REPLICATION BUDGET IN FULL

The replicated axes decompose each paired contrast $\Delta_{dr} = \delta + \gamma_d + e_{dr}$ into interaction σ_γ^2 and replicate noise σ_e^2. The interaction dominates in both domains: median $\mathrm{ICC}_{\mathrm{rep}}$ 0.75 over TabArena's 44 contrasts (nine folds on all 51 datasets, thirty on 17; $\bar{R}$=16; IQR 0.63–0.85) and 0.66 on LTSF-own, whose 8 rolling-origin blocks overlap by up to H steps and so understate σ_e; since the penalty

is $1 + \mathrm{ICC}_{\mathrm{rep}}(R-1)$ that inflates the LTSF numbers, which is *anti*-conservative, so the headline leans on TabArena, whose folds resample the training set rather than the test region.

Breadth dominates replication (Cochran, 1977; Raudenbush, 1997): under a budget of $C = D \cdot R$ training runs, $\mathrm{Var}(\hat{\delta}) = (\sigma_\gamma^2 R + \sigma_e^2)/C$ increases in R, so the penalty of R replicates relative to $R{=}1$ on $R\times$ more datasets is $(\sigma_\gamma^2 R + \sigma_e^2)/(\sigma_\gamma^2 + \sigma_e^2) = 1 + \mathrm{ICC}_{\mathrm{rep}}(R-1)$. The identity is classical (Mustahsan et al., 2025); we contribute the constants for domains where both axes cost a training run, so $R{=}1$ binds (for LLM evals resampling is cheap and the advice inverts, Miller, 2024). From median variance components (implied $\mathrm{ICC}_{\mathrm{rep}}$ 0.76 and 0.49; per-method medians would give the *larger* $2.3\times$/$7.0\times$, so we report the smaller) the penalty is $2.5\times$ at $R{=}3$ and $7.8\times$ at $R{=}10$ on TabArena, $2.0\times$/$5.4\times$ on LTSF-own (Figure 4): the three-to-ten-seeds convention buys noise reduction at several times the price of breadth, which would both shrink variance and starve the winner's curse. Replicates keep two uses we do not dispute: identifying σ_e, and guarding against single-split accidents.

A.10 Why the closed form overshoots: heterogeneity or correlation?

The optimism of a selected maximum has two sources that the $\sqrt{2\log M}$ envelope conflates: candidates are correlated, and candidates differ in true effect (so the arg-min is partly signal). We separate them non-parametrically on each real matrix. Row-shuffling, i.e. permuting each method's row of per-dataset scores independently, destroys all cross-method correlation while preserving every method's marginal; row-centering ($Z \to Z - \overline{Z[m,\cdot]}$) destroys the true effect differences while preserving the correlation. Measuring the split-half optimism of the winner in all four combinations ($4{,}000$ draws each) gives:

Table 9: What drives the closed form's overshoot. Measured split-half optimism with the true effect differences present or nulled, crossed with the real cross-method correlation ("real corr") or with independence ("indep.", row-shuffled). Each factor is read at both corners, because the two interact.

Matrix	real μ real corr	null μ real corr	real μ indep.	null μ indep.	correlation alone	heterogeneity alone
TabArena	.0191	.1347	.0598	.1936	$1.44\times$	$7.06\times$
GIFT-Eval	.0273	.1984	.0702	.2357	$1.19\times$	$7.28\times$
fev-bench	$-.0001$	.0863	.0544	.1247	$1.44\times$	—
LTSF-own	.0003	.1499	.0005	.1920	$1.28\times$	—

"Correlation alone" is the null-μ column ratio (independent / real correlation); "heterogeneity alone" is null-μ over real-μ at the real correlation. On the two champion matrices the real-μ optimism is indistinguishable from zero, so the ratio is undefined there. Table 9 reads as follows. A closed form that assumes equal true effects is wrong by a factor of about seven, while one that assumes independence is wrong by 1.1–$1.4\times$. The emphasis belongs on heterogeneity, not on correlation.

