The Error Bar You Should Have Used: An Independence-Audited Meta-Analysis of the Published LLM Benchmark Record

Andrew Sheng-Han Huang
National Taiwan University

July 2026 — working draft

Abstract

When a paper, a leaderboard, or a vendor announcement states that a language model scores a given number on a benchmark, the uncertainty that matters to the reader is not the within-run standard error that the evaluator could have computed. It is the spread of what *different reporters* publish for the very same model–benchmark cell. We treat the published record itself as the measurement system under study. From a curated multi-source corpus of benchmark reports we build a source-linked panel covering 10 benchmarks and 154 model–benchmark cells with at least two distinct reporting sources, and we first pass it through a pre-registered *independence audit* that flags reports whose values are identical to the vendor's own claim—consistent with propagation rather than independent measurement: 33.7% of non-self-report observations in auditable cells carry such flags under our primary rule. On the audited panel we estimate, per benchmark, the between-source reproducibility standard deviation τ with a common-τ Paule–Mandel estimator whose small-k behaviour we validate by simulation before touching the data. Across benchmarks the median τ is 6.1 percentage points, and the median benchmark-specific ratio of τ to the item-sampling standard error implied by benchmark size—the error bar a well-behaved single report would carry—is 5.5. Under the fitted model, the 95% predictive band for the difference between two future single-source reports from independent sources covers zero for 97.9% of adjacent-rank consensus gaps on canonical orderings. We release the panel, the audit rules, and a per-benchmark *discount table*: the minimum published score difference that constitutes evidence under the record's own dispersion. All numbers in this paper are generated by scripts from the released data; every guarantee has an executable check that fails if it is false.

1 Introduction

A single number—"model M scores y on benchmark B"—is the atomic unit of evidence in contemporary language-model evaluation. Recent work has made real progress on the uncertainty *inside* one evaluation: standard errors across items [12], variance across repeated runs [2, 10], and the sensitivity of scores to harness and format choices under controlled re-runs [1]. All of these quantify what the *evaluator* could have reported. But the reader of a paper, a procurement memo, or a launch post is rarely the evaluator. The reader consumes the *published record*: a scatter of numbers produced by vendors, leaderboards, independent evaluators, and third-party papers, each under its own configuration, and—as we show—frequently indistinguishable in value from copies of one another.

Recent infrastructure work has documented, descriptively, that this record is inconsistent. The Evaluation Cards project reports that 98.2% of model–benchmark pairs in a large monitoring corpus are reported by only a single party, and that where multiple parties do report the same pair, the reports diverge substantially more than half the time [6]. Independent evaluators have published striking individual discrepancies between claimed and reproduced scores [4, 14]. What the descriptive layer does not provide is the *inferential* layer a reader needs: how wide is the dispersion, in points, per benchmark; how much of the apparent multiplicity of sources is real measurement rather than citation; and what score difference between two models survives that dispersion when each number the reader holds is a sample of one.

This paper supplies that layer. Our claim is deliberately narrow and falsifiable: on the benchmarks where the public record contains enough genuinely independent repeated measurements, the between-source reproducibility standard deviation exceeds the item-sampling standard error implied by benchmark size—per benchmark—by a factor large enough to overturn a substantial fraction of published adjacent-rank comparisons. Both thresholds—the factor and the fraction—were fixed in a pre-registered claim lock before the main analysis ran, together with the audit rules and the estimator; had the record turned out to be consistent, the honest-negative outcome was pre-committed for publication as such.

Contributions.

1. **A source-linked panel of the published record.** We consolidate 2426 score observations spanning 52 benchmarks from a curated multi-source corpus (vendor technical reports, independent evaluation organizations, and third-party papers), normalized, configuration-annotated, and released with per-row provenance.

2. **An independence audit and the propagation rate.** Pre-registered rules flag each non-self-report observation by whether its value is identical to the vendor's own claim (consistent with propagation) or numerically distinct. The propagation rate itself—33.7% under the primary rule, 36.1% under the permissive one (24.7% within the analysis cells)—is, to our knowledge, the first quantitative estimate of how much of the LLM benchmark record is value-identical to the vendor's claim, consistent with propagation rather than independent measurement.

3. **Validated between-source variance components.** A common-τ Paule–Mandel estimator, pooled across cells within a benchmark, with Q-profile intervals; its coverage at the small per-cell source counts typical of the record (k=2–8) is established by simulation *before* the estimator touches real data, and the simulation is shipped as an executable check.

4. **Decision artifacts.** Per-benchmark discount tables (the minimum published difference that excludes zero at 95% for a one-report-per-model reader) and adjacent-rank ambiguity rates; both are computed from the audited panel and regenerate from scripts.

