Score Bridges: Auditing LLM Benchmark Version Migrations with Test Equating

Andrew Sheng-Han Huang
National Taiwan University

Abstract

Benchmark versions are routinely replaced, repaired, and re-scored, and a stale score from an old version can readily be placed beside a current score from its successor as though the two shared a scale. We audit what happens when that comparison is made, casting it as a score-linking problem with models as the common persons. On a version-locked MMLU-to-MMLU-Pro migration audit, our registered estimand finds that 40.5% of decisive naive comparisons reverse once the stale score is passed through a leave-one-model-out common-person bridge. This flip rate is, by our own characterization theorem, a measure of how far the two scales sit apart where comparisons actually happen — scale displacement, not a rate of wrong conclusions — and within each version the model orderings agree strongly (Kendall $\tau \geq 0.84$ across all audited panels): the hazard is the cross-version comparison itself, not a reshuffling of model quality. The flip rate survives family-aware, item-resampling, leave-family-out, add-back, category, bridge-choice, and error-scale-calibrated tie-band sensitivity analyses. A companion diagnostic on all ordered pairs, taking the current artifact's ordering as reference, falls from 42.4% to 6.0% after bridging. Two further audits support the finding without being pooled with it: a 343-model panel-expansion audit of the same v1 origin against the v2 leaderboard artifact reproduces the pattern (38.4%), and a 19-model GSM8K-to-GSM-Plus pilot is directionally consistent (21.0% point estimate; intervals cross the pre-registered bar at this small n). We provide a reproducible audit design with three verified protocol properties and two finite-sample results — an exact characterization of what the flip rate measures, and a distribution-free bound whose chief consequence is negative: no such certificate exists at the registered panel size — together with reporting rules and a transition-card template for maintainers and authors who cite cross-version deltas. All panels are selected public artifacts; we make no population claim about all models or all benchmark refreshes.

1 Introduction

Benchmark refreshes are now normal infrastructure. A benchmark may repair labels, replace items, rebalance categories, revise answer formats, or ship a new evaluation harness. MMLU-Pro, MMLU-Redux, and HumanEval+ illustrate different forms of this practice: a more demanding successor, a re-annotation effort, and a strengthened execution-based suite [7, 14, 24]. The usual comparison across such releases is numerically simple: put an older score beside a newer score and infer a difference in model quality. It is also unjustified unless the reported scores are linked to a common scale.

Score linking is the mature response to this problem in educational measurement. Its purpose is not to declare two forms identical, but to map scores so that a stated interpretation is preserved across forms [11]. Under the standard linking taxonomy [3, 9], a map between two instruments that differ in content *and* administration protocol is a concordance, not an equating in the strict sense; decades of concordance practice (for example ACT–SAT tables) exist precisely because institutions compared scores from different instruments as though they shared a scale. We use "equating" in the operational sense throughout and flag this taxonomy point explicitly: our bridges are concordance-grade score maps, and what is new here is not the bridge machinery but the audit estimand and decision protocol built on it. The same logic applies when models are the common persons and benchmark releases are the forms. A public leaderboard creates a particularly sharp exposure: a stale score for one model can be placed beside a current score for another model although each score was produced by a different versioned instrument. We do not quantify how often this comparison is made in the wild; the audit

quantifies its consequence when it is made.

This paper studies that failure mode as an *as-deployed migration audit*. It does not infer that a successor benchmark is intrinsically harder or that a within-version ranking has reversed. Instead, its estimand asks whether a naive stale-old-versus-current-new comparison agrees with the ordering after a bridge learned from common models. The distinction matters: our audited old and new sides differ in both content and scoring protocol.

Contributions. This paper makes four contributions.

1. We design a pre-specified *audit* of deployed benchmark-version migrations: a decision-level estimand, the decisive-pair flip rate; a leave-one-model-out evaluation protocol for concordance-grade common-person score bridges (LOMO); and a tie-band eligibility rule. The bridge machinery itself is textbook [11, 15]; the audit design and estimand are the contribution, and the estimand is robust to swapping the equipercentile bridge for a linear one (Section 4).
2. We establish three verified protocol properties (held-out bridges exclude held-out responses; the SEE tie band excludes ambiguous-by-declared-band comparisons; primary uncertainty resamples paired models) and two finite-sample results: an exact characterization of what the flip rate measures (Theorem 2) and a distribution-free sup-norm bound adapted from standard empirical-process tools [2, 16] whose chief consequence is negative — it certifies nothing at the registered panel size, making the small-panel cap arithmetic (Theorem 1).
3. We audit a public MMLU-to-MMLU-Pro migration. Among decisive naive comparisons, 40.5% reverse after bridging, while within-version orderings agree strongly ($\tau = 0.87$): the migration displaces the scale rather than reshuffling model quality. The registered flip rate is stress-tested against family structure, item sampling, a coverage-floor add-back, category stratification, bridge choice, and an error-scale-calibrated tie band. Two separately constructed audits then probe generality without pooling: a 343-model expansion against the v2 leaderboard artifact reproduces the pattern (Section 5.1), and a GSM8K-to-GSM-Plus pilot provides directionally consistent evidence on a different migration mechanism (Section 5.2).
4. We report a substantiated negative scope finding: on a clean-label refresh audit (MMLU to MMLU-Redux-2.0, 40 models), the registered flip rate is exactly 0.0% and the equipercentile bridge *degraded* held-out prediction (-46.8% vs. identity) — equating machinery can actively hurt where no scale shift exists. We translate both findings into operational reporting rules and a shipped transition-card template.

2 Related Work

Equating, concordance, and psychometrics. Classical test equating, scaling, and linking distinguish score conversion from the stronger claim that forms support interchangeable interpretations [3, 9, 11]; the SEE tradition for equipercentile conversions goes back to Lord [15]. The LLM-psychometrics literature has imported parts of this toolkit, and we differentiate per work. Federiakin [4] propose IRT-based leaderboard designs whose latent scale absorbs new items via within-frame calibration; they do not audit a deployed cross-version comparison of published observed scores. Zhou et al. [26] re-analyze benchmarks with IRT to critique item quality within a version. Li et al. [12] build adaptive testing over a fixed item bank. Ye et al. [25] survey this space and name cross-version comparability as open rather than solved. None of these ships a concordance-with-uncertainty analysis for a deployed version migration of published scores, which is the object audited here; our delta is the decision-level estimand and its protocol, not the use of psychometrics in LLM evaluation, which is by now established.

Cross-instrument comparison in NLP. Placing NLP test sets and models on common scales predates the current wave: Vania et al. [22] compare test sets with IRT using a shared model panel, and Rodriguez et al. [21] bring IRT to leaderboard analysis. Benchmark agreement testing [19] quantifies ordering disagreement between benchmark pairs over shared panels; our flip rate is likewise an ordering-disagreement statistic, but between two *decision rules* (naive versus bridged) on a version pair, under a pre-specified eligibility band, rather than between two benchmarks. Cross-benchmark

common-scale placement [10, 20] builds predictive bridges across instruments. Score linking across instruments is thus old in education and established in NLP; what is new here is the pre-registered audit estimand for a deployed version migration, its LOMO/SEE decision protocol, and the audited magnitudes.

Benchmark refreshes. MMLU established a broad multitask evaluation convention [8]. Its successors and repairs make versioning salient: MMLU-Pro extends the benchmark with a revised task design, MMLU-Redux corrects and re-annotates a subset, and HumanEval+ adds stronger tests to code-generation evaluation [7, 14, 24]. Successor papers typically publish paired old/new scores for a shared model panel and quantify the per-model shift — the GSM-Plus release [13] does exactly this — so the raw material for a bridge often ships with the successor. What they do not supply is the decision-level audit: whether a stale-versus-current comparison agrees with the bridged ordering under a pre-specified eligibility rule. That estimand, not the detection of a score shift, is this paper's object. Our clean-refresh audit (Section 6) is useful precisely because it separates a cleaning-style refresh from the composite migrations audited here.

