Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CleanScore: Black-Box Benchmark Audits with Negative Controls and Sensitivity Bounds

Jeffery Opoku

jefferyopoku491@gmail.com

University of Texas Rio grande valley

David Banahene

Florida International University

Research Article

Keywords:

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10884991/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

CleanScore: Black-Box Benchmark Audits with Negative Controls and Sensitivity Bounds

Jeffery Opoku
School of Mathematical and Statistical Sciences
The University of Texas Rio Grande Valley, Edinburg, Texas, USA
jeffery.opoku01@utrgv.edu

David Banahene
Department of Biostatistics, Robert Stempel College of Public Health and Social Work
Florida International University, Miami, Florida, USA

Abstract

Public benchmark scores may reflect skill, prior exposure to the questions, or both, and for most models the training data are unknown. We present CleanScore, a black-box audit using scored outputs only. Each benchmark question becomes a parent item with one public form and two independently written fresh forms preserving its numbers, facts and answer. The audit reports an interval for the public-form advantage rather than a verdict, and a private negative-control bank with an explicit transport radius separates exposure from ordinary form mismatch. A registered controlled-exposure experiment detects planted exposure and stays quiet under fresh-form exposure. A registered audit of five open models on 200 GSM8K and 200 ARC-Challenge items finds no exposure-consistent advantage, bounding surface-form inflation below five points. Registered positive controls then bound what such a null can mean. Leaking an item raises accuracy on paraphrases the model never saw almost as much as on the leaked wording, leaving 52% to 110% of the effect invisible to a paraphrase audit. On ARC a planted 49-point advantage shows an observable gap of -0.020, and about 20 points survive rewriting stem and options, across four training seeds. A surface-form null bounds far less than the phrase contamination audit implies.

1 Introduction

Public benchmark scores shape model rankings, product claims, and research decisions. Their meaning is less clear when benchmark questions may have appeared in training data. For most closed models, nobody outside the lab can inspect the training corpus, so a score can mix two things: the model's ability to solve the task, and an advantage from having seen the same or closely related material before.

Several tools address this. Clean-Eval rewrites evaluation sets, PaCoST compares model confidence on original and counterpart questions, ConStat corrects for difficulty with reference models, DICE inspects internal states, and DyePack plants markers before a benchmark is released. Each answers a real question. But most of them need something an ordinary user of an API does not have: token probabilities, model internals, a set of reference models that are known to be clean, or control of the benchmark before it became public. And most of them return a verdict, a flag or a p-value, rather than a statement of how large the inflation could be and how sure the auditor can be about it.

CleanScore asks a narrower question that can be answered from scored outputs alone. For a fixed model and benchmark, is accuracy on the public form of a question higher than accuracy on independently written forms that require the same answer and the same reasoning? This difference is observable. Its interpretation is not automatic: a fresh form can be slightly harder, easier, or less natural than the original. The central problem is therefore not only estimating the gap but separating what the data identify from what depends on an

assumption about form mismatch. CleanScore makes that assumption explicit, bounds it with a never-public negative-control bank, and reports the conclusion across a visible range of transport radii.

This paper makes six contributions, each backed by an experiment rather than a promise.

1. A black-box matched-item estimand for the public-form advantage on a finite benchmark, with a conservative bounded-difference interval and a finite-population interval reported side by side (Sections 3 and 6).

2. A negative-control sensitivity interval with an explicit benchmark-to-control transport radius, and an exact placebo-label identity for checking the scoring pipeline (Section 4).

3. A registered controlled-exposure validation in two model families showing that the design detects planted exposure to public forms and stays quiet under exposure to fresh forms (Section 5).

4. A registered real-benchmark audit of five open models on GSM8K and ARC-Challenge, with 1,500 verified matched forms and private control banks, reported under a claim rule fixed before any model was queried, together with a pre-declared robustness rerun (Section 7).

5. Head-to-head comparisons on the same items with a fresh-set score and a confidence-based detector, including a case where the detector and the audit disagree (Section 7.6), and a variance-adaptive distribution-free interval, with simulated coverage, that repairs the one regime where the registered bound is uninformative (Section 7.5).

6. A budgeted-audit analysis showing how many audited items are needed for stable intervals and a stable model ranking (Section 7.8), a detection-rate calibration that says what size of contamination an audit of a given size can find (Section 7.9), a randomization test that restores power where a distribution-free interval has none, with a measurement of what its exactness costs when forms are not exchangeable (Section 7.10 and Proposition 6), a registered positive control that measures how much of a contamination effect a paraphrase audit can see (Section 7.11), a second registered positive control that failed its own manipulation check and is reported as such with the cause identified (Section 7.12), a third that corrects the cause and plants 49 points of contamination the audit cannot see (Section 7.13), a fourth that removes the one assumption the third still rested on and measures how much of that contamination survives a total rewrite (Section 7.14), and a negative control that runs the same fine-tuning with no leak at all and finds the same statistic within two points of zero (Section 7.15), and a released item bank and code base.

One of those results qualifies all the others, so we state it at the outset rather than saving it for Section 7.11. Four models from four families were deliberately contaminated with the public wording of a hundred audited items. Leaking an item raised accuracy on the leaked wording by 26 points against an uncontaminated control group of items, and raised accuracy on paraphrases the model had never seen by 20 points. Only the difference between those two, about 6 points, is what a paraphrase audit measures at all; Figure 1 shows the split.

The consequence is not confined to CleanScore. An audit built on comparing an original question against a rewrite of it measures the visible part only, so a null from such an audit bounds surface-form inflation tightly and bounds total contamination much more weakly than the phrase contamination audit suggests. We report the nulls of Section 7 with that reading attached. The mechanism is a property of what a model learns rather than of any one benchmark. The size of the invisible share is measured on GSM8K; on ARC a separate registered control shows the same blindness for a different reason, because that benchmark's rewrites hold the options fixed and a memorised answer string survives them intact. Section 8 says what each leaves open.