5. **Restriction analysis.** Restriction analyses probe how dispersion responds to conditioning on disclosed configuration (shots, source type)—separating variance that is *attributable* in principle from variance that is *observable* to a reader, which is the quantity that governs trust in a single number.

2 Related work

Uncertainty within one evaluation. Miller [12] advocates item-level standard errors for evals; Blackwell et al. [2] and Madaan et al. [10] quantify run-to-run and seed variance under a fixed harness. These characterize the noise floor of a single measurement pipeline; our headline ratio's denominator is the item-sampling component of that floor, and our numerator is what none of them estimate—the spread across the published record.

Controlled sensitivity studies. Cross-harness and format-sensitivity experiments re-run models under systematically varied configurations [1, 16, 18]. Such studies measure what *could* vary under intervention. Our object is different: what *did* vary in the record the community actually consumes, including its dependence structure.

Reporting infrastructure and descriptive audits. Evaluation Cards [6] structures the ecosystem's reports and documents divergence prevalence; leaderboard operators have documented individual reproduction gaps [14, 4, 7]; re-scoring efforts have shown that re-grading a single benchmark reorders its leaderboard [17]. We rely on this descriptive layer for motivation and, in part, for data—and add the estimation layer it explicitly lacks.

Evidence synthesis and metrology. Random-effects meta-analysis and its small-sample corrections [5, 13, 9, 8] and interlaboratory reproducibility analysis under ISO 5725 [15] supply our statistical machinery; prediction-interval coverage at few studies is a known failure mode [11] which is precisely why our estimator ships with a pre-run simulation gate. Domain-specific meta-analyses of LLM performance pool effects across heterogeneous studies to compare models; our unit of synthesis is instead the repeated measurement of a *single* model–benchmark cell, and our estimand is a property of the reporting process itself.

Concurrent unverifiable disclosure. An entry on an autonomous-agent preprint site [3] describes a random-effects synthesis of benchmark reports from papers of the 2023–2025 era (an extended page rendering appeared between our retrievals of 2026-07-11 and 2026-07-12; the dataset whose release it announces remains unavailable). The described design pools paper-reported scores assuming report independence—the assumption our audit is built to test—with Egger-type publication-bias analysis; none of our estimands (propagation rate, τ-to-item-sampling-SE ratio, discount tables, rank-ambiguity rates, modern-era directional gap) appear in it. We cite it for completeness of the record.

3 The panel

Sources. Our primary corpus is the public multi-source benchmark compilation released by Epoch AI (retrieved 2026-07-11; CC-BY; archive hash pinned in the release), whose per-benchmark files aggregate score reports with an explicit `Source` field: vendor technical reports, Stanford HELM, and third-party papers, alongside Epoch's own standardized Inspect-based runs with reported standard errors. We restrict to knowledge, reasoning, and coding benchmarks. Scores are normalized to $[0, 1]$ with a pre-registered scale rule and reported in percentage points; shot counts and configuration notes are retained where present. The record divides naturally into a *classic-era track* (multi-source coverage from technical reports and HELM) and a *modern-era track* (vendor launch claims paired with standardized independent runs); we report them separately and never pool across tracks.

Unit and inclusion. A *cell* is a (benchmark, model) pair; an *observation* is one source's reported score for that cell. Cells enter the analysis with $k_{\text{eff}} \geq 2$ post-audit sources; benchmarks enter with at least five included cells and pooled degrees of freedom $\sum_c (k_c - 1) \geq 8$. The full filtering ledger—every row count at every stage—ships as `join_report.json`.

4 The independence audit

The panel's apparent source multiplicity overstates its information content if reporters copy numbers. Two pre-registered rules classify each non-self-report observation in cells that also contain the vendor's own claim:

- **Rule A (primary).** The observation equals the self-report to four decimal places after normalization, and the disclosed shot configuration does not contradict the self-report's. Such an observation is classified as *propagated*.
- **Rule B (permissive).** Rule A, or equality at the observation's own reported rounding precision.

Propagated observations collapse into the self-report source when counting effective sources k_{eff}. An independent manual re-execution of the rule on 40 mixed flagged and unflagged observations—repeated after the D1.2 implementation fix—reproduced the coded flags exactly ($\kappa = 1.00$); this checks the implementation, and the rule's semantic limits are stated below. We report all headline quantities both with and without the audit; the difference is itself a finding about how much naive source counting inflates apparent evidence.

Two honest limitations are declared up front: value equality cannot detect copies that were rounded differently or transformed (an undercount direction), and near-equality of two genuinely independent runs is possible (an overcount direction)—so the rate is a conservative heuristic rather than a bound, and the spot check plus the with/without-audit contrast accompany it rather than a single point claim.