M_{eff} needs the same treatment. A participation ratio computed from a sample correlation matrix on D units is biased downward even for genuinely independent methods; row-shuffling gives that floor. Measured against it, TabArena is 4.02 against a floor of 23.94 ± 0.64, GIFT-Eval 6.35 against $53.35{\pm}0.51$, fev-bench 5.70 against $17.84{\pm}0.32$, LTSF-own 2.69 against $9.53{\pm}0.34$. The narrowing is large and real (3.1–$8.4\times$, and 6.0–$8.4\times$ on the two large matrices), but it is not the M/M_{eff} ratio of 4–$19\times$ a reader would compute from the raw numbers, and Figure 3 should be read with the floor in mind.

A.11 The family-blocked LTSF cell, enumerated

With six families, holding in two of them admits exactly $\binom{6}{2} = 15$ distinct realizations, so this cell can be enumerated rather than sampled and its Monte-Carlo standard error is beside the point. The exact retention is -0.347 (against -0.303 from 600 random draws). What varies is not the sampling: the 15 configurations range from -7.3 (Exchange+Weather, where the mechanism picks the short-lookback linear model) to $+5.5$ (Electricity+Exchange), and the mechanism selects five different winners across them (NLinear in 8, Linear-L in 4, and Linear-S, SeasonalNaive and Ridge once each). Leave-one-configuration-out gives a jackknife SE of 0.52; leave-one-*family*-out is more interpretable and more damning: dropping ETT gives -0.89, Electricity -0.66, Traffic -0.46, Weather -0.29,

ILI -0.02, Exchange $+0.28$. A quantity that moves by more than one unit of retention when a single source series is removed is not identified at six families. We report it because the *direction* is consistent: blocking always costs retention here, and the three-family cell is $+0.39$ against an i.i.d. 0.85, but the magnitude is a warning about unit definitions, not an estimate, and no conclusion in this paper rests on it.

A.12 Effect-distribution robustness of the plug-in

The synthetic grid draws $\mu_m \sim \mathcal{N}(0, \tau^2)$ in every cell, and the variance-matched plug-in matches a second moment, so the grid cannot see a mis-specified *shape*. Holding the grid's reference cell fixed (M=30, D=24, σ_g=0.5, ρ_c=0.3, 1,200 simulations) and replacing only the effect vector:

Table 10: Effect-distribution robustness of the variance-matched plug-in: naive and SCB bias, RMSE, and the shrink factor c under four synthetic effect distributions at M=30, D=24.

true μ	naive bias	SCB bias	RMSE naive	RMSE SCB	shrink c
Gaussian, the grid's own prior	$+.055$	$-.034$	.121	.127	0.72
one champion + tied pack	$+.011$	$-.118$	.106	.178	0.31
two near-champions	$+.056$	$-.064$	.111	.135	0.43
point mass (no differences)	$+.173$	$+.006$	.187	.071	0.03

These are *conditional* biases at a fixed μ; averaged over the Gaussian prior SCB is calibrated (App. A.6), and the null cell shows it removing a $+.173$ artifact almost exactly. But conditional on one genuine champion, $\hat{\tau}^2 = \mathrm{Var}_m(\hat{\mu}) - \bar{\hat{v}}$ subtracts sampling variance from a spread that is mostly that champion's real lead, c collapses, and the bootstrap "finds" optimism that is not there.

What c flags, and what it cannot. Two sweeps in the same cell (Table 11; `code/shrink_threshold.py`) put a footing under the $c < 0.8$ line quoted in §5. Moving one entry's lead over a tied pack (sd 0.02) from 0 to 0.6 log units, c rises from 0.04 to 0.66 while SCB's conditional bias runs from the null regime's $+0.007$ through an over-correction that peaks at -0.114 when the champion stands about 1.5 paired standard errors clear of the pack (lead 0.35, $c = 0.29$, tied count 5.9) and recedes to -0.025 once the champion is unmistakable (3.6 standard errors, $c = 0.66$). The naive report is the better estimate from lead 0.25 upward, SCB below it. Drawing Gaussian effects afresh with spread τ from 0 to 0.4 instead, c covers the same range (0.03 to 0.97) while SCB stays within ± 0.011 of unbiased throughout. A low c by itself is therefore not a failure signal. It records how much the plug-in was shrunk, and heavy shrinkage is the right response when there is nothing to find; what it certifies is how much room a shape error has. On the four real matrices c is 0.984, 0.970, 0.983 and 0.995: the applied shrinkage is within 3% of the identity, the plug-in is nearly the observed column, and no effect shape can move the correction far. Below $c \approx 0.8$ the correction is shape-sensitive. Neither c nor the tied count separates the harmful champion shape from the harmless Gaussian one reliably: lead 0.20 and $\tau = 0.05$ print similar c and tied counts, with SCB biases of -0.058 and -0.011. Discarding A_{SCB} there would be the wrong response, since the naive report is the worse of the two in most low-c cells of both sweeps. We therefore ask for both numbers. In ten of the sixteen sweep cells A_{NAIVE} and A_{SCB} bracket the truth, and in the other six one of them is within 0.008 of it.