Uncertainty in leaderboard conclusions. Recent work argues that evaluation claims should carry error bars, dependence diagnostics, and uncertainty-aware comparisons [17]. Benchmark compression likewise uses item information to estimate scores with fewer observations [20]. Those questions concern precision on a chosen scale. Score Bridges addresses a prior question: whether two published values are on a commensurate scale at all.

3 Methods

3.1 Common-person equipercentile bridge

Let x_m be model m's old-version score and y_m its new-version score. The paired models are the common persons. For a held-out model m, let $\widehat{F}_{\text{old},-m}$ and $\widehat{F}_{\text{new},-m}$ denote the empirical score distributions of all remaining paired models. The held-out equipercentile bridge is

$$\widehat{g}_{-m}(x) = \widehat{F}^{-1}_{\text{new},-m}\left(\widehat{F}_{\text{old},-m}(x)\right). \tag{1}$$

This conversion preserves percentile position in the observed common-person panel. It is an operational bridge, not a causal decomposition of why the two releases differ.

Our outcome is a stale-old-versus-current-new comparison. For an ordered pair (i, j), the naive decision uses the sign of $x_i - y_j$; the equated decision uses the sign of $\widehat{g}_{-i}(x_i) - y_j$. The latter evaluates the old score without using model i's own new-version responses. We call a pair a misordering when the two decisions disagree after the tie rule below.

3.2 LOMO protocol and no-leakage property

Protocol Property 1 (LOMO equating excludes held-out new responses)**.** *For any held-out model m, the bridge $\widehat{g}_{-m}$ in Equation (1) is measurable with respect to paired responses of models other than m. In particular, it does not use m's new-version item responses.*

Proof. Both empirical distributions in Equation (1) are constructed after deleting m. The inverse empirical distribution is a deterministic function of the retained new scores. Composing it with the retained old distribution therefore cannot depend on any deleted response. Applying the resulting function to x_m uses the held-out old score only. □

The purpose is predictive discipline, not a claim of independent models. Without deletion, a model can help determine the map used to convert its own old score, making the bridge appear more accurate than a score conversion available for a genuinely stale entry.

Scope of the tie band. LOMO guarantees leakage-free bridge predictions only. The registered primary estimand uses a full-panel SEE tie band, so a held-out model's new-version responses can affect its eligibility band even though they do not enter its bridge prediction. We retain that registered definition and report a fold-specific-SEE robustness check below; it is not a claim that the full primary pipeline is leakage-free.

3.3 SEE tie band and counted decisions

Let s_{ij} be the estimated SEE for the equated contrast between the old score of i and new score of j. With pre-specified multiplier 1.0, define the eligibility indicator

$$A_{ij} = \mathbf{1}\left\{\left|\widehat{g}_{-i}(x_i) - y_j\right| > 1.0 s_{ij}\right\}. \tag{2}$$

Only pairs with A_{ij} are counted. This makes the estimand explicitly conditional on being distinguishable at the bridge's stated error band.

Protocol Property 2 (SEE tie exclusion prevents boundary misclassification)**.** *For every counted pair, the equated contrast is separated from zero by more than its declared SEE band. Consequently, no pair whose equated sign is ambiguous under that band is included in the discordance numerator or denominator.*

Proof. Membership in the counted set is exactly the strict inequality in Equation (2). Its complement includes every contrast whose absolute value is at most the band, including zero. The discordance indicator is evaluated only after multiplication by A_{ij}, so those boundary cases contribute nothing. □

3.4 Model-level uncertainty

The scientific unit for external generalization is a paired model record, including its old score, new score, family identity, and response-derived bridge inputs. The primary interval therefore resamples model records and recomputes the bridge and pairwise estimand in every replicate. Item resampling is retained as a sensitivity analysis because it describes finite-item perturbations conditional on this panel.

Protocol Property 3 (Paired-model bootstrap targets the stated estimand)**.** *Under exchangeability of paired model records for the target public panel, a model-level bootstrap propagates uncertainty from which paired models are observed into both bridge fitting and cross-version comparison. An item-only bootstrap conditions on the observed model panel and cannot provide that model-population uncertainty.*

Proof. The estimand is a functional of the empirical distribution of paired model records: deleting a record changes the bridge and the ordered comparisons. Resampling those records perturbs this empirical distribution and recomputes the full functional. Resampling items while holding records fixed perturbs score measurement only, leaving the model empirical distribution unchanged. Thus the latter is a different, conditional target. □

3.5 Finite-sample accuracy of the bridge

The protocol properties above discipline *how* the bridge is used. A separate question is what accuracy the bridge itself can be guaranteed at a given panel size. The tools are standard: the object $\widehat{g}_n = \widehat{F}^{-1}_{\text{new}} \circ \widehat{F}_{\text{old}}$ is a monotone reparametrization of the two-sample shift function, for which distribution-free bands go back to Doksum [2], and our bound composes the DKW inequality with Massart's constant [16] with an inverse-Lipschitz step; the contribution is the adaptation to the deployed interpolated estimator and the activation-threshold corollary, not new machinery. Let $g = F^{-1}_{\text{new}} \circ F_{\text{old}}$ denote the population bridge and $\widehat{g}_n$ its plug-in estimator built from the two marginal empirical distributions of n paired models.

Theorem 1 (Finite-sample sup-norm bridge error)**.** *Assume the paired records are iid, and that F_{new} admits a density $f_{\text{new}} \geq c > 0$ on an interval extending η beyond $[F^{-1}_{\text{new}}(p_{\text{lo}}), F^{-1}_{\text{new}}(p_{\text{hi}})]$ on each side, for levels $0 < p_{\text{lo}} < p_{\text{hi}} < 1$. Let $\bar{\varepsilon}_n(\delta) = \sqrt{\log(4/\delta)/(2n)}$ and require the strict margin $\bar{\varepsilon}_n(\delta) < c\,\eta$. Then with probability at least $1 - \delta$,*

$$\sup_{x \in I_n} \left|\widehat{g}_n(x) - g(x)\right| \;\leq\; \frac{2}{c}\sqrt{\frac{\log(4/\delta)}{2n}}, \qquad I_n = \left\{x : F_{\text{old}}(x) \in [\, p_{\text{lo}} + \bar{\varepsilon}_n,\; p_{\text{hi}} - \bar{\varepsilon}_n \,]\right\}.$$

Proof sketch. Apply the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant to both marginal empirical distributions ($\delta/2$ each) and decompose $\widehat{g}_n - g$ into a quantile-estimation term and a level-estimation term; a quantile-stability lemma and the $1/c$-Lipschitz property of F_{new}^{-1} on the density-floored window each contribute $\bar{\varepsilon}_n/c$. Appendix C gives the lemmas, the boundary analysis at generalized-inverse edges (where the margin must be strict), and the version for the deployed mid-rank interpolated estimator, whose radius inflates to $\bar{\varepsilon}_n + \frac{1}{2n}$. Each leave-one-model-out bridge satisfies the same bound with $n-1$ in place of n. □

The bound is distribution-free apart from the density floor c on the occupied score interval, a standard equating assumption. Its most useful consequence is negative and exact:

Corollary 1 (Activation threshold; raw, full-panel estimator)**.** *The hypotheses of Theorem 1 (density floor, strict margin, nonempty I_n) can be met by some score distribution only if $\bar{\varepsilon}_n(\delta) < \frac{1}{4}$, equivalently $n > 8\log(4/\delta)$: the interior window must carry probability mass at least $2\bar{\varepsilon}_n$ and each density-floor flank mass exceeding $\bar{\varepsilon}_n$, so total probability forces $4\bar{\varepsilon}_n < 1$. At $\delta = 0.05$ the raw full-panel certificate first activates at $n = 36$. At the audited panel size $n = 21$, no instantiation exists at 95% confidence for any score distribution; the attainable confidence has unattained supremum of about 71.0%.*