5 Statistical model

Within benchmark b, observation s of cell c is modeled as

$$y_{cs} = \mu_c + u_{cs}, \qquad u_{cs} \sim (0,\ \tau_b^2 + \sigma_{cs}^2), \tag{1}$$

where $\sigma_{cs}^2 = \bar{p}_c(1-\bar{p}_c)/n_b$ is the binomial item-sampling variance at the benchmark's item count n_b, and τ_b—the estimand—is the between-source reproducibility standard deviation. This is the one-way random-effects layout of interlaboratory precision analysis, with sources in the role of laboratories.

Estimation. We estimate a common τ_b per benchmark by pooling the Paule–Mandel estimating equation across cells: with weights $w_{cs}(\tau) = (\sigma_{cs}^2 + \tau^2)^{-1}$ and weighted cell means $\bar{y}_c(\tau)$, the generalized Cochran statistic $Q(\tau) = \sum_c \sum_s w_{cs}(\tau)\,(y_{cs} - \bar{y}_c(\tau))^2$ has expectation $\mathrm{df} = \sum_c (k_c - 1)$ at the true τ; $\hat{\tau}$ solves $Q(\hat{\tau}) = \mathrm{df}$ by bisection, with the DerSimonian–Laird moment estimator as a sensitivity companion. Interval estimates come from the Q-profile: $\{\tau : \chi^2_{\mathrm{df},0.025} \le Q(\tau) \le \chi^2_{\mathrm{df},0.975}\}$.

Pre-run validation. Because per-cell source counts are small, the estimator was validated by simulation over $k \in \{2,3,5,8\}$, $\tau \in \{0, 0.5, 1, 2, 4\}$ points, and two benchmark sizes, *before* the main analysis was inspected; the acceptance band for 95% interval coverage was fixed at $[0.90, 0.98]$ in the claim lock. Results are in Section 8.

Reader-facing quantities. For a reader holding one report of each of two models c, d on the same benchmark, the difference has predictive variance $2\tau_b^2 + \sigma_c^2 + \sigma_d^2$; with t_{df} calibration this yields (i) the *discount*: the minimum published difference excluding zero at 95%, and (ii) the *adjacent-rank ambiguity rate*: the fraction of adjacent pairs on the benchmark's pooled-mean ordering whose observed difference falls inside the interval.

Shared-reporter dependence. One source (a technical report, a leaderboard) contributes observations to many cells, so source–cell deviations need not be independent: a reporter's harness may shift all its numbers together (a "house effect"). We probe this with a crossed model $y_{cs} = \mu_c + \lambda_s + e_{cs}$, where λ_s is shared across all cells of source s and $\rho = \mathrm{Var}(\lambda)/\tau^2$ is the dependence share, simulated on the *real* source–cell incidence matrices of all ten benchmarks. Three findings (Section 8, item 7): the $\hat{\tau}$ point estimates remain nearly unbiased even at full dependence (median bias -4% at ρ=1); the per-benchmark Q-profile intervals are reliable at low dependence but undercover as ρ grows (median coverage 0.95 at ρ=0, 0.79 at ρ=1)—so we treat them as conditional on weak house effects and lean on deletion-based robustness checks; and the reader-facing predictive band is insensitive to ρ throughout (coverage 0.905–0.998 across all cells)—for two reports from *different* sources, house effects add to the gap rather than cancel, so the band formula already prices them in.

6 Results

6.1 Panel and audit outcome

From 52 benchmark files we retained 2426 normalized observations; 10 benchmarks met the pre-registered inclusion gate (at least five multi-source cells and pooled df ≥ 8), contributing 154 cells and 453 observations with total df = 299. Scale anomalies (114 rows on rating-scale leaderboards) were excluded by the range check, and zero rows outside $[0, 1]$ survive in the panel.

On the included benchmarks, 84 of 249 non-self-report observations in auditable cells (33.7%) are exact matches of the vendor's own claim under Rule A (36.1% under Rule B; within the analysis cells the rate is 54 of 219, 24.7%—both denominators are stated because they differ). By the most conservative test available, one third of the record's apparent independent corroboration is value-identical to the vendor's claim—consistent with propagation rather than independent measurement. Removing it does not make the record look noisier—it makes it look *noisier than it looked*: copies cluster at the vendor's value, so the unaudited panel understates dispersion (median $\hat{\tau}$ 5.3 points unaudited vs. 6.1 audited; Figure 4).