Table 11: What the shrink diagnostic c flags. Same cell as Table 10 (M=30, D=24, σ_g=0.5, ρ_c=0.3; 1,200 simulations per row, bias SE ≤ 0.005). Top: one champion leading a tied pack by the stated amount. Bottom: Gaussian effects of spread τ, redrawn each simulation. "tied" is the mean tied count (winner included); z the mean separation of the observed #1 from #2 in paired standard errors.

true μ	shrink c	tied	z	naive bias	SCB bias
champion + pack, lead 0.00	0.04	20.3	0.3	+0.173	+0.007
champion + pack, lead 0.05	0.04	20.2	0.3	+0.169	+0.003
champion + pack, lead 0.10	0.05	19.4	0.3	+0.160	−0.005
champion + pack, lead 0.15	0.06	18.0	0.4	+0.144	−0.021
champion + pack, lead 0.20	0.11	15.5	0.6	+0.103	−0.058
champion + pack, lead 0.25	0.16	11.7	0.8	+0.070	−0.084
champion + pack, lead 0.30	0.21	8.5	1.1	+0.038	−0.110
champion + pack, lead 0.35	0.29	5.9	1.5	+0.017	−0.114
champion + pack, lead 0.45	0.47	2.1	2.3	+0.005	−0.081
champion + pack, lead 0.60	0.66	1.1	3.6	−0.002	−0.025
Gaussian, $\tau = 0.00$	0.03	21.0	0.3	+0.175	+0.008
Gaussian, $\tau = 0.05$	0.15	17.7	0.4	+0.148	−0.011
Gaussian, $\tau = 0.10$	0.57	11.7	0.5	+0.114	−0.009
Gaussian, $\tau = 0.15$	0.78	7.9	0.6	+0.088	−0.001
Gaussian, $\tau = 0.25$	0.92	4.5	1.0	+0.059	+0.003
Gaussian, $\tau = 0.40$	0.97	2.7	1.6	+0.034	+0.000

A.13 The SimpleTM transcription

Table 1 of Chen et al. (2025) was extracted from the camera-ready PDF with layout-preserving text extraction and verified cell-by-cell; both MSE and MAE blocks are analyzed (11 methods: SimpleTM, TimeMixer, iTransformer, CrossGNN, RLinear, PatchTST, Crossformer, TiDE, TimesNet, DLinear, SCINet; 8 datasets; horizons averaged; lookback 96). MAE results mirror MSE: margin over runner-up $3.03\% \to 3.02\%$, $M_{\text{eff}} = 2.6$, $D_{\text{eff}} = 6.7$. We emphasize that this re-analysis takes the published numbers at face value; it corrects suite-level selection, not any upstream asymmetry in how those numbers were produced.

A.14 Where the residual on the packed matrix comes from

SCB's residual on TabArena is positive at every audited size, and §5 reports the corrected margin as an upper bound because of it. Two explanations call for opposite remedies, so we separate them by construction. The plug-in could be the wrong *shape*: it is an affine deconvolution, matching the effect distribution's second moment and nothing else. Or the plug-in could simply be *noisy*: $\hat{\tau}^2$ is estimated from the same D' units the estimator is given, and at D'=8 that estimate is poor and floored at zero. Table 12 adds four arms to the audit, all scored on the same draws against the same disjoint-holdout truth. Two are estimators a practitioner could run: a two-component Gaussian mixture and a Kiefer–Wolfowitz NPMLE deconvolution (Kiefer & Wolfowitz, 1956), fitted on a grid by the EM iteration of Jiang & Zhang (2009), each supplying a plug-in whose empirical distribution matches the fitted G rather than only its variance. Two are oracles that read the full matrix, which the estimator may not see, and exist only to bound what any plug-in could achieve: one takes τ^2 from the full matrix, the other uses the full-matrix mean vector itself.