Scope of the activation numbers. Corollary 1 is stated for the raw full-panel estimator. The same mass accounting with the appropriate radius gives stricter thresholds for the quantities actually deployed: the implemented interpolated estimator first activates at $n = 39$; a single implemented LOMO fold at panel size $n = 40$; and simultaneous control of all n folds at $n = 71$ (raw) or $n = 75$ (implemented). The paper's sixty-model evidence bar is therefore *not* presented as a simultaneous deployed-LOMO certificate; it is independently motivated by panel selection and family dependence, and Corollary 1 shows even the weakest (raw, full-panel) form of the guarantee is unavailable at the audited $n = 21$. Expanding the panel is a qualitative change in what can be guaranteed, not merely a precision improvement: at $n = 60$ the raw 95% certificate has half-width $(2/c)\,\bar{\varepsilon}_{60}$ with $\bar{\varepsilon}_{60} = 0.191$. Empirical coverage of the bound and its looseness (DKW-based certificates are conservative) are reported in Appendix C. Two practical notes: for any pair whose equated contrast exceeds the bound, the equated decision sign provably matches the population-bridge sign (decision stability); and the operational precision statement for this panel remains the bootstrap SEE; the theorem states what is certifiable without distributional knowledge, and the two are complementary.

3.6 What the flip rate identifies

The registered flip rate is not a generic disagreement rate; it identifies a specific geometric functional of the bridge's deviation from the identity map. Write $\widehat{y}_i = \widehat{g}_{-i}(x_i)$ for the LOMO prediction and s_i for the SEE at x_i (strictly positive under the registered noise-floor setting), and for each model define the *displacement band* $B_i = [\, x_i \wedge \widehat{y}_i,\ x_i \vee \widehat{y}_i \,]$ and the *tie collar* $C_i = [\, \widehat{y}_i - \kappa s_i,\ \widehat{y}_i + \kappa s_i \,]$ with $\kappa = 1.0$.

Theorem 2 (What the flip rate measures)**.** *For every tie-band-eligible ordered pair (i, j), the naive and equated decisions disagree if and only if $y_j \in B_i$. Consequently, whenever at least one pair is eligible, with ν_i the empirical distribution of $\{y_j : j \neq i\}$,*

$$\textit{flip rate} = \frac{\sum_i \nu_i(B_i \setminus C_i)}{\sum_i \nu_i(\mathbb{R} \setminus C_i)} :$$

the eligibility-weighted mass of comparison targets falling inside the bands swept between stale and bridged placements. (With no eligible pair the implementation reports zero, and the ratio is read with the same convention.)

Proof. Fix an eligible pair and write $a = x_i$, $b = \widehat{y}_i$, $t = y_j$. Eligibility gives $|b - t| > \kappa s_i \geq 0$ strictly, so $t \neq b$ and $\operatorname{sgn}(b - t) \neq 0$. If $t \in B_i$ then either $t = a$, making $\operatorname{sgn}(a - t) = 0 \neq \operatorname{sgn}(b - t)$, or t lies strictly between a and b, making the two signs opposite nonzero; both disagree. If $t \notin B_i$, both differences are nonzero with equal sign. The ratio form follows by counting: eligibility is exactly $y_j \notin C_i$, and $b \in C_i$ always, so the collar exclusion subsumes $t \neq b$. □

Theorem 2 fixes the interpretation of the headline value: among the 375 decisive ordered pairs, 40.5% have their comparison target inside the displacement band — the naive and equated decisions yield different signs (opposite nonzero signs in the band's interior, or tie-versus-non-tie at the x_i endpoint; the audited panel had no endpoint hits). Three consequences follow. If the bridge agrees with the identity at every occupied old score, the flip rate is exactly zero; the machine check on the audited panel confirms this. Holding the eligibility pattern fixed, pointwise enlargement of the displacement bands can only increase the flip count (the unconditional map from displacement to flip rate is not monotone, and we do not claim it). And the flip rate is an empirical, collar-conditioned occupancy functional of displacement from the *identity convention* — not a measure of bridge error against truth: a perfectly estimated bridge on a genuinely shifted migration *should* produce a high flip rate. On the audited panel the band characterization reproduces the registered count 152/375 exactly, with zero boundary-case occurrences.

4 The MMLU-to-MMLU-Pro Audit

4.1 Panel construction and estimand

The old side is the Open LLM Leaderboard v1 per-item MMLU artifact; the new side is the TIGER-AI-Lab published MMLU-Pro evaluation artifact. We canonicalized model names, joined paired records using text hashes, and required a pre-intersection coverage rate of at least the declared coverage floor. The dense intersection contains 3931 new-side items for 21 paired open-weight models; the old side contains 11338 items. The coverage rule removes a near-floor record before forming the dense matrix. The generated panel table below summarizes family counts without exposing per-model scores.

The audit's primary unit is an ordered pair of models. There are 420 ordered pairs, of which 375 remain outside the pre-registered SEE band. The primary comparison is intentionally asymmetric: a reader with an old score for model i asks how it compares to a current score for model j. It is not a within-version rank-reversal claim.

Terminology. Two quantities recur and must not be conflated. The *primary flip rate* (40.5%; 152/375) is the fraction of *decisive* ordered pairs—those outside the SEE tie band—whose naive decision disagrees with the equated decision. The *all-pair error* pair (42.4% → 6.0%; 178/420 → 25/420) compares each decision rule against the *realized ordering* of the paired panel over *all* ordered pairs, without tie filtering — where the realized ordering is $\mathrm{sgn}(y_i - y_j)$, both models' actual new-version scores. Taking that ordering as the reference is a convention (the current artifact's own ranking), not a claim that the new version measures true quality; and because the bridged placement $\widehat{g}_{-i}(x_i)$ is precisely the bridge's held-out estimate of y_i, the equated all-pair error is close to a re-expression of held-out bridge fit and is favorable to bridging largely by construction. The flip rate is the registered estimand and receives the full robustness grid; the all-pair pair is a same-denominator before/after diagnostic. Table 1 summarizes the panel.

4.2 Headline audit result

Among decisive naive comparisons, 40.5% reverse after bridging; the companion all-pair error falls from 42.4% to 6.0%. Both are properties of the audited public migration, not evidence that the new benchmark alone changed difficulty. Two diagnostics fix what these numbers mean. First, the within-version orderings of the panel agree strongly across versions (Kendall $\tau = 0.87$ between old-side and new-side aggregate-score orderings): the migration *displaces the scale* far more than it reshuffles model quality, exactly the reading Theorem 2 assigns to the flip rate. Second, the estimand is robust to the bridge family: replacing the equipercentile bridge with the linear one changes the flip rate from 40.5% to 42.7%, so the headline is not a product of the equipercentile machinery. On held-out prediction the linear bridge is in fact slightly better here (83.1% vs. 80.7% RMSE reduction relative to identity); we retain equipercentile as the registered bridge because it was pre-specified and assumption-free about the score relation's shape, and we report the linear bridge throughout as the ablation. Figure 1 shows why comparing unlinked raw scores can produce a reversed decision; Figure 2 shows the conversion evaluated by the audit.