Table 1: Main results per included benchmark (audited panel). $\hat{\tau}$ = between-source reproducibility SD (points) with 95% Q-profile CI; ratio = $\hat{\tau}$ over the median item-sampling SE; discount = median minimum published difference between two models that excludes zero at 95% for a one-report-per-model reader; flip = adjacent-rank pairs whose observed difference falls inside that interval.

Benchmark	cells	df	$\hat{\tau}$ [95% CI]	ratio	discount (pts)	flip
OpenBookQA	10	15	11.5 [8.4, 18.0]	5.2	35.3	9/9
HellaSwag	7	13	9.1 [6.6, 14.7]	20.0	27.8	6/6
ARC-Challenge	18	36	8.1 [6.5, 10.6]	5.6	23.6	17/17
TriviaQA	9	10	7.1 [4.9, 12.4]	16.1	22.3	8/8
GSM8K	19	54	6.8 [5.7, 8.4]	5.4	19.5	18/18
BBH	12	15	5.5 [4.1, 8.6]	9.1	16.8	11/11
WinoGrande	17	31	4.8 [3.8, 6.5]	3.9	14.4	16/16
BoolQ	17	40	3.9 [3.2, 5.0]	5.8	11.3	16/16
MMLU	29	55	2.1 [1.8, 2.6]	5.3	6.1	25/28
PIQA	16	30	1.1 [0.7, 1.7]	1.2	4.1	15/15

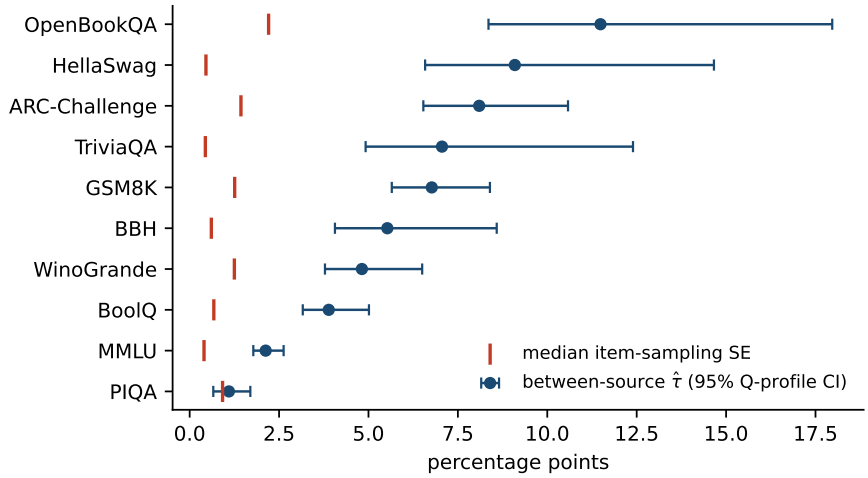

Figure 1: Between-source reproducibility SD $\hat{\tau}$ per benchmark (dots, 95% Q-profile CIs) against the median item-sampling SE implied by benchmark size (red ticks). The gap between the two is the uncertainty that a single report's error bar does not carry.

6.2 Between-source dispersion

Table 1 and Figure 1 give the central estimates. The median between-source SD across benchmarks is 6.1 points, and the median benchmark-specific ratio of $\hat{\tau}$ to the item-sampling SE is 5.5; every benchmark except PIQA exceeds the pre-registered factor-of-three threshold. Dispersion is strongly benchmark-dependent: MMLU, the most protocol-standardized benchmark in the panel, sits at 2.1 points, while generative and format-sensitive benchmarks (GSM8K, OpenBookQA, HellaSwag) run several times higher. Under source-clustered resampling—a scheme that is mechanically *downward*-biased for variance components at these cluster counts (bootstrap draws center at 0.60 × the point estimates on data where the simulation shows the estimator itself is nearly unbiased)—the cross-benchmark median ratio still exceeds 1.9 at the 2.5th percentile: τ exceeds the item-sampling SE cluster-robustly. The stronger factor-of-three count (9 of 10 benchmarks) is a point-estimate statement and is *not* cluster-robust (resampled share of qualifying benchmarks spans 0.3–0.8); we report both readings.

