THE SHORTLIST ILLUSION: WHY LEADERBOARD SEPARATION FAILS CANDIDATE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Public LLM leaderboards rank hundreds of models across macro benchmark frames, and practitioners routinely rely on these aggregate scores to guide high-stakes model selection and enterprise deployment among a small shortlist of candidate models. We show that macro-level benchmark performance fundamentally fails to transfer to candidate shortlists, creating a pervasive shortlist illusion. Using cohort-conditional separation—the fraction of within-cohort model pairs distinguishable beyond an operational performance margin—we audit 9,103 model–date records for 4,108 models across 133 snapshots of the Open LLM Leaderboard v2. We isolate three distinct failure mechanisms that misinform deployment decisions: (1) Non-identification: separation is fundamentally undefined without naming the comparison cohort; on the same MUSR snapshot, a 35-model frame separates 69.6% of pairs, whereas a 5-model candidate shortlist separates only 20.0%. (2) Selection-induced compression: filtering top-K models restricts the observed score range, collapsing discriminative power; on MATH-Hard, separation plummets from 91.1% across the full frame to 0% among the top 5 models. (3) Small-sample noise: while random sub-cohorts are unbiased in expectation, small candidate sets suffer severe dispersion, reducing top-model selection to near-chance agreement. We translate these findings into an actionable, zero-cost six-field reporting contract and recommend cohort conditionality as an indispensable auditing requirement to curb misleading model cards and prevent wasteful deployment choices.

1 INTRODUCTION

Benchmark numbers are routinely interpreted as immutable facts about model capabilities. We ask a more urgent, practical question: can a macro leaderboard report reliably guide selection among the specific candidate models a practitioner actually intends to deploy? As the MUSR example shown in Figure 1, on the 2025-03-13 snapshot, the same scores and strict $>$ 3-percentage-point rule separate 414/595 pairs (69.6%) in the 35-model eligible frame but only 2/10 pairs (20%) in a prespecified five-model candidate cohort whose observed scores happened to lie in the 42–46% range. Nothing about the benchmark, snapshot, scores, or threshold changed; only the candidate cohort changed.

This gap carries major real-world consequences. In production settings, teams never deploy an entire public archive; enterprise budgets, licensing, latency, and hardware constraints invariably narrow evaluation down to a shortlist of 3–5 viable models. Substantial infrastructure budgets, hardware acquisitions, and engineering cycles are routinely committed based on fractional leaderboard margins, under the unspoken assumption that leaderboard separation carries over to the candidate shortlist. Yet a report that omits the cohort silently asks an archive-wide summary to stand in for a local, candidate-level decision. We show this assumption induces a critical blind spot in LLM benchmarking. While the evaluation literature has been meticulous about auditing task suites, submission policies, prompts, judges, and inference protocols (Dehghani et al., 2021; Liang et al., 2022; Srivastava et al., 2022; White et al., 2025; Singh et al., 2025; Sclar et al., 2024; Potamitis et al., 2025; Zheng et al., 2023; Dorner et al., 2025; Zhang et al., 2026; Dror et al., 2018; Hariri et al., 2026; Kübler et al., 2026; Shen et al., 2026) (reviewed in Section 2), the compared model set itself has remained entirely unexamined—either held fixed, grown as a side effect, or left implicit. We invert this paradigm: controlling every standard source of evaluation variation, we isolate the cohort as the sole manipulated factor.

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.

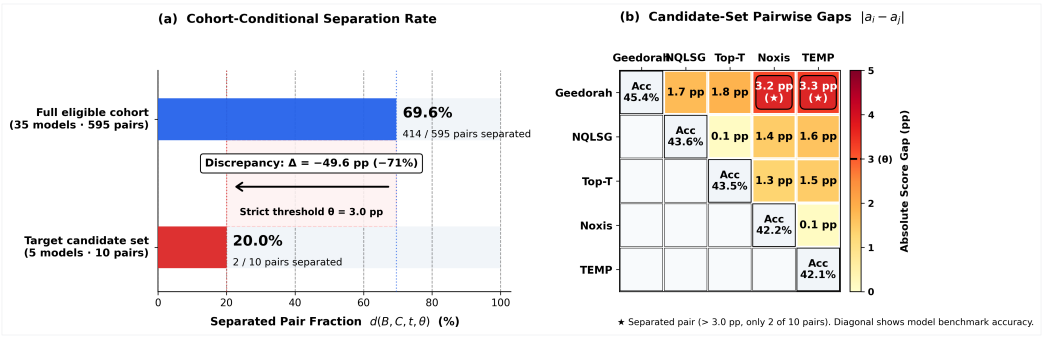

Figure 1. The same benchmark, snapshot, and threshold; two cohorts; a 49.6-point difference in the report. MUSR on the 2025-03-13 snapshot under the strict $>$ 3-percentage-point rule. The 35-model eligible cohort separates 414/595 pairs (69.6%); a prespecified five-model candidate set whose observed scores happened to lie in the 42–46% range separates 2/10 (20%).

We make four contributions to evaluation methodology and auditing practice. First, we formalize transportability in terms of cohort-conditional separation and its associated transfer gap; Propositions 1–3 establish that cohort conditionality is a problem of identification and post-selection inference rather than a definitional afterthought. Second, we empirically distinguish three cohort regimes in OLLB v2 (full eligible frames, score-selected top-K frontiers, and candidate target sets) and show that top-K selection compresses cohort-conditional separation by up to 36.6 percentage points. Third, we audit 26,600 target cohorts across 133 snapshots and find that winner predictions under small shortlists degrade to near-chance agreement, exposing the acute risk of using macro-level rankings for micro-level selection. Fourth, we translate these findings into an actionable, zero-overhead six-field reporting contract that restores auditable accountability to model cards and averts wasteful deployment choices.

2 Related Work

2.1 Benchmark suites, protocols, and run sensitivity

HELM, BIG-bench, and lm-evaluation-harness established broad multi-task suites and standardized protocols (Liang et al., 2022; Srivastava et al., 2022; EleutherAI, 2023), and contamination-resistant designs such as LiveBench refresh tasks to preserve discriminative power (White et al., 2025); the four benchmarks in our archive were produced under distinct task and scoring protocols (Suzgun et al., 2023; Rein et al., 2023; Hendrycks et al., 2021; Sprague et al., 2024). Meaning-preserving format perturbations move accuracy by up to 76 points (Sclar et al., 2024), multiple-choice presentation has its own failure modes (Balepur et al., 2025), repeated runs of one configuration produce overlapping score distributions that can silently flip rankings (Potamitis et al., 2025), and analogous protocol sensitivity appears outside NLP, where the choice of distance measure alone materially changes measured k-NN performance (Abu Alfeilat et al., 2019). This literature changes *what* is measured and reports per-model variance for a fixed compared set; we hold that fixed and ask how much of an aggregate comparison claim is carried by the set itself.