The answer is unambiguous in both directions. Oracle-τ^2 leaves TabArena's residual exactly where it was ($+0.018 \to +0.018$ at D'=8), so estimation noise in $\hat{\tau}^2$ explains none of it; oracle-μ removes it ($+0.018 \to -0.002$, $z = -0.7$), so the shape of the plug-in explains all of it. That is the good news for future work, and Table 12 bounds the prize: on TabArena the oracle is both unbiased and the lowest-RMSE arm of the six, and on the other three matrices it is within 1% of the best. The bad news is that neither richer plug-in collects it. Both the mixture and the NPMLE overshoot on *every* matrix, by $z = -6.0$ to -9.7, turning a positive residual of $+0.018$ into a negative one of -0.027 and a champion's honest -0.003 into -0.039. The mechanism is the one §5 already names for James–Stein shrinkage: a deconvolution fitted for per-element accuracy attributes too much of the observed spread to noise, returns a G that is too narrow, and a too-narrow plug-in leaves the bootstrap winner too much room to be lucky, so the estimated correction is too large. Nonparametric

deconvolution does not escape that trap at $D' \approx 8$; it falls into it harder. Among the four plug-ins we tried, variance matching is the only one whose residual is within Monte-Carlo error on three of the four matrices, which is why it remains the shipped default and why the TabArena number stays an upper bound.

Table 12: Bias against disjoint-holdout truth for six plug-in choices, all scored on the same 600 draws. Top block: TabArena at three suite sizes. Bottom block: all four matrices at D'=8. Oracle arms read the full matrix and are not available to a practitioner; they bound what a perfect plug-in could do.

TabArena, plug-in	D'=8	D'=16	D'=24
NAIVE	$+0.0467$ ($z=+15.3$)	$+0.0268$ ($z=+13.7$)	$+0.0194$ ($z=+11.5$)
SCB (varmatch)	$+0.0183$ ($z=+6.6$)	$+0.0080$ ($z=+4.4$)	$+0.0032$ ($z=+2.0$)
SCB, 2-component mixture	-0.0257 ($z=-8.8$)	$+0.0034$ ($z=+2.0$)	$+0.0086$ ($z=+5.5$)
SCB, NPMLE deconvolution	-0.0272 ($z=-9.7$)	-0.0052 ($z=-3.2$)	-0.0035 ($z=-2.4$)
SCB, oracle τ^2	$+0.0181$ ($z=+6.7$)	$+0.0086$ ($z=+4.8$)	$+0.0034$ ($z=+2.1$)
SCB, oracle μ	-0.0018 ($z=-0.7$)	$+0.0009$ ($z=+0.5$)	$+0.0008$ ($z=+0.5$)

Bias at D'=8	NAIVE	SCB	mixture	NPMLE	oracle τ^2	oracle μ
TabArena	$+0.0467$	$+0.0183$	-0.0257	-0.0272	$+0.0181$	-0.0018
GIFT-Eval	$+0.1587$	-0.0209	-0.0549	-0.0722	$+0.0313$	$+0.0111$
fev-bench	$+0.0331$	-0.0029	-0.0299	-0.0390	$+0.0049$	$+0.0008$
LTSF-own	$+0.0103$	-0.0006	-0.0173	-0.0171	-0.0000	-0.0032

A.15 Does the interval we recommend reporting actually cover?

§7 asks benchmark authors to print A_{SCB} with a bootstrap interval, which is only worth asking if the interval is calibrated. Getting there took two corrections, and both are instructive.