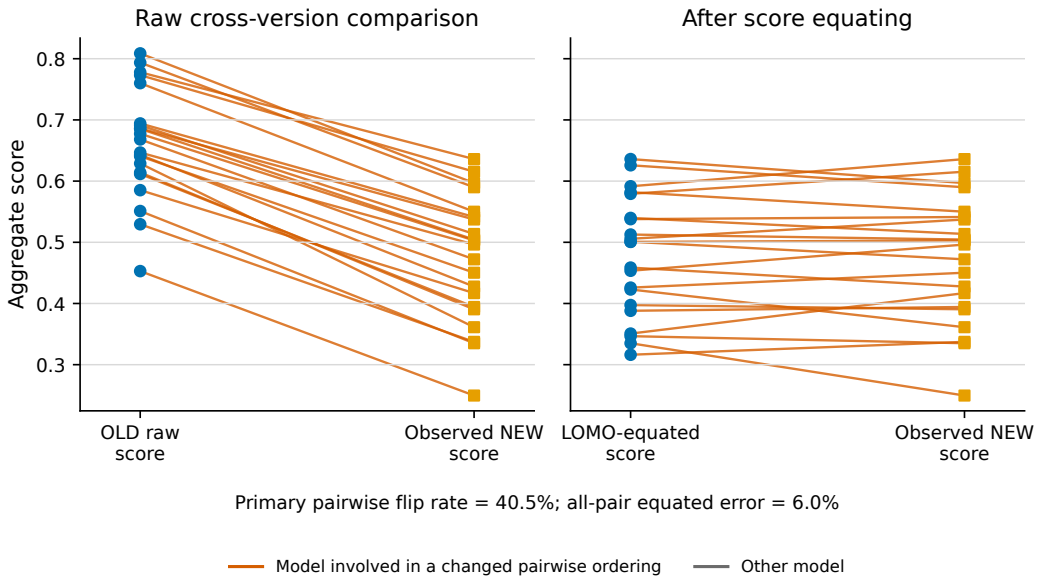

Figure 1: Cross-version comparisons can cross when an old-scale score is placed beside a new-scale score. The figure visualizes the audited stale-old-versus-current-new decision, not a within-version ranking claim.

4.3 Robustness grid

The primary interval is model-level. We report three resampling views of dependence and disclose the reliability of each. The iid model bootstrap spans [33.8%, 48.3%]. The family-cluster bootstrap spans [36.4%, 50.3%], but it resamples only 5 clusters — below any threshold for trustworthy cluster inference — and its lower bound in fact sits *above* the iid lower bound, so it cannot be defended as the more conservative interval; we report it as a dependence-aware view, not as primary protection. An organization-level clustering (8 clusters) gives [37.9%, 51.4%]. All three lower bounds sit far above the pre-registered bar, and none of these clusterings is guaranteed conservative at these cluster counts. Half-item resampling remains between 38.8% and 42.5%. Adding back the coverage-floor model gives a flip rate of 40.2% (167/415 decisive pairs on the 22-model panel); its all-pair companion is 42.0% → 6.7% (naive versus equated error against the realized ordering, 462 pairs) — the two metrics are reported separately and never joined. With LOMO predictions but a source-fold-specific SEE tie band, the flip rate is 40.6% (152/374); this near-null check addresses the leakage direction only — it varies *which* fold estimates the band, not the band's *scale*, which Section 5.3 treats separately. Widening the tie band itself barely moves the estimand: calibrating the collar to the held-out error scale (2.0×SEE; Section 5.3) gives 39.8% (349 decisive). Every *flip-rate* cell in the grid stays far above the pre-registered minimum-effect bar of 15.0% (the all-pair equated companion is a different metric and is not measured against that bar). Table 2 and Figure 3 report the full grid.

4.4 Category heterogeneity

The result is not driven by a single retained category. Category-level primary flip rates range from 18.8% in psychology to 49.0% in law, and equated misordering stays low throughout. These are descriptive stratifications of the same public artifact, not independently powered population estimates. Table 3 gives all categories, and Figure 4 makes the spread visible.

Table 1: Paired-panel summary by model family.

Model family	Models
Llama	7
Mistral	6
Qwen	3
Yi	2
other	3
Total paired models	21
OLD items	11338
NEW dense items	3931

Table 2: Run 3 sensitivity grid for the primary flip rate.

Cell	Point estimate	Interval or range
Baseline	0.4053	IID CI [0.3375, 0.4829]
Family-cluster bootstrap	0.4053	CI [0.3645, 0.5032]
Leave-family-out	–	[0.3863, 0.4390]
Item half-subsample	0.4070	[0.3883, 0.4251]
Add back coverage-floor model	0.4024	point estimate
Per-category	–	[0.1879, 0.4896]
Fold-specific SEE	0.4064	152/374 decisive pairs

Table 3: Per-category primary flip rates and equated misordering.

Category	Flip rate	Equated misordering	NEW items
business	0.3150	0.1095	287
computer science	0.3639	0.1071	129
economics	0.2559	0.1167	506
engineering	0.4699	0.1167	158
history	0.3919	0.0643	190
law	0.4896	0.1238	551
math	0.1994	0.1548	139
other	0.4034	0.0690	637
philosophy	0.4476	0.0833	322
physics	0.3592	0.1262	336
psychology	0.1879	0.1095	676
excluded-domains-aggregate	0.2957	0.0929	813

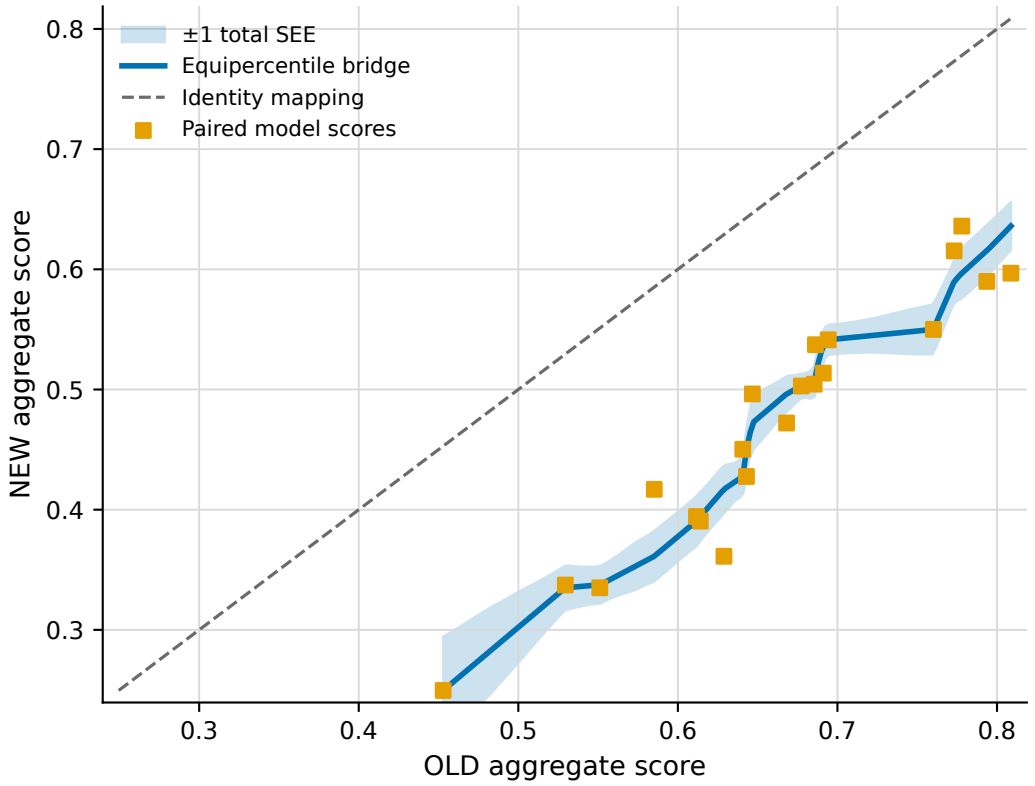

Figure 2: The leave-one-model-out bridge maps an old-version score to the new-version scale using only other paired models.