2.2 Judges, rankings, and leaderboard audits

Model-based judges introduce protocol-dependent variation (Zheng et al., 2023; Bavaresco et al., 2025; Li et al., 2025), and debiasing a judge cannot buy more than a factor of two in label efficiency (Dorner et al., 2025). A parallel line asks whether rankings are valid at all: multi-task benchmarks face an inherent diversity–stability trade-off (Zhang & Hardt, 2024), training on the test task confounds comparison (Dominguez-Olmedo et al., 2025), and identical benchmark-specific fine-tuning harmonizes rankings across 24 benchmarks (Zhang et al., 2026). Audits of live leaderboards document how submission policy reshapes public ranks (Singh et al., 2025), and re-evaluating open-ended evaluation exposes redundancy sensitivity in Elo-based systems (Liu et al., 2025); such com-

parison long predates LLMs (Baru et al., 2013). All of these characterize stability or validity *across* benchmarks, protocols, or policies, for a cohort that is fixed or grows as a side effect.

2.3 Inference and reporting standards

Guidance on significance testing and power (Dror et al., 2018; Card et al., 2020), Bayesian posteriors replacing Pass@k with credible intervals (Hariri et al., 2026), paired tests resolving 0.3-point degradations under a controlled false-positive rate (Kübler et al., 2026), certification with imperfect judges (Shen et al., 2026), and frameworks for ranking pairwise human judgments (Ameli et al., 2025) all sharpen inference for a *specified* comparison; classical test theory, IRT, and DIF provide item-level tools for test quality and invariance (Lord, 1980; Hambleton et al., 1991; Nunnally & Bernstein, 1994; American Educational Research Association, 2014), and the measurement model itself encodes normative commitments (Jacobs & Wallach, 2021). Our estimand is deliberately not a test but the descriptive object such a test would be applied to, and our point is that this object is indexed by the cohort. Reporting guidelines (Dodge et al., 2019; Gallifant et al., 2025) demand transparent specification of model, data, and evaluation detail but assume item-level responses or metadata a public aggregate archive does not publish; our schema is designed for exactly the data a leaderboard actually releases.

Across these lines, the compared cohort is fixed, grows as a side effect, or stays implicit. We invert the design: everything the field has learned to control is controlled, and the one factor public reports never name is the one we vary.

3 Framework

3.1 An estimand indexed by the cohort

Let B be one benchmark, C a finite observed cohort of models, and t an archive snapshot date. The archive stores aggregate scores in percentage points; we write $a_i(B,t)$ for the recorded score of model $i \in C$. For a non-negative operational threshold θ, also in percentage points, cohort-conditional separation is

$$d(B,C,t,\theta) = \frac{1}{\binom{|C|}{2}} \sum_{\substack{i<j \\ i,j\in C}} \mathbb{1}\left[|a_i(B,t) - a_j(B,t)| > \theta\right], \tag{1}$$

for $|C| \geq 2$. Missing scores are excluded before the cohort is formed. The indicator uses an absolute gap and a strict $>$ rule, so d is a descriptive fraction of unordered model pairs with threshold-separated recorded scores. We refer to d as the separation rate. A worked three-model example is given in Appendix A.

We use $\theta = 3$ percentage points as the primary operational rule and write d_3 for the resulting quantity; $\{1, 2, 5\}$ percentage points are sensitivity analyses, reported for all three mechanisms in Appendix Table 5. For $\theta > 0$, $d = 0$ means only that the observed score range is at most θ, not that all scores are identical. Score spread and d are two summaries of the same score vector. While our primary empirical case study evaluates accuracy on the Open LLM Leaderboard v2, the estimand is metric-agnostic: $a_i(B,t) \in \mathbb{R}$ equally represents Pass@k rates on code generation (Chen et al., 2021; Hariri et al., 2026) or Bradley-Terry Elo ratings on preference arenas (Chiang et al., 2024).

3.2 What the estimand does and does not identify

Three properties explain why cohort conditionality is a reporting problem rather than a definitional curiosity; proofs are in Appendix A. They are also the three failure mechanisms this paper separates, and they differ in kind. An **unnamed cohort** leaves the report *unidentified*: (B,t,θ) does not determine d (Proposition 1). A **score-selected cohort** can exhibit *selection-induced distortion*: selecting with the score under evaluation restricts the range that score can take, and no uniform-subset unbiasedness result applies (Proposition 2). A **uniformly drawn cohort** is unbiased in expectation but has *finite-cohort sampling variance*: the remaining error is dispersion under the stated sampling design (Proposition 3). Identification failure, selection-induced distortion, and noise have different

remedies, so keeping them apart is the organizing claim of the paper; Sections 5.1–5.3 measure one each.

Proposition 1 (Non-identification without the cohort) *Fix B, t, and θ, and let M be the eligible cohort at that benchmark and date. If the score range of M exceeds θ, there exists a sub-cohort $C_1 \subseteq M$ with $d(B, C_1, t, \theta) = 1$. If in addition M contains two models whose scores differ by at most θ, there exists a sub-cohort $C_2 \subseteq M$ with $d(B, C_2, t, \theta) = 0$. Consequently d is not a function of (B, t, θ) alone, and d is not monotone under cohort inclusion: $C' \subset C$ implies no ordering between $d(B, C', t, \theta)$ and $d(B, C, t, \theta)$.*

Proposition 1 formalizes the transport problem: a score gap does not travel because the reported quantity is not fully specified until the cohort is named. In our reported MUSR case, the five candidate models have scores $\{42.06, 42.20, 43.52, 43.65, 45.37\}$ and yield $d_3 = 2/10 = 20\%$, whereas the 35-model eligible frame yields $414/595 = 69.6\%$ on the same snapshot. This is a specified-candidate contrast, not a score-selected top-K contrast.

Proposition 2 (Post-selection has no uniform-subset guarantee) *Let $C \subseteq F$ be any sub-cohort whose score range is at most θ. Then $d(B, C, t, \theta) = 0$ for every benchmark B and snapshot t. Because the top-m cohort is a deterministic function of the scores rather than a uniform draw, no uniform-subset unbiasedness guarantee applies, and the failure does not vanish as the cohort grows: for every $m \geq 2$ there are frames with $|F| > m$ whose top-m cohort has $d(B, C, t, \theta) = 0$ while $d(B, F, t, \theta)$ approaches $24/25$ at $\theta = 3$.*

The post-selection result is a deterministic range-restriction statement: it does not assert that every top-m cohort has zero separation, but it shows that score-based selection can make a zero report available at every cohort size. The empirical C2 analysis therefore tests the direction and magnitude of this compression, rather than treating the proposition as a universal equality.

Proposition 3 (Uniform sub-cohorts are unbiased) *Let F be a sampling frame of n models with valid scores and let $S \subseteq F$ be a uniformly random subset of size m with $2 \leq m \leq n$. Then*

$$\mathbb{E}\big[d(B, S, t, \theta)\big] = d(B, F, t, \theta).$$

Moreover $d(B, S, t, \theta)$ takes values on the lattice $\{k/\binom{m}{2} : k = 0, \ldots, \binom{m}{2}\}$, so a five-model report moves in increments of 10 percentage points.

The proof is a one-line symmetry argument: every pair in F enters S with equal inclusion probability $\binom{m}{2}/\binom{n}{2}$. In the language of finite population sampling (Cochran, 1977; Thompson, 2012), $d(B, S, t, \theta)$ is a symmetric degree-2 U-statistic evaluated over a finite population without replacement.