The first is studentization: each bootstrap draw's error is divided by that draw's own standard error before the quantiles are taken, and the quantiles are rescaled by the observed standard error. At $D' \approx 8$ that is worth five to nine coverage points over the plain percentile interval (Table 13). The second is symmetry. The textbook basic-bootstrap inversion is equal-tailed, and here it misbehaves: the bootstrap error distribution is skewed, so the equal-tailed interval's *midpoint* lands well above its own point estimate—on TabArena at D'=8, about $+0.030$ above, which is more than the entire correction SCB had just applied, so that interval quietly gives the correction back and its midpoint is a worse estimate of the held-out truth than the uncorrected report is. Symmetric percentile-t endpoints, $A_{\text{SCB}} \pm q_{0.90}(|t|) \cdot \widehat{\text{se}}$, are centred on the estimate by construction, are the more accurate construction in theory (Hall, 1988), and cover better on all four real matrices. They are what we recommend and what `code/selection.py` returns.

The textbook remedy for a skewed bootstrap distribution is the bias-corrected and accelerated percentile interval (BCa; Efron, 1987), so Table 13 carries it as a fifth arm, built on the same single-level replicates with its acceleration from a jackknife over units (`scb_bca_interval`). It makes matters worse, not better: 0.675 and 0.758 where the symmetric-t interval covers 0.865 and 0.905, below the symmetric-t interval's raw coverage on all twelve real cells and at or below every other SCB arm's on nine of them. The reason is the direction of its correction. The acceleration is negligible here ($|a| \leq 0.04$ at D'=8 on all four matrices), so BCa reduces to a median-bias shift, and that shift is positive ($z_0 \approx +0.2$ to $+0.4$): the replicates sit below the point estimate, so BCa moves both endpoints *up*, towards the uncorrected report, and its midpoint lands 0.02 to 0.09 above A_{SCB}, on TabArena about the size of the whole correction. This is the same defect as the equal-tailed interval's, reached from the other direction. A construction that reads its centre off the skewed replicate distribution gives the correction back, whereas endpoints centred on the estimate cannot.

Table 13 measures coverage where the truth is known exactly, on the generative model of App. A.6, in two regimes. When the methods genuinely differ (τ=0.25) the interval covers 0.865 against a nominal 0.90 and the naive t interval a paper would print today covers 0.843: neither is exact, and the naive one buys its coverage with 3% more width. When the top is *packed* (τ=0.05, M=45), the regime this paper is about where the reported margin is mostly selection, the interval covers 0.905 and the naive interval covers 0.607 while being 25% wider. An interval that ignores the selection

event does not fail by being too narrow; it fails by being centred in the wrong place, and width does not repair that where the bias is a multiple of the standard error.

On real matrices the truth must be proxied by the complement mean, which is itself an average of $|B|$ units, so a perfectly calibrated interval misses it some of the time; we report raw coverage and a yardstick-adjusted figure that adds the complement's own standard error in quadrature. Raw coverage for the recommended interval is 0.69–0.93 and the adjusted figure 0.89–1.00. We draw no comparative conclusion from these cells, and say why. The complement's standard error is roughly half the interval's half-width at these sizes, so it swamps the 0.02–0.03 of bias that separates the arms: the last column gives the naive interval SCB's width, which is the comparison that would isolate centring, and it comes out at or above SCB's own coverage on every matrix. That is what a yardstick with this much noise looks like, not evidence that the naive centring is better. Table 13's upper block, where the target carries no noise at all, is the one that separates them, and it is the block the recommendation rests on.

Table 13: Coverage of a nominal-90% interval. Upper block: the synthetic model of App. A.6, where μ is drawn so the target needs no proxy; mean width in parentheses. *Packed* is the regime this paper is about. Lower block: real matrices against the disjoint complement, printed as raw / yardstick-adjusted, the adjustment adding the complement's own standard error in quadrature. "NAIVE" is the ordinary t interval around the winning column; "NAIVE@w" gives it SCB's width, isolating centring from width. "BCa" is Efron's bias-corrected and accelerated percentile interval on the same replicates, with the acceleration from a jackknife over units.