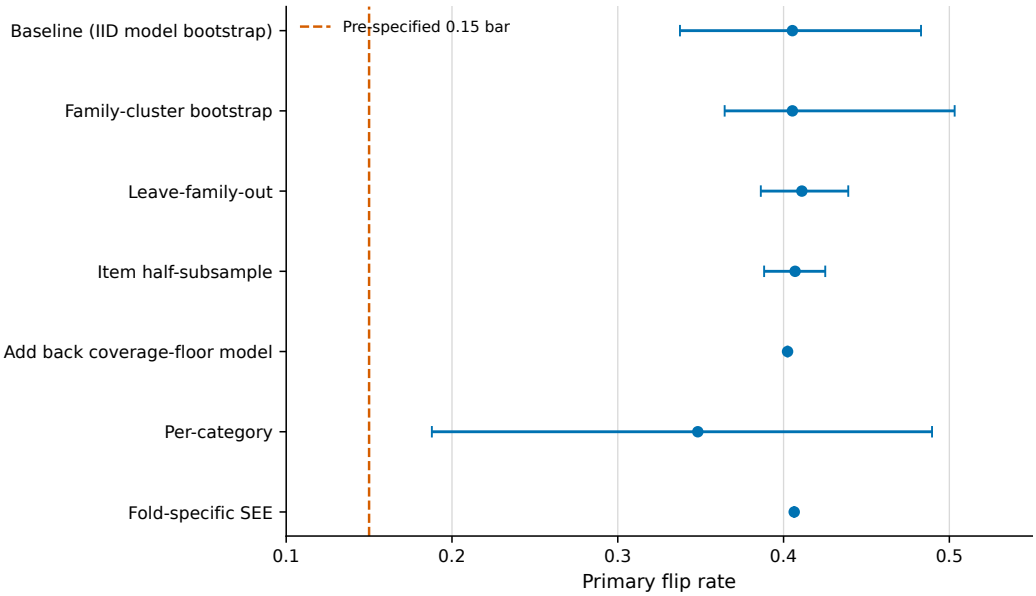

Figure 3: Sensitivity analysis across resampling, family, coverage, and category perturbations. The vertical reference is the pre-registered minimum-effect bar.

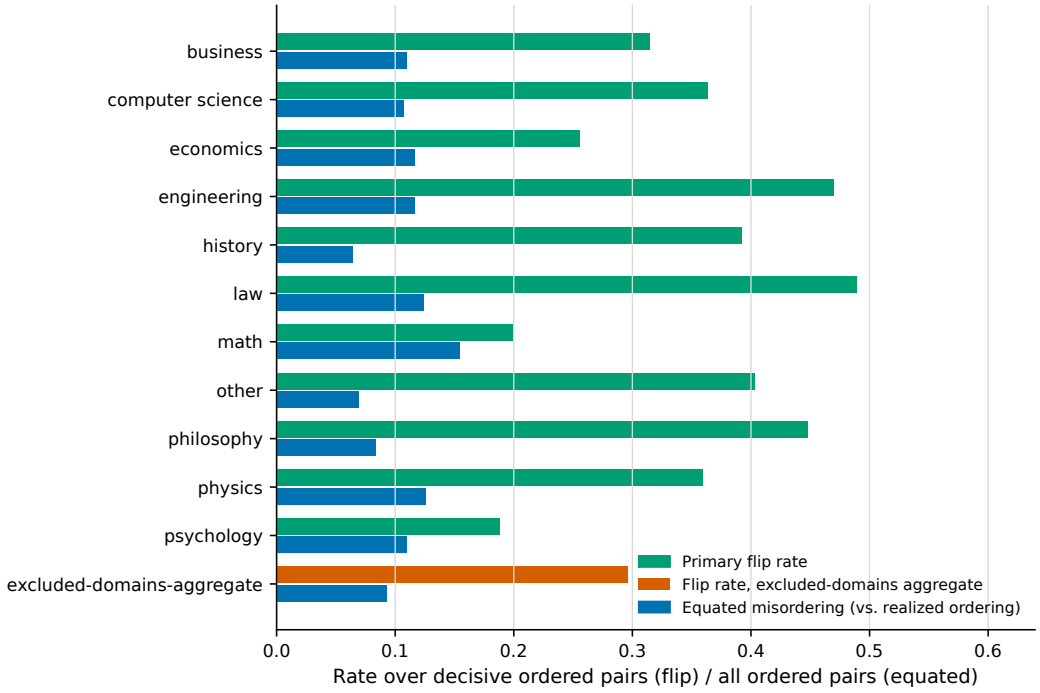

Figure 4: Per-category primary flip rate (green; excluded-domains aggregate hatched) and equated misordering against the realized ordering (blue) in the audited dense intersection.

5 Panel Expansion and Replication Audits

The registered audit of Section 4 is deliberately locked to its pre-registered 21-model panel. This section reports two additional, separately constructed audits that test whether its finding survives more models and a different migration type. Neither audit modifies the claim lock; each is analyzed with the same frozen statistical core, the same estimand construction, and seed 20260711. The three audits concern three *different* as-deployed migrations with heterogeneous scopes, and are never pooled: the registered audit pairs loglikelihood MMLU scoring with a generation-plus-extraction MMLU-Pro artifact (content and protocol both change); the expansion audit pairs two harness-produced artifacts of different item banks and harness versions; and the GSM-Plus audit compares binary-correctness records from a single archive whose old- and new-side evaluation pipelines are documented only at the archive level. Known changes are stated per audit rather than assumed identical across audits.

5.1 Panel expansion: the same origin against the v2 leaderboard artifact (343 models)

The Open LLM Leaderboard v2 [5] publishes gated per-item MMLU-Pro details; precise artifact provenance (repository names, revisions, and result-file hashes) is frozen in the audit manifests that accompany this paper, rather than in the leaderboard's mutable front page. Exact canonical name matching between v1 and v2 details repositories yields 397 candidate models; 345 had complete per-item artifacts on both sides, and 2 were removed by a doc-id universe check that caught truncated evaluation runs, leaving $n = 343$ paired models (old side 14042 MMLU items; new side 12032 MMLU-Pro items; no missingness by construction). The new side here is the `leaderboard_mmlu_pro` harness artifact, *not* the TIGER generation artifact of Section 4: this is a different composite migration of the same v1 origin, so agreement between the two audits is evidence that the misordering phenomenon is not an artifact of one particular new-side protocol — it is not a pooled estimate of a single quantity.

The pattern reproduces at 16 × the panel size. Among decisive comparisons, 38.4% flip after equating (42776/111512 of 117306 ordered pairs); naive all-pair error falls from 38.9% to 5.5%; the held-out bridge improves prediction RMSE by 86.7% over the identity mapping. The iid model-bootstrap interval narrows to [36.3%, 40.4%]. Two dependence-aware cluster bootstraps agree: resampling 8 name-based architecture families gives [36.2%, 45.5%], and resampling 137 publishing organizations gives [35.0%, 41.9%]; every lower bound stays far above the pre-registered 15.0% bar. The org-level clustering is the headline dependence check here (many clusters, no remainder bucket); the 8-family clustering is demoted to a coarse sensitivity view because its name-based assignment leaves a large unattributed remainder cluster (49% of the panel) and few-cluster bootstraps are not guaranteed conservative in either direction. The estimand is again robust to bridge choice (linear-bridge flip rate 35.3%) and to calibrating the tie collar to the held-out error scale (6.1×SEE: 36.1% over 91245 decisive pairs), and the within-version orderings again agree strongly across versions ($\tau = 0.84$).

At $n = 343$, the sample-size activation thresholds of Corollary 1 are all crossed (raw 36; implemented 39; simultaneous deployed-LOMO 75). Crossing the thresholds is a statement about n only, and we deliberately quote *no* plug-in certificate width here. The reason is directional: assumption (A3) requires a *lower* bound on the new-score density over the whole window, and every spacing-based value measured on this panel that exceeds the true floor produces a half-width that is *optimistically too small*. Concretely, the flattering proxies (average-density 2.13, median-local 4.38) yield half-widths of 0.075 and 0.036, while the one floor-respecting spacing proxy (widest-gap) yields a half-width of 2.0 — more than four times the occupied range 0.47, i.e. vacuous. A certified width would require a valid lower confidence bound on the density over the full window, which we do not have; until then the honest summary is that the sample-size conditions are met and the density-floor condition is unverified, so the distribution-free certificate remains unavailable in practice at every panel size used in this paper. The iid record assumption likewise remains an idealization here as in the registered audit.