The consequence is the paper's most diagnostic result: **a cohort-wide separation report is unbiased in expectation for a uniformly drawn candidate set, but its realized transport error can be large.** Whatever a practitioner loses by substituting a leaderboard-wide number for a uniformly sampled shortlist is finite-sample variance under this design, not systematic distortion. For $3 \leq m < n$, pairs sharing a model have nonnegative covariance, strictly positive except in the finite case $(n, m) = (4, 3)$; the $m = 2$ boundary has $\pi_1 = 0$. The finite-population covariance structure also includes disjoint pairs (Appendix A.4). Thus the effective sample size need not equal the nominal $\binom{m}{2}$ pair count, and variance can be substantial for shortlists with $m \leq 10$, where the lattice is coarsest as well. The remedy the theory implies is accordingly not a better benchmark alone, but an explicitly stated cohort and denominator.

Proposition 2 and Proposition 3 are deliberately adjacent: selection changes the estimand through the membership rule, whereas uniform sampling preserves the frame value only in expectation and leaves finite-cohort sampling variance under the stated frame and without-replacement design. The three empirical sections follow the corresponding reporting order: C1 establishes the benchmark-native frame value, C2 measures score-selected compression, and C3 measures uniform-sampling dispersion. Reporting conventions that conflate them are therefore not merely imprecise; they report different functionals.

3.3 THREE COHORT CONSTRUCTIONS

We evaluate the same separation functional under three constructions. The **benchmark-native full-cohort** report uses $C = M_{B,t}$, the models with valid scores for benchmark B at date t; it describes score geometry at a snapshot. The **score-selected top-K** report uses $C = F^K_{B,t}$, selected by descending score within the same benchmark and date, with deterministic identifier-based tie breaking; it exposes post-selection range restriction. The **target-set** report uses $C = S \subseteq F^{\mathrm{cc}}_t$ when a candidate cohort is specified; F^{cc}_t is the complete-case common frame with valid scores on all four benchmarks and is used only as the sampling frame. Target-set values are compared with the benchmark-native full frame, not silently with the common frame. A target set of size $|S|$ contributes $\binom{|S|}{2}$ pairs; the constructions differ in membership determination and reference frame, while the separation functional itself is unchanged.

For two cohorts evaluated on the same benchmark, snapshot, and threshold, we write their descriptive contrast as

$$\Delta(C_1, C_2 \mid B, t, \theta) = d(B, C_1, t, \theta) - d(B, C_2, t, \theta). \tag{2}$$

A positive value means C_1 contains a larger fraction of threshold-separated pairs. Contrasts are interpretable only when B, t, and θ are held fixed, which is precisely the discipline our design imposes.

Canonical benchmark order. Whenever two benchmarks at the same date return identical d_3 we break ties in the canonical order $\mathrm{BBH} \succ \mathrm{GPQA} \succ \mathrm{MATH\text{-}Hard} \succ \mathrm{MUSR}$. This convention is declared once here and reused by the winner-agreement tables and figures; it is independent of any subsequent leaderboard re-ranking.

Transport diagnostics. Let $F_{B,t}$ denote the eligible full frame and $S \subseteq F_{B,t}$ a target cohort. We define the signed transfer gap

$$\mathrm{TG}(B, S, t, \theta) = d(B, S, t, \theta) - d(B, F_{B,t}, t, \theta), \tag{3}$$

and the target-size transport error

$$\mathrm{TE}_m(B, t, \theta) = \mathbb{E}_{S:|S|=m}\left[|\mathrm{TG}(B, S, t, \theta)|\right]. \tag{4}$$

The signed gap diagnoses direction; the absolute error measures how far a frame-level report can be from an m-model target cohort. For a multi-benchmark diagnostic, we call the benchmark maximizing descriptive d_3 the *separation winner*; ties follow the canonical benchmark order of Section 3.3, and winner mismatch is a migration diagnostic.

4 DATA AND METHODS

4.1 STUDY DESIGN

Our primary analysis is a controlled comparison of cohorts: for each within-benchmark contrast we hold the benchmark, snapshot, score construction, units, threshold, and strictness rule fixed, so that cohort membership is the intended varying factor. The C3 analysis additionally distinguishes the complete-case sampling frame from the benchmark-native reference and reports their small composition gap. These are finite, cross-sectional descriptions of released scores; they are not estimates of benchmark validity, model capability, or downstream utility.

The unit of analysis is a benchmark–date–cohort report: we construct the eligible model set for a benchmark and snapshot, then compute all unordered pair gaps within it. We use the strict rule $|a_i - a_j| > 3$ percentage points throughout, retaining 1, 2, and 5 as sensitivity analyses. We never impute a missing score, and we keep the pair denominator explicit rather than treating overlapping pairs as independent observations.

4.2 THE OLLB V2 ARCHIVE

The primary archive is the cached Open LLM Leaderboard v2 collection of 4,108 unique models and 9,103 model–date records for BBH, GPQA, MATH-Hard, and MUSR (Hugging Face, 2024).

We use this archive because it is dated, publishes per-model standard errors, and exposes the same aggregate scores that practitioners read; it does not provide item-level responses, so our analysis remains deliberately at the aggregate-score level. The raw archive spans 223 dates. We retain the 133 dates shared by all four benchmark groups and require at least ten eligible models per benchmark, yielding 532 benchmark/date observations.

Preprocessing follows a fixed provenance rule rather than a result-dependent filter. We resolve duplicate model–date records using the documented latest-record policy, preserve the selected record in the provenance manifest, exclude records without a valid aggregate score, and perform no score imputation. We then intersect the retained dates across benchmarks for the primary analysis. For C3, the complete-case frame additionally requires a valid score on all four benchmarks; this frame is used only for the target-set simulation and is reported separately from benchmark-specific eligibility.

4.3 Cohort construction protocols

For C1, we compute d_3 on each benchmark-specific full eligible cohort and give every retained date equal weight. We obtain descriptive intervals for date-level summaries with 2,000 replicates of a five-consecutive-date block bootstrap, which preserves local temporal dependence without pretending that adjacent snapshots are independent. For C2, we sort eligible models by recorded score, break ties deterministically by model identifier, select the top $K \in \{5, 10, 20, 50\}$, and evaluate d_3 on those same selected scores. This deliberate reuse of the selection score is the post-selection operation described by Proposition 2.

For C3, we evaluate a prespecified illustrative five-model set whose observed MUSR scores happened to lie in the 42–46% range on 2025-03-13, and sample 200 target sets without replacement per eligible date and target size $\{5, 10, 20, 50\}$ from the complete-case intersection frame. This produces 26,600 size-5 sets. For each target set, the benchmark winner is the argmax of d_3 over the canonical benchmark order declared in Section 3.3; the reference benchmark-native full-frame winner is the argmax of benchmark-native full-frame d_3 values. Winner agreement is the fraction of target sets for which these two benchmark winners match, and winner-mismatch gaps are computed in percentage points. The complete-case frame values are retained separately to quantify any frame-composition difference.