Synthetic, exact truth	SCB symmetric-t	SCB equal-tail-t	SCB percentile	SCB BCa	NAIVE
$\tau = 0.25$, $M = 30$	0.865 (0.439)	0.840 (0.439)	0.792 (0.352)	0.675 (0.324)	0.843 (0.453)
$\tau = 0.05$, $M = 45$	0.905 (0.363)	0.905 (0.364)	0.883 (0.308)	0.758 (0.343)	0.607 (0.455)

Matrix	D'	SCB sym-t	SCB eq-t	SCB pct.	SCB BCa	NAIVE	NAIVE@w
GIFT-Eval	8	0.89/0.90	0.89/0.91	0.85/0.86	0.79/0.83	0.87/0.90	0.89/0.90
GIFT-Eval	16	0.88/0.91	0.87/0.91	0.83/0.87	0.76/0.83	0.87/0.91	0.85/0.90
GIFT-Eval	24	0.84/0.90	0.84/0.89	0.82/0.88	0.77/0.83	0.81/0.88	0.83/0.88
LTSF-own	8	0.88/0.97	0.82/0.91	0.69/0.85	0.79/0.88	0.86/0.94	0.89/0.99
LTSF-own	16	0.75/0.94	0.68/0.87	0.66/0.89	0.66/0.86	0.75/0.94	0.78/0.96
LTSF-own	24	0.69/0.96	0.61/0.87	0.59/0.92	0.59/0.87	0.63/0.93	0.68/0.96
TabArena	8	0.92/0.99	0.87/0.95	0.79/0.94	0.69/0.85	0.81/0.95	0.93/0.98
TabArena	16	0.93/1.00	0.80/0.88	0.77/0.92	0.61/0.80	0.82/0.94	0.96/1.00
TabArena	24	0.92/1.00	0.81/0.91	0.70/0.86	0.58/0.80	0.71/0.93	0.94/0.99
fev-bench	8	0.89/0.89	0.89/0.91	0.75/0.78	0.84/0.87	0.88/0.89	0.90/0.91
fev-bench	16	0.87/0.93	0.88/0.92	0.84/0.89	0.87/0.90	0.89/0.93	0.90/0.94
fev-bench	24	0.86/0.92	0.87/0.92	0.83/0.90	0.83/0.91	0.86/0.92	0.87/0.92

A.16 How much does the family definition matter?

Family labels are an a priori judgement and §4 leans on them, so Table 14 re-runs the blocked audit under a ladder of defensible partitions of each matrix, always against the matched i.i.d. control at the same unit counts. The decisive column is the *shuffled* null: a random partition with exactly the same number of families and the same family sizes, which destroys any real within-family correlation while preserving the coarseness of the draw. If the blocking cost were an artifact of drawing units in big chunks, the shuffled arm would reproduce it. A data-driven partition, clustering units by their empirical error covariance, is the natural alternative to a priori labels; we do not use it because it would define the units from the very outcomes being audited, a selection this paper would then have to audit in turn, and because a partition that tracks the residual correlation by construction leaves the shuffled null with nothing to test. The ladder of a priori partitions below is our substitute.

It does not, anywhere: the shuffled cost is within its own Monte-Carlo spread of zero in every cell of every matrix. What the partitions change is how much genuine structure they capture. On GIFT-Eval, grouping by source dataset costs $+0.098$ to $+0.110$ of retention and grouping by source-and-frequency costs $+0.038$ to $+0.060$, while partitions that do *not* track the source (by sampling frequency, or by forecast horizon) cost nothing (-0.072 to $+0.014$). The cost appears exactly when

the partition tracks provenance and vanishes when it does not, which is the specificity check the claim needs. TabArena, the control, shows no cost under any partition (-0.020 to $+0.020$).

Our own LTSF grid is where this matters most, because the paper's sharpest blocked result lives there and the paper's own caveat is that six families cannot identify a magnitude. The ladder is reassuring about the direction and unforgiving about the number. Pooling the four ETT views into one family costs $+1.289$; splitting ETT by sampling frequency costs $+0.874$; treating all nine source datasets as separate families still costs $+0.842$; blocking by horizon instead costs $+0.877$. So the conclusion that counting configurations as datasets manufactures a champion does not depend on pooling ETT. The retention *level* does: it is -0.365 under the paper's partition but $+0.025$ once ETT is split, so the headline that the champion goes negative is a property of the ETT-pooled reading, while the blocking cost of roughly 0.8 to 1.3 is common to all of them. We state the cost, not the sign, as the finding.