5.2 Second migration type: GSM8K to GSM-Plus (19 models)

GSM-Plus extends GSM8K with adversarial perturbations of its problems [1, 13]; the GSM-Plus authors' public archive reports per-item binary correctness for 19 models on both sides (1319 old

items; 10552 new items; no missingness), including two closed API models alongside open-weight ones. This is an *adversarial-variant extension* — a third migration mechanism, distinct from both MMLU-to-MMLU-Pro audits — and its scale is pilot-sized. The point estimates are directionally consistent with the MMLU audits: 21.0% of decisive comparisons flip (65/309 decisive comparisons, among 342 total ordered pairs); naive all-pair error falls from 23.1% to 5.3%; the linear and equipercentile bridges improve held-out RMSE by 66.6% and 63.2%. The uncertainty, however, is wide at $n = 19$: the iid interval is [11.4%, 33.0%] and the family-cluster interval [9.4%, 39.8%], and *both lower bounds cross below* the pre-registered 15.0% bar. The point estimate is also the most sensitive of the three audits to the tie-band and bridge checks: calibrating the collar to the held-out error scale (2.3×SEE) gives 19.5% over 267 decisive pairs, and the linear bridge gives 16.7% — still above the bar as point estimates, but with no margin to spare. Within-version orderings again agree strongly ($\tau = 0.91$). We therefore read this audit as directionally consistent pilot evidence that the phenomenon extends beyond MMLU-style migrations — not as a demonstration that the minimum-effect threshold holds robustly for this migration. These numbers must not be pooled with either MMLU audit: the migration mechanism, item bank, and panel are all different.

5.3 What the large panel exposes about SEE

SEE under-covers on every audited panel, and this bears directly on the estimand's construction. On the registered panel itself, 2×SEE covers only 76.2% of held-out predictions; on the 343-model expansion panel the figure is 59.8% (and 89.5% on the GSM-Plus pilot). The shortfall was thus visible at the registered scale, not a large-panel discovery. Its consequence is not confined to prediction: by Theorem 2 the flip rate is a collar-conditioned functional, so an understated SEE narrows the tie collar, admits pairs into the decisive set that an honestly scaled band would have excluded, and — at a below-one-half baseline flip rate — biases the registered estimand *upward*. Protocol Property 2 guarantees only that counted pairs exceed the *declared* band; it cannot certify the band's scale. We therefore quantify the exposure directly: recomputing each audit with the collar inflated to its held-out error scale (SEE ×2.0/2.3/6.1 for the registered, GSM-Plus, and expansion audits) moves the flip rates to 39.8%, 19.5%, and 36.1% respectively — leaving every conclusion intact, with the thinnest margin on the GSM-Plus pilot. As to the source of the under-coverage: one *plausible hypothesis* is a population-invariance component (model-to-model heterogeneity around any single bridge, which need not shrink with n); the residual may also reflect bridge misspecification or bias, and these audits cannot identify the decomposition. SEE must not be read as a full predictive interval for an individual model's new score, and a band calibrated for the realized error scale — as in the sensitivity above — is the appropriate default for future audits.

6 Scope: When Equating Is Not Needed

Equating should not become a ritual attached to every dataset edit. Our first audit targeted two cleaning-style refreshes: MMLU to MMLU-Redux-2.0 [7] and GSM8K to GSM8K-Platinum [23], on a 40-model panel (seed 20260702, same frozen core and decision construction). The registered flip rate was exactly 0.0% on both migrations. More telling, equating actively hurt: on the Redux refresh the equipercentile bridge *degraded* held-out prediction RMSE by -46.8% relative to the identity mapping (the linear bridge improved it by only 3.5%), and the equated all-pair error (4.0%) was *worse* than the naive one (3.1%). This negative result is a scope boundary with teeth: when a refresh mainly repairs labels while preserving the effective score scale, fitting a bridge adds noise without removing bias, and raw comparison is adequate after the usual uncertainty checks.

The contrast with the present audit is instructive. A refresh can be cleaning-like, or it can combine a new item bank, prompt convention, answer format, decoder behavior, and scoring artifact. Only the latter composite case is supported here. We recommend measuring, rather than assuming, whether a bridge is needed for each migration.

7 Recommendations

For benchmark maintainers. Each public refresh should ship a version-transition card — an instance, for version migrations, of the documentation-artifact genre established by model cards and

datasheets [6, 18]. It should state the old and new scoring protocols, the model panel used as common persons, inclusion and coverage rules, a machine-readable equating table, held-out prediction diagnostics, and an uncertainty interval whose resampling unit matches the claimed population; prediction bands should be calibrated to the realized held-out error scale rather than the raw SEE (Section 5.3). The card should distinguish a content-only repair from a composite benchmark-plus-protocol migration. To make this adoptable rather than aspirational, the audit materials ship a card template, a worked example filled with the registered audit's values, and a single-file bridge tool in the supplementary directory;[1] the tool consumes two per-item files and emits the equating table, held-out diagnostics, and interval by calling the same frozen core used here. For closed API models, per-item outputs are generally unavailable and models are deprecated; there the card degrades gracefully to a maintainer-run common-person panel (the maintainer evaluates a fixed open panel on both versions and publishes the bridge), and where even that is impossible the honest card states that no bridge exists and stale scores should not be compared.

For model authors and leaderboard users. A cross-version delta should name both versions, state whether the scoring protocol changed, and either cite a maintained bridge or state that no bridge was available. Authors should not put an old raw score and a new raw score into a ranking claim without this disclosure. When a comparison is close to the bridge uncertainty band, the appropriate conclusion is unresolved rather than a forced ordering. Table-level results should report a conditional estimand: which pairs were eligible after the tie rule and what population the interval addresses.

For future audit studies. The next decisive data collection is a larger paired panel evaluated under one harness, prompt policy, extraction rule, and answer key — the single-axis contrast that would separate protocol effects from item-bank effects. We state plainly why this paper does not contain it: every audit here is zero-inference (it re-reads published per-item artifacts), and no public artifact pairs the same item bank under two protocols, or two item banks under one protocol, for a usable panel — the v2 leaderboard does not publish original-MMLU per-item results, so even the 343-model panel cannot isolate either axis without new model inference. The expansion panel does, however, fix the exact model set and item banks such a run should use. Until then, the useful claim is pragmatic: audited public migrations can make naive stale-versus-current comparisons misleading.

8 Limitations and Disclosures

Composite migration, not difficulty attribution. This is a benchmark-plus-protocol migration. The old side uses loglikelihood multiple-choice scoring, while the new side uses chain-of-thought generation with answer extraction. The audit therefore supports only an *as-deployed* migration claim; it cannot attribute the observed cross-version misordering to pure item-bank difficulty shift.

Selected public panel. The paired panel contains 21 open-weight models from the available recent public artifacts. It is a selective sample, with limited score-band and era coverage, so extrapolation to closed systems, later models, or another migration is not warranted. The expansion and replication panels of Section 5 are equally selective: the expansion panel is leaderboard participants (open-weight), and the replication panel is one archive's 19-model roster, which includes two closed API models. Larger n reduces sampling noise, not selection.

Artifact-dependent scoring and item intersection. New-side scores rely on TIGER official outputs. The dense intersection contracts the available item pool to 3931 items, and recorded answer-key conflicts were not independently adjudicated. The result should therefore be read as an audit of the published artifact, not a re-certified authoritative answer key.