Finally, we propagate the archived per-model standard errors only as a sensitivity analysis. For each pair, we approximate the standard error of the score difference by the square root of the two squared standard errors, assume independence, and retain a pair only when the two-sided normal-approximation lower bound for the absolute gap exceeds 3 percentage points. This procedure is not a capability test; it shows how much the descriptive separation rate changes when the archive's reported measurement uncertainty is taken at face value.

5 Results

We organize the results around three cohort constructions. C1 establishes the benchmark-native full-cohort reference; C2 quantifies the selection-induced compression produced by score-based top-K selection, as formalized in Proposition 2; and C3 measures the finite-cohort variance that remains under uniform target-set sampling, as formalized in Proposition 3.

5.1 C1: Separation tracks score dispersion, not benchmark quality

Table 1 calibrates the estimand over 532 benchmark/date observations. Score standard deviation is positively associated with d_3 for every benchmark, with Spearman correlations from 0.50 [0.34, 0.65] for BBH to 0.88 [0.81, 0.93] for MUSR. The association is expected because both quantities come from the same score vectors; we report it as calibration of the threshold rule, not as a benchmark-quality finding. The median within-cohort gap makes the same point numerically, at 10.9 percentage points for BBH but 3.4 for GPQA: the fixed 3-point rule is permissive for BBH and MATH-Hard and demanding for GPQA and MUSR, where 45.5% and 36.5% of pooled pairs fall below 3 points (Appendix Table 4 gives the full binning). It is an operational reporting convention, and every contrast below holds it fixed.

Table 1. C1 calibration across the 133 retained snapshots (532 benchmark/date observations). ρ is the descriptive Spearman association between within-cohort score SD and d_3, with 95% five-consecutive-date block-bootstrap intervals (2,000 replicates). Median gap is the pooled within-cohort median absolute pair gap. Point d_3 is the date-equal-weighted full-cohort value.

Benchmark	$\rho(\mathrm{SD}, d_3)$ [95% int.]	Median gap (pp)	Point d_3
BBH	0.50 [0.34, 0.65]	10.9	78.4%
GPQA	0.82 [0.73, 0.89]	3.4	52.3%
MATH-Hard	0.74 [0.61, 0.83]	7.5	74.7%
MUSR	0.88 [0.81, 0.93]	4.2	60.6%

Table 2. Score-selected top-five conditioning across 133 snapshots. Full and Top-5 are equal-weighted means of date-level d_3; Δ is Top-5 minus Full (pp), computed from unrounded means, so the rounded columns can differ from it by up to 0.1 pp.

Benchmark	Full d_3	Top-5 d_3	Δ (pp)	95% interval
BBH	78.4%	43.3%	-35.1	[-40.3,-30.0]
GPQA	52.3%	25.7%	-26.6	[-31.0,-22.1]
MATH-Hard	74.7%	60.5%	-14.1	[-18.3,-10.0]
MUSR	60.6%	23.9%	-36.6	[-40.8,-32.7]

Table 3. Empirical test of Proposition 3 at $|S| = 5$ (26,600 sampled sets). Gap is mean target-set d_3 minus benchmark-native full-cohort d_3; the size-5 frame-composition gap is at most 0.002 percentage points on average. Disp. is within-date SD.

Benchmark	**Gap (pp)**	**95% interval**	**Disp. (pp)**
BBH	+0.04	[-0.15, +0.22]	13.9
GPQA	+0.11	[-0.12, +0.34]	18.5
MATH-Hard	+0.02	[-0.17, +0.21]	15.0
MUSR	-0.01	[-0.20, +0.17]	17.0

The uncertainty adjustment changes the level, not the cohort-conditionality. Propagating the archive's own per-model standard errors moves the C1 level substantially on two of four benchmarks: GPQA falls from 52.3% to 20.2% and MUSR from 60.6% to 21.1%, retaining only 35.9% and 33.5% of pairs, whereas BBH and MATH-Hard fall by 8.0 and 9.8 points. The independent-standard-error rule is a sensitivity analysis—a pair survives only when the lower bound of its 95% interval for the gap already exceeds 3 points—and ignores covariance from shared evaluation items. It therefore need not produce a guaranteed lower bound; the direction of conservatism depends on the omitted covariance. The adjustment disciplines the *level*: on GPQA and MUSR, d_3 is not a count of statistically resolved differences, which is why θ is never a significance threshold here. It does not touch cohort conditionality, since Sections 5.2–5.3 recompute the same functional on the same score vectors at the same θ; the interaction with finite-cohort dispersion remains open.

5.2 C2: SCORE-SELECTED TOP-K COHORTS—SEPARATION COMPRESSION

The running MUSR example above changes only the cohort. C2 isolates a second mechanism: selecting the cohort with the score being evaluated. On the final retained snapshot (2025-03-13), the MATH-Hard full cohort separates 542/595 pairs (91.1%), whereas the five highest-scoring models separate 0/10 (0%). This is not the same cohort as the illustrative MUSR candidate set: the former is score-selected, while the latter is specified independently of the reported score. For MUSR, the score-selected top five instead separate 8/10 pairs. The MATH-Hard contrast is a selection effect.

Across all 133 retained snapshots, the equal-weighted top-five report is lower than the full-cohort report by 35.1, 26.6, 14.1, and 36.6 percentage points for BBH, GPQA, MATH-Hard, and MUSR; Table 2 reports the summary and Figure 3 plots both series snapshot by snapshot. The final-snapshot MUSR increase therefore does not contradict the across-date mean decrease; they summarize different date-level observations. As $K \to |M_{B,t}|$, the top-K cohort converges to the eligible frame and the contrast tends to zero. Intermediate-K contrasts need not be monotone, however, and the $K \in \{5, 10, 20, 50\}$ profiles use different eligible-date sets; Figure 4 and Appendix B.2 therefore report descriptive profiles rather than a monotone curve.

5.3 C3: TARGET COHORTS (THE VARIANCE MECHANISM)

5.3.1 SMALL CANDIDATE SETS INTRODUCE SUBSTANTIAL TRANSPORT DISPERSION

We first show an illustrative, prespecified five-model set on the final retained snapshot, 2025-03-13; Appendix B.3 records its fixed-set definition and eligibility details. Its ten pairs make the report

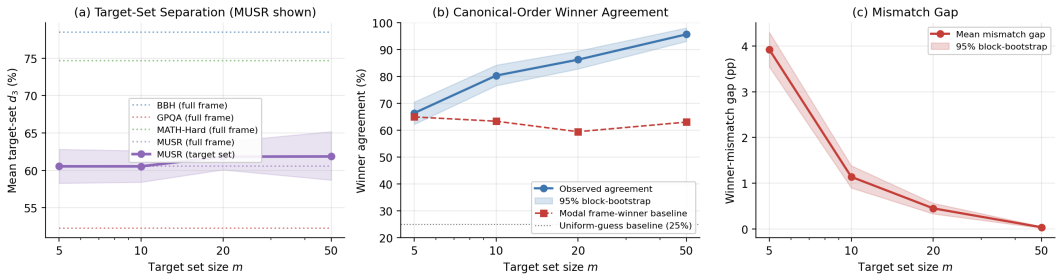

Figure 2. Transport failure is a small-cohort phenomenon. Panel (a) shows the mean target-set d_3 across uniformly drawn MUSR sets, with the four benchmark-native full-frame values shown as dotted horizontal references; the per-benchmark size-5 target diagnostics are given in Table 3. Panel (b) plots canonical-order winner agreement rising from 53.1% at size 5 to 95.4% at size 50, with the 95% five-consecutive-date block-bootstrap band; the size-5 interval contains 50%. Panel (c) shows the winner-mismatch gap with the same bootstrap band, contracting from 3.92 pp at size 5 to 0.03 pp at size 50. The agreement and mismatch trends are the empirical content of Proposition 3.

coarse, with 10-point increments, and it is not intended to be representative. The observed 42–46% MUSR score range is a characterization of this fixed set, not its selection rule.