Table 14: Family-partition sensitivity. For each matrix, each defensible partition, and each hold-in family count: the blocking cost (matched i.i.d. retention minus blocked retention) and the same quantity under a random partition with an identical size profile ($\pm$ SD over 20 shuffles). G is the number of families, u the mean units held in. The cells are an independent 600-draw pass, so they differ from Table 3's by Monte-Carlo error: on the LTSF two-family cell, whose SE is ≈ 0.15, this pass reports a cost of $+1.29$ and retention -0.365 against $+1.23$ and -0.30 there, both straddling the exact enumerated -0.347 of App. A.11.

Partition	G	fam. held in	u	blocking cost	shuffled null
TabArena					
dataset (paper)	51	2	2.0	-0.008	$+0.003^{\pm 0.039}$
dataset (paper)	51	8	8.0	-0.011	$+0.021^{\pm 0.083}$
dataset (paper)	51	17	17.0	$+0.020$	$-0.008^{\pm 0.098}$
singleton (i.i.d.)	51	2	2.0	-0.020	$+0.009^{\pm 0.027}$
singleton (i.i.d.)	51	8	8.0	-0.011	$-0.020^{\pm 0.083}$
singleton (i.i.d.)	51	17	17.0	-0.014	$+0.016^{\pm 0.108}$
GIFT-Eval					
dataset (paper)	28	2	7.0	$+0.022$	$-0.011^{\pm 0.046}$
dataset (paper)	28	4	13.5	$+0.110$	$-0.002^{\pm 0.035}$
dataset (paper)	28	9	30.9	$+0.098$	$-0.001^{\pm 0.025}$
dataset+freq	55	2	3.5	$+0.046$	$+0.006^{\pm 0.037}$
dataset+freq	55	9	15.7	$+0.060$	$-0.006^{\pm 0.033}$
dataset+freq	55	18	31.8	$+0.038$	$-0.001^{\pm 0.022}$
frequency	10	2	19.0	-0.018	$+0.007^{\pm 0.045}$
frequency	10	3	29.4	$+0.014$	$+0.007^{\pm 0.054}$
horizon	3	2	64.1	-0.072	$-0.004^{\pm 0.075}$
singleton (i.i.d.)	97	2	2.0	-0.009	$-0.004^{\pm 0.046}$
singleton (i.i.d.)	97	16	16.0	$+0.004$	$+0.000^{\pm 0.030}$
singleton (i.i.d.)	97	32	32.0	-0.012	$-0.003^{\pm 0.024}$
fev-bench					
leading token	31	2	5.9	$+0.043$	$-0.013^{\pm 0.078}$
leading token	31	5	14.6	$+0.045$	$-0.006^{\pm 0.033}$
leading token	31	10	29.2	$+0.023$	$-0.009^{\pm 0.032}$
singleton (i.i.d.)	91	2	2.0	$+0.030$	$+0.020^{\pm 0.058}$
singleton (i.i.d.)	91	15	15.0	$+0.016$	$-0.005^{\pm 0.030}$
singleton (i.i.d.)	91	30	30.0	-0.008	$-0.005^{\pm 0.019}$
source (paper)	51	2	3.6	$+0.012$	$-0.025^{\pm 0.083}$
source (paper)	51	8	14.4	$+0.006$	$-0.004^{\pm 0.039}$
source (paper)	51	17	30.6	$+0.014$	$+0.002^{\pm 0.018}$
LTSF-own					
horizon	8	2	9.0	$+0.877$	$-0.060^{\pm 0.219}$
singleton (i.i.d.)	36	2	2.0	-0.000	$+0.017^{\pm 0.145}$
singleton (i.i.d.)	36	6	6.0	$+0.027$	$+0.017^{\pm 0.079}$
singleton (i.i.d.)	36	12	12.0	-0.014	$+0.012^{\pm 0.075}$
source, ETT by frequency	7	2	10.4	$+0.874$	$+0.045^{\pm 0.117}$
source, ETT pooled (paper)	6	2	12.0	$+1.289$	$-0.050^{\pm 0.135}$
source, ETT split	9	2	8.0	$+0.842$	$+0.037^{\pm 0.128}$
source, ETT split	9	3	12.0	$+0.313$	$-0.001^{\pm 0.082}$

A.17 Are the counts inflated by near-duplicate method columns?