Protocol revision and evidence scope. The estimand, population, minimum effect, kill condition, and headline were fixed in an internally version-locked, hash-pinned protocol *before* analysis — not deposited in a public registry, so we describe them as internally pre-specified rather than publicly pre-registered. One revision is disclosed: the protocol's original at-least-forty-model frame was revised by signed internal authorization to the present at-least-21-model public panel. Corollary 1 makes the smallness of that panel arithmetic rather than merely procedural: at $n = 21$ no 95% distribution-free

[1] `transition_card_template.md` and `make_transition_card.py`.

sup-norm certificate exists even for the raw full-panel bridge, and the simultaneous deployed-LOMO version first activates at $n = 75$. The 343-model expansion audit of Section 5.1 is an independent, non-pre-specified panel that satisfies the sample-size condition. Our evidence grade is therefore plainly stated: selected public panels of open-weight (and, in one pilot, two closed) models, heterogeneous composite migrations, and no population claim — these are observational audits, and we do not extrapolate beyond the audited artifacts.

Density-floor proxies and SEE scale. Spacing-based density values are plug-in proxies, not the population lower bound required by assumption (A3); proxies that exceed the true floor make plug-in widths optimistically small, the floor-respecting proxy renders the certificate vacuous on the expansion panel, and no certified guarantee is claimed at any n in this paper (Section 5.1). Separately, SEE under-covers held-out predictions on every audited panel (76.2% registered; 59.8% expansion; 89.5% pilot), which both disqualifies it as a predictive interval and — because the tie collar is built from it — biases the registered flip rate upward at a below-one-half baseline; the error-scale-calibrated sensitivity in Section 5.3 quantifies that exposure directly and leaves every conclusion intact, with the thinnest margin on the GSM-Plus pilot.

Dependence and interval choice. Model families are clustered, so an iid bootstrap can be over-optimistic in principle. But no clustering of the registered panel has enough clusters for trustworthy cluster inference (5 families; 8 organizations), and the family-cluster lower bound in fact exceeds the iid lower bound, so we claim no conservativeness for any interval; we report all three resampling views (Section 4) and note they agree in keeping every lower bound far above the bar. None of them converts this selected panel into an independent random sample of all language models.

Headline revision after cross-model derivation review. A cross-model derivation review found the prior headline invalid because it joined a tie-filtered naive–equated flip rate to an all-pair equated truth error rate. Under a signed, project-logged data-forced revision (`M14`, 2026-07-11), we replaced that cross-metric arrow with the same-denominator all-pair error comparison and report the registered flip rate separately.

9 Conclusion

Benchmark version migration is a measurement problem before it is a leaderboard problem. On one public, composite MMLU-to-MMLU-Pro migration, 40.5% of decisive naive comparisons reverse after bridging — and because within-version orderings agree strongly across versions ($\tau = 0.87$), that number measures scale displacement, not a reshuffling of model quality: the hazard is the act of comparing across versions, and it survives family, item, category, add-back, bridge-choice, and calibrated-band checks. The companion all-pair error falls from 42.4% to 6.0% against the current artifact's ordering. The pattern reproduces on a 343-model expansion against the v2 leaderboard artifact (38.4%), a GSM8K-to-GSM-Plus pilot is directionally consistent (21.0% point estimate, intervals crossing the bar at $n = 19$), and a clean-label refresh audit shows the flip rate is exactly zero where no scale shift exists — three different migrations and one negative control, never pooled, each with its uncertainty stated. Every panel is a selected public artifact and each migration carries its own composite of content and protocol changes; we claim no more than the audits show. The practical prescription is modest and actionable: publish a transition card (a template and tool ship with this paper), bridge stale scores before comparing them with current scores, and disclose when the data do not support a common-scale claim.

A Audit Replay Order

The audit is replayable in a fixed order: construct the paired artifact from published per-item records; apply the canonicalization and text-hash join (reporting the join rate before any statistic); enforce the coverage rule before forming the dense intersection; fit each bridge after leaving its target model out; impose the pre-specified SEE tie rule; and recompute the complete estimand inside every resample. This order matters. Changing the item intersection after inspecting a result, or fitting a bridge with the target model included, changes the estimand rather than merely improving its precision.

The public audit materials should preserve every decision required to replay that sequence: source identifiers and checksums, canonical model names, family labels, coverage diagnostics, the retained-question manifest, answer-key conflict records, prompt and extraction policies, and the seed used for resampling. A transition card that contains only aggregate scores cannot support a stale-score bridge, because it hides the common-person alignment and the source of the uncertainty interval.

B Reproducibility Statement

Every empirical statistic quoted in the prose, figures, and tables enters the manuscript through generated macros produced by one deterministic generator (`paper/scripts/make_artifacts.py`, seed 20260703). The generator recomputes the registered audit by calling the frozen statistical core (sha256 prefix 735bec0ce34a, cross-checked at generation time against the hash recorded inside the replication audit's result file) and reads the robustness grid and the Section 5 audit results verbatim from their frozen derived result files. Those derived inputs were each produced by seeded analysis scripts (seed 20260711) that call the same frozen core, and each analysis was executed twice with byte-identical result files before its numbers entered the generator; the generator itself regenerates byte-identical outputs across repeated runs, verified by hashing two independent executions. The expansion audit's result file does not embed the core hash internally; its provenance is established by the generation-time cross-check above and by the audit scripts pinned in the repository. Data-acquisition (fetch) scripts are resumable network code and are documented rather than claimed bit-reproducible. A machine-readable map from every macro to its value and source file accompanies the manuscript (`paper/generated/numbers.json`). The expansion audit retains no benchmark question text: per-item records are reduced to binary correctness at fetch time.

The estimand, population, minimum effect, kill condition, and headline were fixed in an internally version-locked, hash-pinned protocol before the manuscript was written (not a public registry; see Section 8), and one revision to the headline presentation—replacing an invalid cross-metric arrow with the same-denominator comparison reported here—was made under a signed, disclosed internal authorization following an adversarial cross-model derivation review (three review rounds; all findings closed). The two finite-sample theorems were added under a second signed, logged authorization and passed the same discipline: a dedicated adversarial re-derivation per theorem (whose first round produced counterexamples that forced strict margin conditions and an explicit score-support assumption), a full-suite re-derivation of all five formal results, and a post-fix confirmation pass. The three protocol properties and two theorems of Section 3 were thus independently rederived and probe-tested. The finite-sample bound was additionally stress-tested by a seeded coverage simulation (seed 20260711) whose configurations satisfy the theorem's margin conditions by construction (Appendix C). Per-item source data are public third-party artifacts; no new model inference was run for this audit.

C Finite-Sample Theory: Lemmas, Implementation, and Simulation

This appendix proves Theorem 1, states the version matching the deployed estimator, and reports the coverage simulation. Throughout, $F^{-1}(p) = \inf\{t : F(t) \geq p\}$ and $\bar{\varepsilon}_n(\delta) = \sqrt{\log(4/\delta)/(2n)}$.

Assumptions. (A0) scores lie in $[0, 1]$ almost surely (automatic for aggregate accuracies); (A1) the n paired records are iid; (A2) F_{old} and F_{new} are continuous; (A3) F_{new} is absolutely continuous on an open interval containing $J = [Q_{\text{new}}(p_{\text{lo}}) - \eta,\ Q_{\text{new}}(p_{\text{hi}}) + \eta]$ with density $f_{\text{new}} \geq c$ a.e. on J, where $Q_{\text{new}} = F_{\text{new}}^{-1}$. The bridge estimator uses only the two marginal empirical distributions, so no dependence assumption is made; the raw theorem uses only (A1) and (A3).

Lemma 1 (Two-sample DKW event)**.** *With probability at least* $1-\delta$*, both marginal empirical CDFs are within* $\bar{\varepsilon}_n(\delta)$ *of their population CDFs in sup norm.*

Proof. Massart's form of the Dvoretzky–Kiefer–Wolfowitz inequality gives $\Pr(\|\widehat{F} - F\|_\infty > \varepsilon) \le 2e^{-2n\varepsilon^2}$ for each side; $\varepsilon = \bar{\varepsilon}_n(\delta)$ makes each right side $\delta/2$, and a union bound (dependence between the marginals is irrelevant) completes the claim. □

Lemma 2 (Quantile stability; strict margin)**.** *Let* F *satisfy (A3) and let* G *be any CDF with* $\|G - F\|_\infty \le \varepsilon$ *where* $\varepsilon < c\eta$ *strictly. Then* $\sup_{p\in[p_{\mathrm{lo}},p_{\mathrm{hi}}]} |G^{-1}(p) - F^{-1}(p)| \le \varepsilon/c$.