For MUSR, the benchmark-native full-cohort report is 414/595 (69.6%) whereas the illustrative target set yields 2/10 (20%). Across 133 dates and 26,600 sampled size-5 target sets, 200 per date, the mean MUSR target-set report is 60.5% [58.3, 62.8]. All intervals below are 95% five-consecutive-date block-bootstrap intervals from 2,000 replicates. The mean absolute transport error is 13.3 percentage points for MUSR at size 5 and falls to 1.4 points at size 50. Across all benchmarks, mean absolute transport errors at size 5 are 10.8–14.8 points, while agreement between each target-set winner and the benchmark-native full-frame winner is 53.1% [49.9, 56.4], 78.2% [74.1, 82.1], 85.8% [82.7, 89.1], and 95.4% [92.8, 98.0] for target sizes 5, 10, 20, and 50; corresponding winner-mismatch gaps are 3.92 [3.54, 4.30], 1.14 [0.89, 1.38], 0.45 [0.33, 0.56], and 0.03 [0.02, 0.05] percentage points. (Figure 2). We make no independent p-value claim because the winner is a descriptive argmax over clustered, correlated summaries; the size-5 agreement interval includes 50%, so the result is compatible with chance-level agreement at this descriptive resolution. The mean signed transfer gap remains below 0.12 percentage points while absolute error is large at small sizes: transfer failure is primarily finite-cohort dispersion under uniform sampling, not a claim about arbitrary deployment shortlists. Coverage narrows with target size, because the complete-case frame admits 133 eligible dates at sizes 5 and 10 but only 86 at size 20 and 25 at size 50; the size-50 estimates therefore rest on 18.8% of the date sample. The sampled sets are uniform draws by construction, while deployment shortlists need not.

5.3.2 Testing Proposition 3: signed gaps are negligible, dispersion is not

The reported analysis retains both the complete-case sampling-frame value and the 200 sampled sets per date, so the finite-population unbiasedness calculation can be checked from the reported records. At size $|S| = 5$, the complete-case frame's d_3 differs from the benchmark-native full-cohort d_3 by at most 0.002 percentage points on average; this frame-composition gap is reported rather than absorbed into the reference. Table 3 reports the target-set minus benchmark-native full-frame gap, which includes both composition difference and sampling variation. Across 26,600 size-5 sets, the mean target-set report differs from the benchmark-native full-cohort report by -0.01 to $+0.11$ percentage points, and every 95% block-bootstrap interval contains zero. The same sets have a within-date dispersion of 13.9–18.5 percentage points against a mean signed gap of at most 0.12. The mean within-date SD, averaged equally over benchmark–date observations, contracts from 16.1 to 1.9 percentage points as target size grows from 5 to 50. This is the mechanism behind the agreement recovery in Figure 2: what shrinks with m is spread, not offset.

6 DISCUSSION

6.1 WHAT THE MEASUREMENT ADDS BEYOND THE DEFINITION

While d depends on C by definition, empirical auditing reveals the *structural decomposition* of that dependence. Three findings carry practical significance. **(i) Magnitude.** On MUSR, holding benchmark, snapshot, and threshold fixed, a 35-model frame separates 414/595 pairs while a five-model candidate set separates only 2/10—a 49.6 pp collapse driven solely by cohort substitution. **(ii) Selection distortion versus dispersion.** Uniform substitution is unbiased in expectation (Proposition 3) but incurs severe dispersion (10.8–14.8 pp error at size five), whereas score selection triggers post-selection range compression of 14.1–36.6 pp (Proposition 2). **(iii) Auditing practice.** The reporting contract turns an otherwise untestable illusion into an auditable protocol.

6.2 WHY A REPORTING PROTOCOL, AND NOT A BETTER BENCHMARK

Benchmark proliferation cannot cure cohort transport failure: candidate-set error is governed by finite-sample dispersion and post-selection range restriction, not benchmark defect. Re-benchmarking on newer datasets merely exports the same unstated-cohort illusion. What restores validity is metadata a bare score cannot convey: the explicit cohort, pairwise denominator, and whether models were score-selected. Emitting the six contract fields (Table 7) incurs zero compute or annotation cost, delivering high-leverage institutional governance for leaderboard integrity.

6.3 PRACTICAL SELECTION AND REPORTING INCENTIVES

Enterprise adoption pre-filters models by commercial licenses, latency, memory, and parameter limits. When these operational filters correlate with capability, they force evaluation into the post-selection regime, collapsing pairwise separation by up to 36.6 pp. Uncontextualized percentage claims exploit this ambiguity, projecting an illusion of universal dominance when models within the viable candidate tier are statistically indistinguishable. Mandating exact denominators (e.g., $2/10$) and selection tags anchors claims to verifiable local contexts, curbing marketing hyperbole and preventing costly, misinformed infrastructure commitments.

6.4 LIMITATIONS AND BOUNDARY CONDITIONS

Proposition 3 analyzes uniform sampling; production shortlists assembled under operational constraints shift toward post-selection compression whenever filtering criteria correlate with benchmark scores. The distortion-versus-variance split is therefore a foundational benchmark for uniform draws, not an unconditional guarantee for arbitrary deployment filters. Furthermore, our audit operates on public aggregate scores because leaderboards omit item-level response matrices; while item-level IRT could model item covariances, our aggregate analysis directly audits the exact evidence artifact practitioners and model cards consume. For candidate shortlists, we recommend embedding the six-field contract (Table 7) into model cards and evaluation releases, governed by three operational rules: **no substitution** , **exact fractions for small sets**, and **label selection**.

7 CONCLUSION

In conclusion, separability is fundamentally a property of the evaluated cohort rather than an invariant feature of a benchmark. Even under strictly controlled benchmarks, snapshots, scores, and decision thresholds, macro-level leaderboard claims fail to transfer to candidate shortlists—undermined by identification failure, selection compression, and finite-cohort variance. For practitioners selecting among top-tier models, this divergence renders aggregate leaderboard guidance little better than random chance (53.1% agreement). To establish a principled foundation for model selection, we advocate for an explicit reporting contract that mandates named cohorts and well-defined pair denominators. As a zero-cost standard, this protocol curbs misleading model-card claims, safeguards engineering investments, and lays the groundwork for trustworthy, auditable evaluation across both static and dynamic agentic benchmarks.