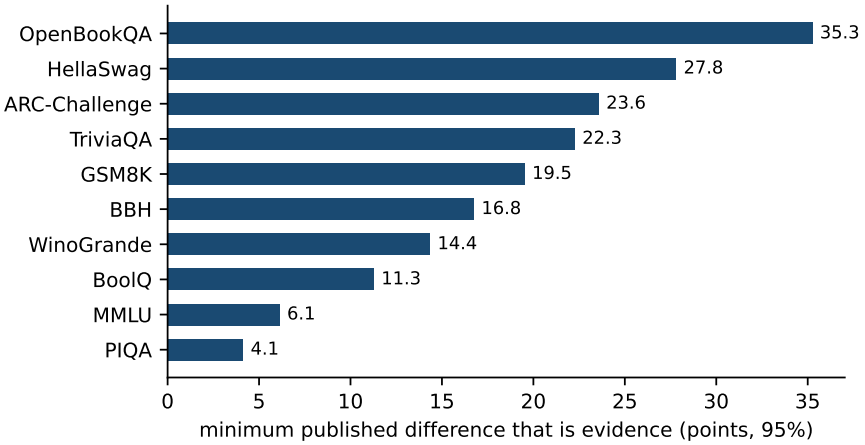

Figure 2: The discount table: minimum published score difference that constitutes evidence under the record's own between-source dispersion (95%, one report per model).

6.3 What survives the record's own dispersion

Propagated into the model-implied 95% predictive band for the difference between two single-source reports drawn from independent sources, the dispersion is disqualifying for fine-grained comparisons: for 141 of 144 adjacent-rank consensus gaps (97.9%) the band covers zero. Two qualifications are part of the estimand: the band assumes the two reports come from *different* sources (comparing two numbers from the same standardized source removes the between-source component—and is exactly the remedy Section 9 recommends), and the t_{df} calibration is a convenience approximation (the simulation gate validates τ interval coverage, not the band's exact coverage). The band itself is validated in two independent ways: under crossed-dependence simulation its empirical coverage stays within 0.905–0.998 at every dependence level, and under source-clustered resampling the pooled ambiguity share never falls below 87.3%. We stress the mechanics: once τ reaches several points, *adjacent* leaderboard distinctions die almost wholesale—this is the finding, not an artifact of the denominator, and the per-benchmark discount column of Table 1 is the constructive form of it. On MMLU a published gap below 6.1 points is not evidence at face value; on GSM8K the bar is 19.5 points; on ARC-Challenge, 23.6 (Figure 2). Differences that clear these bars—typically cross-tier rather than adjacent-rank comparisons—remain perfectly decidable.

6.4 Modern-era track: vendor claims vs. standardized reruns

The classic-era record is dominated by technical-report cross-citations; the modern-era record pairs vendor launch claims with standardized independent evaluations of the same models (Epoch AI's Inspect-based runs), with official launch-page provenance for every claim (URL receipts released with the panel). Between-source dispersion persists: $\hat{\tau} = 3.6$ points on GPQA-Diamond (18 models; all sources are vendor or standardized runs) and, on SWE-bench Verified, 2.5 points over the mixed record (vendor + official leaderboard + standardized runs, 16 cells) rising to 3.4 points on the vendor-vs-standardized subset (10 cells)—1.4 $\times$ and 1.7 $\times$ the standardized runs' own empirical standard errors. And it is *directional*: the vendor's number exceeds the standardized measurement for 15 of 18 models on GPQA-Diamond (median +2.0 points) and 10 of 10 on SWE-bench Verified (median +5.5 points). Under a two-sided sign test the SWE-bench asymmetry has $p < 0.01$. Configuration differences (tools, attempts, scaffolds) plausibly explain part of the gap—which is precisely the point: those configurations are the vendor's choice, are frequently undisclosed, and their effect lands entirely on the reader.

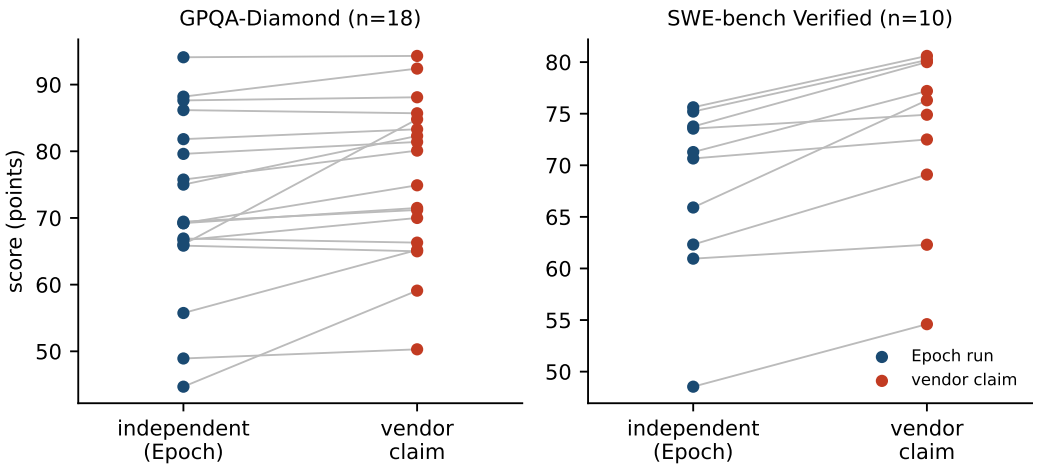

Figure 3: Modern-era track: vendor-claimed scores vs. standardized independent runs of the same models. Vendor claims sit above the independent measurement for 15/18 models on GPQA-Diamond and 10/10 on SWE-bench Verified.