A leaderboard's entries are not always distinct ideas. TabArena's 45 are 16 base learners crossed with `(default)`, `(tuned)` and `(tuned + ensemble)`; GIFT-Eval's 119 include seven Toto sizes, five TTM variants and three Chronos sizes. Since §3 assigns an a priori family label to every evaluation *unit* for exactly this reason, the method axis deserves the same treatment, and two quantities could plausibly be inflated by it: $P(\#1)$, if the reported winner and the complement's winner are often two configurations of one learner, and retention, if a configuration-level leaderboard offers more chances to be lucky than a learner-level one would.

The first concern is measurable and turns out to be empty. Collapsing each entry to its base learner (Table 15) leaves $P(\#1)$ *unchanged to three decimals* on all four matrices at every suite size, because the event it would have to explain never occurs: across $2{,}000$ draws in each of twelve cells, the selected method and the complement's winner are two different configurations of the same learner exactly zero times. TabArena's misses are AutoGluon against REALMLP against TabM against TabPFNv2, genuinely different learners, so none of the 0.39 hit rate is an artifact of counting `CAT (tuned)` and `CAT (tuned + ensemble)` separately.

The second concern is real and we quantify it. Listing one row per learner, scored as that learner's mean over its configurations, raises TabArena's retention from $+0.082$ to $+0.131$ at $D'{=}8$ and from $+0.357$ to $+0.803$ at $D'{=}24$; on GIFT-Eval, where 102 of 119 entries are already singletons, it moves nothing ($+0.744 \rightarrow +0.749$). Part of that gain is mechanical, since a mean over three configurations is less noisy than any one of them and less noise is less winner's curse, so this is not a like-for-like comparison of two estimators but of two *leaderboard designs*. Read that way it is a reporting suggestion with a measured payoff: a tabular leaderboard that reported one row per learner rather than one per configuration would have a top row that transfers roughly twice as well at $D'{=}24$. It is also a caveat on our own M-sweeps and on M_{eff}, both of which count configurations.

Two GIFT-Eval entries carry names suggesting they are not finished submissions (`MoiraiAgent-leaking` and `TSOrchestra-test`). We keep every entry the public leaderboard lists, because any inclusion rule of our own would be a selection we would then have to audit; both sit mid-table (ranks 24 and 39 of 119) and neither is the winner under any scale, so no reported number depends on them.

Table 15: The method axis, blocked by base learner. M_{fam} counts distinct learners. "near-miss" is the share of draws in which the selected method and the complement's winner are different configurations of the *same* learner—the event that would have to be common for $P(\#1)$ to be deflated by duplicate columns. "dedup" re-runs the audit on a leaderboard collapsed to one row per learner, scored as that learner's mean over its configurations. $2{,}000$ draws per cell.

Matrix	M	M_{fam}	D'	$P(\#1)$	$P(\#1$ fam$)$	near-miss	retention	dedup
TabArena	45	17	8	0.387	0.387	0.000	+0.082	+0.131
			16	0.305	0.305	0.000	+0.223	+0.394
			24	0.250	0.250	0.000	+0.357	+0.803
GIFT-Eval	119	102	8	0.338	0.338	0.000	+0.744	+0.749
			16	0.482	0.482	0.000	+0.848	+0.851
			24	0.570	0.570	0.000	+0.895	+0.896
fev-bench	22	22	8	0.559	0.559	0.000	+0.880	+0.880
			16	0.775	0.775	0.000	+0.939	+0.939
			24	0.855	0.855	0.000	+0.962	+0.962
LTSF-own	13	10	8	0.892	0.892	0.000	+0.850	+0.927
			16	0.986	0.986	0.000	+0.969	+0.985
			24	0.959	0.959	0.000	+0.976	+0.976