Ethics Statement

This work investigates leaderboard reporting standards and evaluation methodology using publicly released benchmark aggregates. It involves no human participants, personal identifiable information (PII), or private data. Beyond academic methodology, our findings bear directly on sustainable AI deployment and economic efficiency: in industry and public-sector adoption, committing infrastructure to suboptimal models based on spurious leaderboard separations incurs millions of dollars in wasted compute, energy, and engineering labor. By formalizing why macro separation metrics fail to transport to candidate shortlists and identifying the governing statistical mechanisms, our goal is to enhance scientific integrity, curb inflated marketing claims in model cards, and provide practitioners with auditable standards to prevent economically and operationally wasteful deployment decisions.

Reproducibility Statement

The accompanying anonymous supplementary package contains the public aggregate inputs, versioned analysis artifacts, and a reproduction script for the main numerical results. Sections 4 and 5 specify the archive date range, latest-record deduplication rule, missing-score policy, threshold boundary, target-set sampling design, bootstrap block length, eligible-date coverage, and score-unit conversion. We do not use item-level responses or score imputation. The uncertainty sensitivity treats archived proportion-scale standard errors as inputs and converts them to percentage-point units before adjustment; shared evaluation-item covariance is unavailable, so the direction of conservatism is not guaranteed. The threshold-grid and uncertainty-propagation analyses are explicitly labelled post hoc and are not presented as preregistered analyses.

AI Use Statement

Generative AI tools assisted with language editing, structural revision, consistency checks, and claim-audit triage. Human authors reviewed the AI-assisted changes and take responsibility for the final manuscript, claims, and materials. No generative model was used to generate the data or run the primary analysis.

REFERENCES

Yaser Abu Alfeilat, Ahmad B Hassanat, Osamah Lasassmeh, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, 7(4):221–248, 2019.

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

American Educational Research Association. *Standards for Educational and Psychological Testing*. AERA, 2014.

Nishant Balepur, Rachel Rudinger, and Jordan Boyd-Graber. Which of these best describes multiple choice evaluation with LLMs? a) forced b) flawed c) fixable d) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3394–3418. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.169. URL https://aclanthology.org/2025.acl-long.169/.

Chaitanya Baru, Milind Bhandarkar, Raghunath Nambiar, et al. Benchmarking big data systems and the bigdata top100 list. *Big Data*, 1(1):60–64, 2013.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, et al. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Dallas Card, Peter Henderson, Urvashi Khandelwal, et al. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7513–7523. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.745. URL https://aclanthology.org/2020.emnlp-main.745/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, et al. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

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

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, et al. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. doi: 10.48550/arXiv.2107.07002. URL https://arxiv.org/abs/2107.07002.

Jesse Dodge, Suchin Gururangan, et al. Show your work: Improved reporting of experimental results. In *EMNLP*, 2019.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Ron Dror, Gaurav Baumer, Segev Shlomov, and Roi Reichart. The Hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

EleutherAI. lm-evaluation-harness. Software repository: https://github.com/EleutherAI/lm-evaluation-harness, 2023.

Jack Gallifant, Majid Afshar, Saleem Ameen, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*, 31(1):60–69, 2025. doi: 10.1038/s41591-024-03425-5. URL https://doi.org/10.1038/s41591-024-03425-5.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of Item Response Theory*. SAGE, 1991.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't Pass@k: A Bayesian framework for large language model evaluation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of NeurIPS*, 2021.

Hugging Face. Open LLM leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Jonas Kübler, Kailash Budhathoki, Matthäus Kleindessner, Xiong Zhou, Junming Yin, Ashish Khetan, and George Karypis. When LLMs get significantly worse: A statistical approach to detect model degradations. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Dawei Li, Bohan Jiang, Liangjie Huang, et al. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

Percy Liang, Rishi Bommasani, et al. Holistic evaluation of language models, 2022. arXiv:2211.09110.

Siqi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. Re-evaluating open-ended evaluation of large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Frederic M Lord. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1980.

Jum C Nunnally and Ira H Bernstein. *Psychometric Theory*. McGraw-Hill, 1994.

Nearchos Potamitis, Vansh Ramani, Har Ashish Arora, Dhairya Kuchhal, Lars Klein, and Akhil Arora. ReasonBENCH: Benchmarking the (in)stability of LLM reasoning. *arXiv preprint arXiv:2512.07795*, 2025. doi: 10.48550/arXiv.2512.07795. URL https://arxiv.org/abs/2512.07795. Also presented at the ICLR 2026 P-AGI workshop; not a main-conference paper.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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

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

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.16049. URL https://arxiv.org/abs/2310.16049.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. arXiv:2206.04615.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Steven K. Thompson. *Sampling*. John Wiley & Sons, 3rd edition, 2012.

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

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Guanhua Zhang, Ricardo Dominguez-Olmedo, and Moritz Hardt. Train-before-test harmonizes language model rankings. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

A PROOFS AND VARIANCE DERIVATION

The worked example fixes the pairwise denominator; the proofs then establish non-identification (P1), uniform-subset unbiasedness (P2), and post-selection failure (P3), and the final subsection derives the finite-cohort variance that remains under uniform sampling. In that derivation, n and m are the frame and target-set sizes, $N = \binom{n}{2}$ and $K = \binom{m}{2}$ are their pair counts, and π, π_1, and π_0 are the one-pair, shared-model-pair, and disjoint-pair inclusion probabilities. The empirical frame value and target-set dispersion are reported in Table 3 and Figure 2.

Worked micro-example. Suppose $\theta = 3$ and a cohort of three models records scores $\{40, 41, 50\}$ in percentage points. The three pairs have gaps $\{1, 10, 9\}$, two pairs exceed θ, and $d(B, C, t, \theta) = 2/3$. The same statistic returns $0/3$ on $\{40, 41, 42\}$ and $0/3$ on $\{50, 51, 52\}$; the difference in the first example is due to the score vector, not a changed denominator.

A.1 PROOF OF PROPOSITION 1

Let $R = \max_{i \in M} a_i(B, t) - \min_{i \in M} a_i(B, t)$ and suppose $R > \theta$. Take $i^* \in \arg\max_{i \in M} a_i$ and $j^* \in \arg\min_{i \in M} a_i$ and set $C_1 = \{i^*, j^*\}$. Then $\binom{|C_1|}{2} = 1$ and the single pair has gap $R > \theta$, so $d(B, C_1, t, \theta) = 1$. If additionally there exist $i', j' \in M$ with $|a_{i'} - a_{j'}| \leq \theta$, set $C_2 = \{i', j'\}$; the single pair does not exceed the threshold under the strict rule, so $d(B, C_2, t, \theta) = 0$. Both C_1 and C_2 are sub-cohorts of the same M at the same (B, t, θ), so d is not a function of (B, t, θ) and is therefore not identifiable from the triple (B, t, θ) alone.