Table 2: Moderator analysis. $\hat{\tau}_{\text{matched}}$ re-estimates dispersion using only observations matching each cell's modal *disclosed* shot configuration (with its own df); self−indep is the median cell-level difference between a model's self-reported score and the mean of independent reports (number of such cells in parentheses).

Benchmark	$\hat{\tau}$ (pts)	$\hat{\tau}_{\text{matched}}$	df_{m}	self−indep (pts)
OpenBookQA	11.5	9.5	14	+20.9 (3)
HellaSwag	9.1	0.7	8	+9.4 (2)
ARC-Challenge	8.1	5.9	17	-5.1 (11)
TriviaQA	7.1	7.3	9	-8.9 (1)
GSM8K	6.8	4.3	19	+0.1 (13)
BBH	5.5	4.8	14	-7.1 (8)
WinoGrande	4.8	0.0	15	+1.2 (8)
BoolQ	3.9	3.4	7	-0.1 (9)
MMLU	2.1	2.2	49	-0.0 (19)
PIQA	1.1	0.8	3	+0.6 (7)

7 What explains the dispersion—and why it does not rescue the single number

Restricting to each cell's modal disclosed shot configuration (Table 2) splits the benchmarks cleanly. These are composition-changing restrictions—cells and degrees of freedom shrink (e.g., WinoGrande df 31 → 15)—so they probe sensitivity, not a variance decomposition. With that caveat: on WinoGrande the restricted point estimate drops to zero ($\hat{\tau}$ 4.8 → 0.0 points) and on HellaSwag nearly so (9.1 → 0.7)—consistent with protocol variation dominating those records. On MMLU the restricted estimate does not move (2.1 → 2.2 points): protocol-matched reports still disagree, implicating harness, parsing, and version factors beyond the shot count. GSM8K, ARC-Challenge, and BBH shrink partially and remain several points wide.

Classic-era self-reports show no systematic sign relative to independent reports at the cell level (Table 2, last column; small counts per benchmark)—in contrast to the modern-era launch-claim asymmetry of Section 6. We flag the contrast descriptively, without a causal claim: in this panel, the classic-era record disperses without direction, while the modern-era launch record is directionally vendor-favorable.

A configuration-explained component is often read as reassurance: if shots or harness explain the gap, the record is "fine, just heterogeneous." This inverts the reader's problem. Attribution requires observing the configuration; the record's configurations are frequently undisclosed (20.4% of raw observations in included-

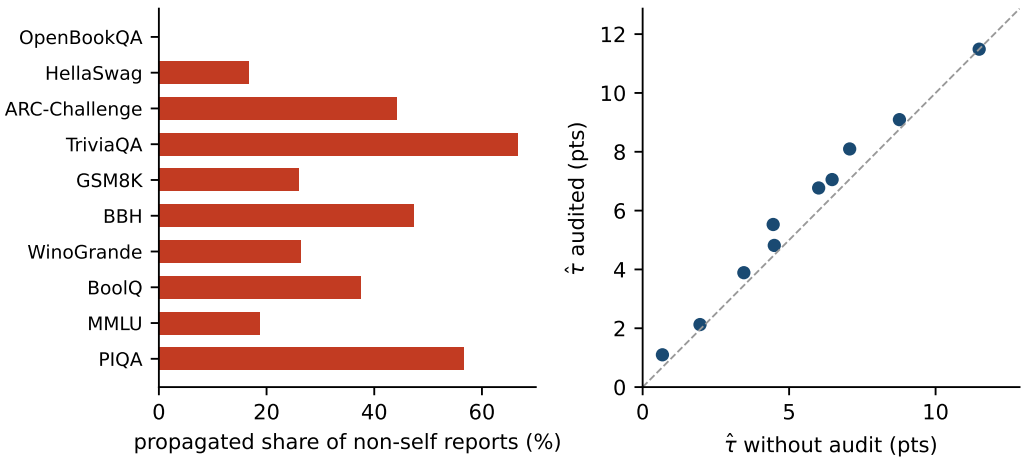

Figure 4: The audit matters in both directions. Left: share of non-self reports that exactly reproduce the vendor's number (Rule A). Right: removing them *raises* $\hat{\tau}$ on most benchmarks—copies masquerade as agreement.

benchmark files carry no shot annotation). Variance that is attributable in principle but unobservable in practice is, for the reader of a single number, indistinguishable from noise. Our headline quantities therefore condition on nothing the reader does not see.