Proof. Fix p, set $q = F^{-1}(p)$ and $a = \varepsilon/c < \eta$; by (A3), F is continuous and strictly increasing on J and $F(q) = p$. *Upper*: $q + a \in J$ and $F(q+a) \ge p + \varepsilon$ by integrating the floor, so $G(q+a) \ge p$ and $G^{-1}(p) \le q + a$. *Lower*: we show $F(x) < p - \varepsilon$ for every $x < q - a$, whence $G(x) \le F(x) + \varepsilon < p$ and $G^{-1}(p) \ge q - a$. If $x \in J$, then $[x, q] \subseteq J$ and $F(q) - F(x) \ge c\,(q - x) > c\,a = \varepsilon$. If $x < Q(p_{\mathrm{lo}}) - \eta$, monotonicity and the flank integral give $F(x) \le F(Q(p_{\mathrm{lo}})) - c\eta \le p_{\mathrm{lo}} - c\eta < p - \varepsilon$, using the strict margin. Strictness is essential: at $\varepsilon = c\eta$ a perturbing CDF can jump exactly to level p on a flat region outside J and the conclusion fails. □

Lemma 3 (Inverse Lipschitz)**.** *Under (A3),* $|Q_{\mathrm{new}}(p) - Q_{\mathrm{new}}(p')| \le |p - p'|/c$ *for* $p, p' \in [p_{\mathrm{lo}}, p_{\mathrm{hi}}]$.

Proof. Both quantiles lie in J and $|p - p'| = \int f_{\mathrm{new}} \ge c\,|Q_{\mathrm{new}}(p) - Q_{\mathrm{new}}(p')|$. □

Proof of Theorem 1. On the event of Lemma 1, for $x \in I_n$ the estimated level $\widehat{p} = \widehat{F}_{\mathrm{old}}(x)$ satisfies $|\widehat{p} - F_{\mathrm{old}}(x)| \le \bar{\varepsilon}_n$ and lies in $[p_{\mathrm{lo}}, p_{\mathrm{hi}}]$. Decompose $\widehat{g}_n(x) - g(x)$ into $[\widehat{F}^{-1}_{\mathrm{new}}(\widehat{p}) - Q_{\mathrm{new}}(\widehat{p})] + [Q_{\mathrm{new}}(\widehat{p}) - Q_{\mathrm{new}}(F_{\mathrm{old}}(x))]$; Lemma 2 (strict margin) bounds the first term by $\bar{\varepsilon}_n/c$ and Lemma 3 the second by the same. □

Proposition 1 (Implemented estimator)**.** *The deployed estimator (mid-rank plotting positions* $(i - \frac{1}{2})/n$*, linear interpolation, range clamping, final clamp to* $[0,1]$*) satisfies the same bound with* $\bar{\varepsilon}_n$ *replaced by* $r_n = \bar{\varepsilon}_n + \frac{1}{2n}$ *throughout (strict margin* $r_n < c\eta$*, interior margin, half-width* $2r_n/c$*), provided additionally (A0), (A2),* $p_{\mathrm{lo}} \ge \frac{1}{2n}$*, and* $p_{\mathrm{hi}} \le 1 - \frac{1}{2n}$.

Proof. Under (A2) sample values are a.s. distinct. The old-side tail-clamped plotting-position map (which need not be a CDF) is within $\frac{1}{2n}$ of the old ECDF everywhere, so implemented levels err by at most r_n and, on I'_n, stay in $[p_{\mathrm{lo}}, p_{\mathrm{hi}}]$. On the new side, define a proper CDF that is 0 below $y_{(1)}$, jumps to $\frac{1}{2n}$ at $y_{(1)}$, linearly interpolates the plotting positions, and jumps from $1 - \frac{1}{2n}$ to 1 at $y_{(n)}$; it is within $\frac{1}{2n}$ of the new ECDF, and its generalized inverse coincides with the implemented interpolated quantile on levels in $[\frac{1}{2n}, 1 - \frac{1}{2n}]$, which contains $[p_{\mathrm{lo}}, p_{\mathrm{hi}}]$ by the side conditions. Lemma 2 applies to that CDF with radius r_n, and the level term inherits the same radius. By (A0) every interpolated prediction lies between observed new scores in $[0,1]$, so the final clamp is inactive; without (A0) the clamp can force deterministic violations, so the support assumption is part of the statement. □

Each LOMO fold uses $n-1$ records, so its bound holds with $n-1$ in place of n; simultaneity over all n folds costs $\delta \mapsto \delta/n$. These substitutions produce the activation thresholds quoted in Section 3.5. The panel is not literally iid — model families cluster — so the bound is the independent-records benchmark, the same disclosure that makes the family-cluster interval primary. Aggregate scores are also discrete grid values; (A2) is an idealization, and a tie block of multiplicity m would inflate the $\frac{1}{2n}$ interpolation constant to $\frac{m}{2n}$ (no aggregate-score ties occur in the audited panel).

Empirical coverage. A seeded simulation (seed 20260711; 1000 replicates per cell; every bridge fit calls the frozen implementation) draws panels from known score distributions — a uniform pair with a linear bridge and a Beta-shaped new side with a nonlinear bridge — at configurations satisfying the strict margin by construction, and measures realized sup-norm error over I'_n against the Proposition 1 bound. All 4 feasible cells achieved full coverage (1000 of 1000; targets $1-\delta$), with median looseness factors between 3.3 and 9.4: DKW-based certificates are conservative, as expected. An independent grid scan over configurations (over two million per scenario) found none satisfying the hypotheses at $n = 21$, $\delta = 0.05$ — an empirical confirmation of Corollary 1 — and the nonlinear scenario is

additionally infeasible at $(n, \delta) = (60, 0.05)$ and $(21, 0.50)$ because the Beta density floor collapses over the η-extended interval; infeasible cells are reported as such rather than forced. Full tables ship with the audit materials.

References

[1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[2] Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2(2):267–277, 1974.

[3] Neil J. Dorans. Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4): 227–246, 2004.

[4] Denis Federiakin. Improving LLM leaderboards with psychometrical methodology. *arXiv preprint arXiv:2501.17200*, 2025.

[5] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM leaderboard v2. Hugging Face, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

[6] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[7] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with MMLU? *arXiv preprint arXiv:2406.04127*, 2025.

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

[9] Paul W. Holland and Neil J. Dorans. Linking and equating. In Robert L. Brennan, editor, *Educational Measurement*. Praeger, 4th edition, 2006.

[10] Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench: A sparse benchmark of reasoning and knowledge in large language models. *arXiv preprint arXiv:2407.12844*, 2024.

[11] Michael J. Kolen and Robert L. Brennan. *Test Equating, Scaling, and Linking: Methods and Practices*. Springer, second edition, 2004. ISBN 978-0-387-40086-0.

[12] Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V. Chawla. Adaptive testing for LLM evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*, 2026.

[13] Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. ACL Anthology 2024.acl-long.163; arXiv:2402.19255.

[14] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023.

[15] Frederic M. Lord. The standard error of equipercentile equating. *Journal of Educational Statistics*, 7(3): 165–174, 1982.

[16] Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

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

[18] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of FAT* 2019*, 2019.

[19] Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Do these LLM benchmarks agree? fixing benchmark evaluation with BenchBench. *arXiv preprint arXiv:2407.13696*, 2024.

[20] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating LLMs with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

[21] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of ACL-IJCNLP 2021*, 2021.

[22] Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In *Proceedings of ACL-IJCNLP 2021*, 2021.

[23] Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*, 2025.

[24] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

[25] Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*, 2026.

[26] Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*, 2025.