Non-monotonicity under cohort inclusion. The hypothesis above does not by itself force a strict-non-monotone example on a single M, so we give explicit nested constructions. With scores $\{0, 1, 4, 5, 6\}$ and $\theta = 3$, the two-model cohorts $\{0, 5\}$ and $\{1, 4\}$ have $d = 1$ and $d = 0$, respectively, illustrating both extremes. The nested cohorts $\{0, 1\} \subset \{0, 1, 5\}$ have $d = 0$ and $d = 2/3$, so adding a model can increase d. Conversely, $\{0, 4, 5\} \subset \{0, 1, 4, 5, 6\}$ have $d = 2/3$ and $d = 5/10 = 1/2$, so adding models can decrease d. Thus no ordering of $d(B, C', t, \theta)$ and $d(B, C, t, \theta)$ follows from $C' \subset C$. □

A.2 PROOF OF PROPOSITION 3

Write $N = \binom{n}{2}$ and $K = \binom{m}{2}$, and let $y_p \in \{0, 1\}$ indicate separation for pair p in the frame F. For a uniformly random S of size m and any fixed pair p,

$$\mathbb{P}(p \subset S) = \frac{\binom{n-2}{m-2}}{\binom{n}{m}} = \frac{m(m-1)}{n(n-1)} = \frac{K}{N}.$$

Therefore

$$\mathbb{E}\big[d(B, S, t, \theta)\big] = \frac{1}{K} \sum_p y_p \, \mathbb{P}(p \subset S) = \frac{1}{K} \cdot \frac{K}{N} \sum_p y_p = \frac{1}{N} \sum_p y_p = d(B, F, t, \theta).$$

The lattice claim is immediate: the numerator is an integer k between 0 and K, so $d = k/K$; for $m = 5$, $K = 10$. □

A.3 PROOF OF PROPOSITION 2

Range restriction. If $\max_{i \in C} a_i(B, t) - \min_{i \in C} a_i(B, t) \leq \theta$, then every pair in C has an absolute gap of at most θ, the strict $>$ rule is never satisfied, the numerator of $d(B, C, t, \theta)$ is zero, and therefore $d(B, C, t, \theta) = 0$ for every (B, t).

The failure does not vanish in m. Fix $m \geq 2$, $\theta = 3$, and an integer k. Partition a frame into a top band of m models all scored 100 and $g = 25$ further bands of k models each, scored $0, 4, 8, \ldots, 96$. Consecutive bands are $4 > \theta$ apart, so a pair is separated exactly when its two members lie in different bands. With $n = m + 25k$ the separated pairs are the $25mk$ elite–rest pairs and the $\binom{25k}{2} - 25\binom{k}{2} = 300k^2$ cross-band pairs below, so

$$d(B, F, t, \theta) = \frac{25mk + 300k^2}{\binom{m+25k}{2}} \longrightarrow \frac{600k^2}{625k^2} = \frac{24}{25} \qquad \text{as } k \to \infty \text{ with } m \text{ fixed.}$$

Table 4. Pairwise score-gap distributions used to contextualize the 3-percentage-point operational threshold. Values pool within-cohort pairs separately by benchmark across the 133 retained snapshots; bins are $[0, 3)$, $[3, 5)$, and $[5, \infty)$ percentage points, so the boundary at 5 pp is included exactly once.

Benchmark	**Mean (pp)**	**Median (pp)**	$[0, 3)$ **pp**	$[3, 5)$ **pp**	$[5, \infty)$ **pp**
BBH	12.4	10.9	17.3%	9.5%	73.1%
GPQA	4.1	3.4	45.5%	22.3%	32.2%
MATH-Hard	10.5	7.5	24.9%	12.0%	63.1%
MUSR	5.0	4.2	36.5%	20.2%	43.3%

The top-m cohort C is the top band, whose range is $0 \leq \theta$, so $d(B, C, t, \theta) = 0$ by the preceding paragraph. Since m is arbitrary, no analogue of the $m \to n$ variance collapse of Proposition 3 applies. More generally g bands give $(g-1)/g$, so the attainable gap is bounded by how many threshold-separated bands fit in the score range, not by the cohort size: the date-averaged profiles in Appendix B.2 are descriptive and converge only in the full-frame limit, while a zero report remains available at every cohort size. □

A.4 Variance of a uniformly sampled target-set report

With the notation above, let $\pi = K/N$. For distinct pairs p, q, the probability that both are included depends only on $|p \cap q|$:

$$\pi_1 = \mathbb{P}(p, q \subset S,\ |p \cap q| = 1) = \frac{\binom{n-3}{m-3}}{\binom{n}{m}} = \frac{m(m-1)(m-2)}{n(n-1)(n-2)},$$

$$\pi_0 = \mathbb{P}(p, q \subset S,\ p \cap q = \emptyset) = \frac{\binom{n-4}{m-4}}{\binom{n}{m}} = \frac{m(m-1)(m-2)(m-3)}{n(n-1)(n-2)(n-3)}.$$

Then

$$\mathrm{Var}\big[d(B, S, t, \theta)\big] = \frac{1}{K^2}\Big[\pi(1-\pi)\sum_p y_p + \sum_{p \neq q} y_p y_q \big(\pi_{|p \cap q|} - \pi^2\big)\Big],$$

where $\sum_p y_p = N\, d(B, F, t, \theta)$ and $\pi_{|p \cap q|}$ equals π_1 for pairs sharing one model and π_0 for disjoint pairs.

Two consequences follow from this finite-population U-statistic variance structure (Cochran, 1977; Thompson, 2012). For $3 \leq m < n$, pairs that share a model contribute a nonnegative covariance term, strictly positive except when $(n, m) = (4, 3)$ because then $\pi_1 = \pi^2$; at the boundary $m = 2$, $\pi_1 = 0$ because no two distinct pairs can be jointly sampled. Disjoint pairs contribute a negative term under sampling without replacement. Thus the pair indicators are not generally exchangeable and the effective sample size need not equal the nominal pair count. As $m \to n$ we have $\pi \to 1$, $\pi_1 \to 1$, and $\pi_0 \to 1$, so the variance vanishes, recovering the frame value. This is the sense in which transport failure is a small-cohort phenomenon, and it is why canonical-order agreement recovers to 95.4% at size 50 in Section 5.3.

B Additional empirical validation

B.1 Threshold calibration and sensitivity

The threshold is an operational reporting rule, not a significance cutoff; the calibration table is included only to show how the fixed rule interacts with benchmark score scales.

Table 5 varies only θ on the reported score matrix and target-set draws. The signs of the C2 contrast and the selection-induced-compression versus dispersion decomposition are stable across $\theta \in \{1, 2, 3, 5\}$.

Across the grid, C1 decreases as θ rises, C2 remains negative for every benchmark, and the size-5 transport error increases. These checks support the mechanism-level conclusions without treating θ as a significance threshold.

Table 5. Threshold sensitivity. C1 is the date-equal-weighted full-cohort d_θ; C2 is the across-date top-five minus benchmark-native full-frame contrast (pp); |TE| is the size-5 mean absolute transport error (pp). Only θ enters; the reported target-set draws are reused so that the comparison isolates the threshold. Size-5 winner agreement at θ=3 is given in the main text.