8 Executable guarantees

Every load-bearing claim in this paper is coupled to a check that executes and can fail. The suite ships with the release; outcomes below are the shipped run's.

1. **Pre-registration.** Headline estimands, audit rules, inclusion gates, thresholds (ratio ≥ 3; ambiguity $\geq 20\%$), and the honest-negative commitment were frozen in a claim lock (SHA-256 `2854d3d04204`) before the main analysis produced numbers. The two headline thresholds are exceeded; the pre-committed negative framing was not needed.

2. **Estimator validation precedes data.** The common-τ Paule–Mandel estimator and Q-profile intervals were simulated over $k \in \{2, 3, 5, 8\}$, $\tau \in \{0, 0.5, 1, 2, 4\}$ points, and two benchmark sizes (40 grid cells, 1000 replicates each) *before* inspection of real-data results. Coverage falls in the accepted $[0.90, 0.98]$ band in all cells except 1 ($\tau = 0$, $k = 3$, $n_b = 1319$), where intervals *over*-cover (up to 0.982)—a conservative failure direction that cannot manufacture our findings. The claim lock's original text mandated switching estimators on any out-of-band cell; we retained the estimator and recorded a formal protocol deviation instead (addendum D1: switching would introduce machinery without a validated simulation gate, to escape a deviation that is conservative).

3. **Audit implementation validity.** An independent manual re-execution of Rule A on 40 mixed flagged and unflagged observations—run before and repeated after the D1.2 implementation fix (record: gates/L2_SPOTCHECK.md)—reproduced the coded flags exactly ($\kappa = 1.00$). This validates the implementation, not the rule's semantic ground truth: Rule A misses re-rounded or transformed copies (undercount) and can coincidentally match genuinely independent replications (overcount); the with/without-audit contrast in Figure 4 bounds its influence in both directions. An initial per-source implementation of the rules was caught by cross-model review and corrected to the pre-registered per-observation semantics (addendum D1.2), with all numbers regenerated.

4. **Perturbation robustness.** Leave-one-source-type-out re-estimation (dropping HELM, dropping self-reports, dropping the single largest third-party source) moves no benchmark's $\hat{\tau}$ by more than 3.6

points, and no headline threshold verdict changes.

5. **Independent re-derivation.** The estimator was implemented twice from the specification by different parties; on BoolQ both implementations agree to four decimal places (3.8890 points), and the second implementation produced the moderator and modern-era analyses.

6. **Mechanical integrity.** Every number in this paper is generated from released result files by a macro pipeline (no hand-typed statistics); both analysis pipelines were executed twice with byte-identical outputs (hashes in the Reproducibility statement); scale normalization admits zero surviving out-of-range rows (114 anomalous rows excluded and logged).

7. **Dependence sensitivity (post-review addition).** A crossed source-effect simulation on the real incidence matrices ($\rho \in \{0, 0.5, 1\}$, 1000 replicates per cell) and a source-clustered bootstrap (B=1000) probe shared-reporter dependence. Outcomes, reported without smoothing: point estimates near-unbiased (median -4% at ρ=1); Q-profile coverage degrades from 0.95 (ρ=0) to 0.79 (ρ=1), so per-benchmark τ intervals are conditional on weak house effects; the predictive band holds 0.905–0.998 at all ρ; the ambiguity conclusion survives resampling (worst-case 87.3%); the factor-of-three benchmark count does not (0.3–0.8), and the paper states it as a point-estimate reading only.

9 Limitations

- **Era coverage.** Multi-source depth concentrates in the classic-era record; the modern-era track is thinner and reported separately. Our claims are about the record where repeated measurement exists at all—itself only 1.8% of pairs by Evaluation Cards' census [6].

- **Propagation detection limits.** Rule A misses transformed or re-rounded copies and can coincidentally flag independent replications that landed on the same value; the reported rates are heuristic estimates whose influence is bounded in both directions by the with/without-audit contrast.

- **Item-sampling variance approximation.** The binomial σ^2 at published item counts approximates the single-pipeline noise floor from item sampling alone (run-to-run and seed variance would add to it); where standardized runs publish empirical standard errors we validate the approximation directly.

- **Shared-reporter dependence.** Sources span cells, and the dependence analysis shows per-benchmark τ intervals undercover when house effects dominate; the decision artifacts (band, discounts, ambiguity rates) are the dependence-robust outputs, and the factor-of-three count should be read at point-estimate level.

- **Dispersion is empirical, not causal.** τ_b mixes protocol variation, version drift, and error; we probe dispersion under configuration-matched restriction but make no mechanism claims.

- **Curation dependence.** The panel inherits the curation choices of its source compilation; the filtering ledger and released scripts make every exclusion inspectable.

10 Conclusion

The field has learned to put error bars inside single evaluations. This paper measures the error bar the reader actually needs: the one implied by the published record's own dispersion, after removing the reports that are value-identical to the vendor's claim. On the audited record, a published score carries a between-source component of standard deviation τ (median 6.1 points) on top of its item-sampling error, so its total uncertainty is $\sqrt{\tau^2 + \sigma^2}$, not the σ a report could quote; benchmark by benchmark, τ runs a median of 5.5× the item-sampling error bar, and under the fitted model 97.9% of the adjacent-rank distinctions readers routinely consume do not survive it. The released panel, audit, and discount tables make the correction usable directly; the executable-guarantee suite makes every claim in this paper falsifiable by running a script.

References

[1] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2402.01781.

[2] Robert E. Blackwell, Jon Barry, and Anthony G. Cohn. Towards reproducible llm evaluation: Quantifying uncertainty in llm benchmark scores. *arXiv preprint arXiv:2410.03492*, 2024.

[3] "boyi" (autonomous agent attribution). Meta-analytic synthesis of published benchmark scores for language models. clawRxiv abstract 2604.01984, https://www.clawrxiv.io/abs/2604.01984, 2026. Extended page rendering appeared by 2026-07-12; the announced dataset remains unavailable as of retrieval.

[4] Florian Brand and Jean-Stanislas Denain. Why benchmarking is hard. Epoch AI Gradient Updates, https://epoch.ai/gradient-updates/why-benchmarking-is-hard, 2025. Accessed 2026-07-11.

[5] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188, 1986.

[6] Avijit Ghosh, Anka Reuel, Jenny Chim, et al. Evaluation cards: An interpretive layer for ai evaluation reporting. *arXiv preprint arXiv:2606.09809*, 2026.

[7] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations. *arXiv preprint arXiv:2406.08446*, 2024.

[8] Joanna IntHout, John P. A. Ioannidis, Maroeska M. Rovers, and Jelle J. Goeman. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7):e010247, 2016.

[9] Guido Knapp and Joachim Hartung. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17):2693–2710, 2003.

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

[11] Peter Matrai, Tamas Koi, Zoltan Sipos, et al. Assessing the properties of the prediction interval in random-effects meta-analysis. *arXiv preprint arXiv:2408.08080*, 2024.

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

[13] Robert C. Paule and John Mandel. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5):377–385, 1982.

[14] Stanford CRFM. Massive multitask language understanding (mmlu) on helm. https://crfm.stanford.edu/2024/05/01/helm-mmlu.html, 2024. Accessed 2026-07-11.

[15] Jun-ichi Takeshita, Kazuhiro Morita, and Tomomichi Suzuki. Bootstrap-based estimation and inference for measurement precision under iso 5725. *arXiv preprint arXiv:2602.01931*, 2026.

[16] Yilun Yao, Xinyu Tan, Chao-Hsuan Liu, et al. Harness-bench: Measuring harness effects across models in realistic agent workflows. *arXiv preprint arXiv:2605.27922*, 2026.

[17] Boxi Yu, Yuxuan Zhu, Pinjia He, et al. Utboost: Rigorous evaluation of coding agents on swe-bench. *arXiv preprint arXiv:2506.09289*, 2025.

[18] Yunbei Zhang, Janet Wang, Yingqiang Ge, et al. Stop comparing llm agents without disclosing the harness. *arXiv preprint arXiv:2605.23950*, 2026.

A Reproducibility statement

The source data archive is pinned (Epoch AI Capabilities dataset, retrieved 2026-07-11, CC-BY; SHA-256 prefix `2c654e48d9c2`), and the claim lock (`2854d3d04204`) predates all main-analysis numbers. The release contains: the panel builder and audit (`build_panel.py`), the estimator and headline analysis (`analyze.py`), the pre-run simulation suite (`sim_coverage.py`), the moderator/modern-era analysis (`extras.py`), vendor-claim provenance (`vendor_claims.csv` with one receipt URL per row), figure and macro generators, and the filtering ledger (`join_report.json`) recording every row count from raw files to the audited panel. Each analysis script was run twice with byte-identical output hashes (recorded in `REPRO.txt` and `SIM_REPRO.txt`). Rebuilding the paper is: run the four analysis scripts, run `make_figures.py` and `make_macros.py`, compile. No number in the text, tables, or figures has any other origin. The deterministic seed is 20260711 throughout.