Mechanism (metric)	Benchmark	θ=1	θ=2	θ=3	θ=5
C1 full d_θ (%)	BBH	90.2	83.6	78.4	68.9
	GPQA	80.9	65.7	52.3	31.6
	MATH-Hard	87.7	80.7	74.7	63.5
	MUSR	84.7	71.0	60.6	41.6
C2 Δ top-5 (pp)	BBH	−21.0	−30.1	−35.1	−42.5
	GPQA	−19.9	−25.8	−26.6	−21.1
	MATH-Hard	−4.5	−10.0	−14.1	−17.6
	MUSR	−21.4	−33.6	−36.6	−32.2
C3 |TE| size 5 (pp)	BBH	7.4	9.5	10.8	12.5
	GPQA	10.3	13.3	14.8	14.8
	MATH-Hard	8.5	10.4	11.7	13.4
	MUSR	8.9	12.0	13.3	14.7

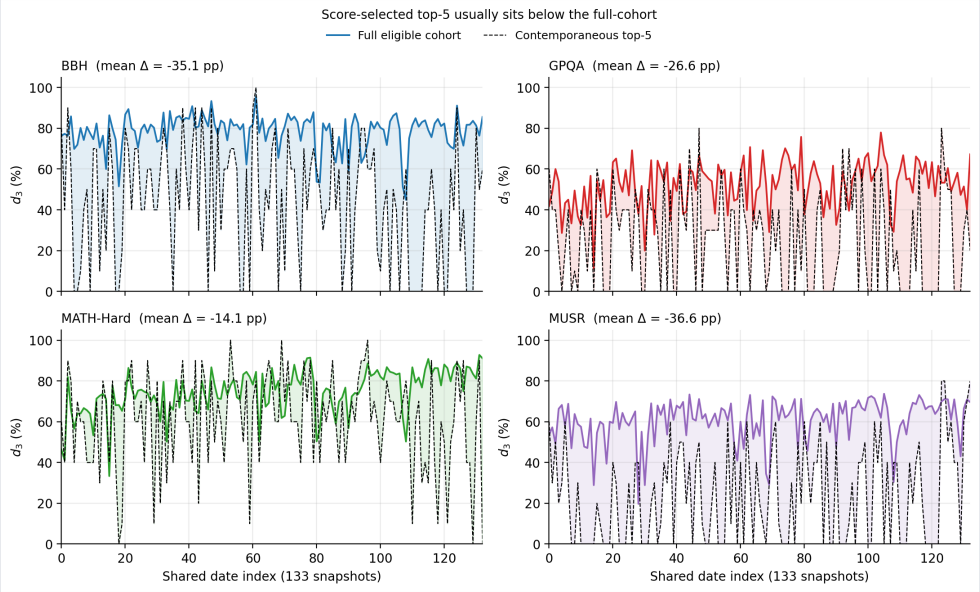

Figure 3. Score-selected top-five reports usually sit below full-cohort reports. Top-five values occur in increments of 10 percentage points because each report contains ten pairs. The horizontal axis is an ordered snapshot index, not a longitudinal time scale.

B.2 Top-K and target-set sensitivity

The top-K curves use $K \in \{5, 10, 20, 50\}$ on their own eligible-date sets. The target-set simulation uses 200 draws per date and size, and neither analysis interpolates unavailable dates or treats θ as a significance cutoff.

B.3 Illustrative target-set report

Table 6 gives the prespecified five-model target set and the benchmark-native full eligible frame on 2025-03-13. The fixed set is not a random sample.

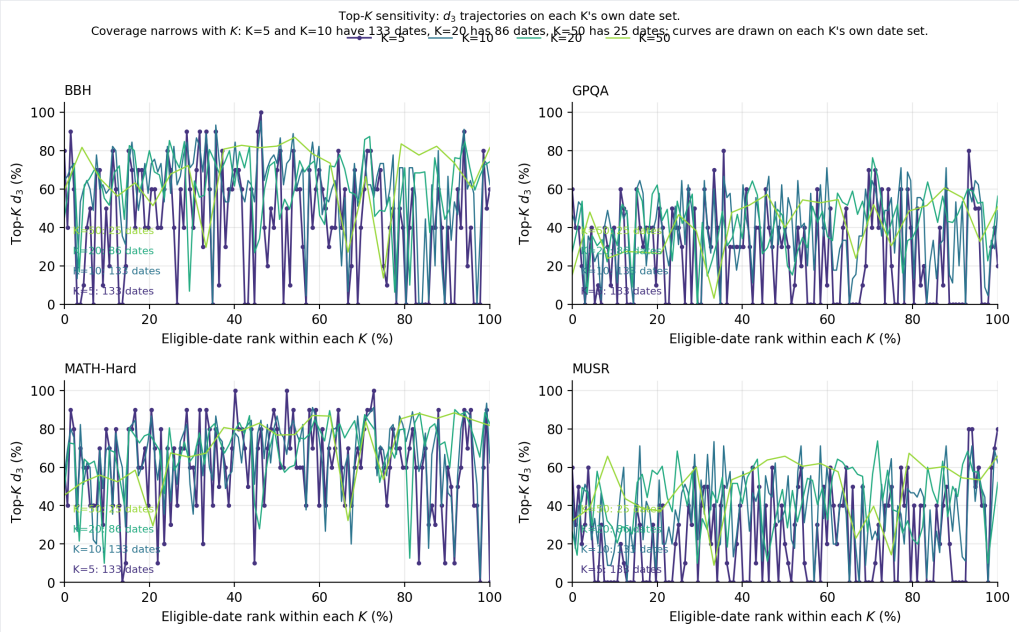

Figure 4. Top-K sensitivity: d_3 trajectories drawn on each K's own eligible-date set. $K = 5$ and $K = 10$ have 133 eligible dates, $K = 20$ has 86, and $K = 50$ has 25; unavailable dates are not interpolated.

Table 6. Illustrative target-set report on the 2025-03-13 snapshot, by benchmark. The target set is the same five models across benchmarks. The benchmark-native full eligible frame is the 35-model benchmark-specific frame on the same date.

Benchmark	Target d_3 (5 models, 10 pairs)	Full d_3 (35 models, 595 pairs)	Gap (pp)	Note
BBH	90.0%	85.5%	+4.5	high-spread cohort
GPQA	70.0%	67.2%	+2.8	intermediate
MATH-Hard	90.0%	91.1%	−1.1	top of range
MUSR	20.0%	69.6%	−49.6	narrow 5-model range

C Reporting contract

Table 7 lists the six fields a cohort-conditional separation claim should carry. For $m \leq 10$, the full separation matrix (≤ 45 pairs) is exhaustive, so the pair denominator makes small-cohort claims auditable without any new benchmark run.

Table 7. Reporting contract for cohort-conditional separation claims. The cohort, threshold, and pair-denominator fields are the minimum needed to audit the C1/C2/C3 mechanisms discussed above; the remaining rows reduce reporting ambiguity.

#	Field	Required specification
1	Benchmark	benchmark identity and score construction
2	Snapshot	archive date or contemporaneous snapshot rule
3	Cohort	membership and selection rule (full eligible, top-K, or target)
4	Threshold	value, unit, strictness ($>$ or $\geq$)
5	Cohort and pairs	cohort size and pair count (the denominator)
6	Score-selected?	whether the cohort was chosen using the reported score