CleanScore is not presented as the first paraphrase audit, the first contamination detector, or proof that a proprietary model used a specific dataset. Its value is a reusable, honest uncertainty statement when only black-box scored outputs are available.

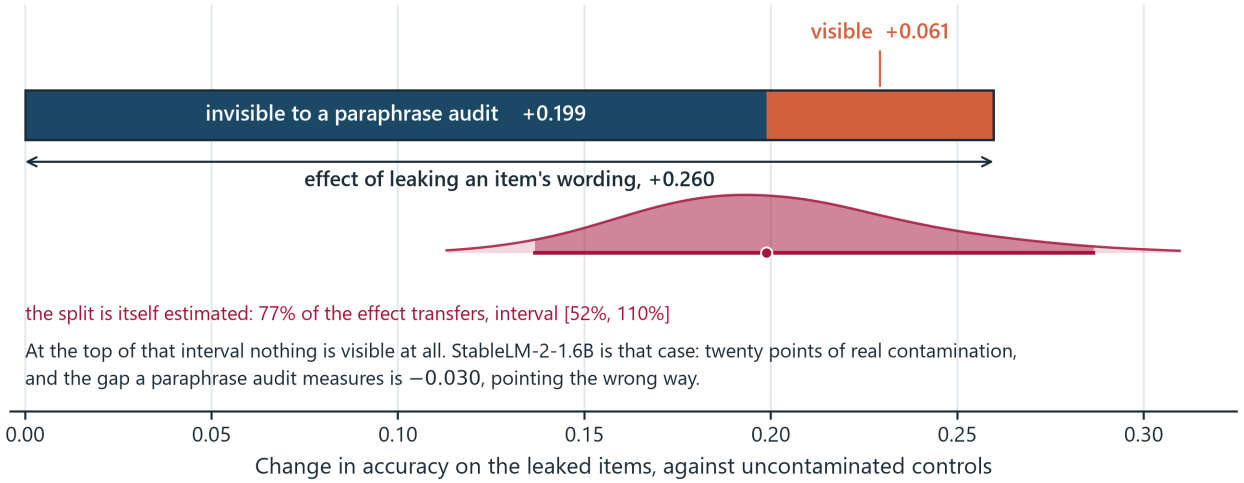


Figure 1: The finding of Section 7.11, stated once at the start because it bounds how much any result in this paper can mean. Leaking an item's wording raises accuracy on that wording, and most of that gain carries over to paraphrases of the same item, where an audit that compares an original against a rewrite cannot see it. The estimated split is 77% invisible, with an interval running from 52% to 110%; at the top of that interval nothing is visible at all.

2 Related work

Clean-Eval builds filtered paraphrased evaluation sets and shows why public benchmark forms can give optimistic results (Zhu et al., 2024). PaCoST compares model confidence on original and counterpart questions (Zhang et al., 2024). ConStat estimates benchmark-specific performance inflation by comparing a target model with reference models and reference benchmarks (Dekoninck et al., 2024). DICE uses internal model states to detect in-distribution contamination during fine-tuning (Tu et al., 2024). Prospective approaches such as DyePack place randomized markers in a benchmark before exposure and can therefore provide stronger evidence when that design is available (Cheng et al., 2025). Closed or partly closed benchmarks, including MMLU-CF, reduce future exposure risk but do not directly audit every existing public benchmark (Zhao et al., 2024).

Closest of all in intent is work that attaches a formal guarantee to a contamination decision rather than returning a bare verdict. Zhang et al. (2026) combine several detectors under an adaptive weighting scheme and control the false discovery rate among the samples a filter keeps, so a user can set in advance how much contamination they are willing to leave in an evaluation set. That is the same instinct as ours, applied at the other end of the pipeline: their guarantee is about which items survive a filter, ours is about how large an advantage could be on items that were kept. The two are complements rather than competitors, and an evaluation team plausibly wants both, one to clean a benchmark and one to say what a clean-looking score still permits. Their approach also needs per-sample detector scores, which an API user auditing a closed model does not have, so the two occupy different access regimes.

Three further lines of work sit close to this paper and shape what it can claim.

The first is black-box contamination testing that needs no reference model and no control of the benchmark. Oren et al. (2024) show that a contaminated model finds a benchmark's canonical example ordering more likely than a shuffled one, and turn that into a test with a provable guarantee from scored outputs alone. The relationship to our Section 7.10 is close and worth stating precisely, because both rest on exchangeability and rest on it differently. Their exchangeability is over the order of examples in a dataset, which is a property of how the benchmark was assembled and is plausible by construction. Ours is over which of a parent's forms is the original, which is a property of the forms themselves and is false whenever a rewrite is even slightly harder than what it replaces. Their test detects that a benchmark was trained on; ours, when it rejects,

cannot separate contamination from form mismatch, which is why Section 7.10 uses it only to strengthen a null. An audit that wants a positive finding from a black box should reach for theirs, not ours.

The second is the observation that paraphrase defeats the decontamination that most benchmark releases actually perform. Yang et al. (2023) show that rephrasing or translating test items slips past n-gram overlap filters, and that a 13B model trained on such variants reaches scores that look like a far larger model's. That result is the mirror image of ours and the two fit together: they show a contaminated model can be built whose contamination string matching cannot see, and we show that once such a model exists, an audit comparing an original against a rewrite cannot see most of it either. Their concern is the decontamination pipeline, ours is the audit that runs afterwards, and a benchmark maintainer needs both.

The third is the question of what a contamination detector's output is worth when someone is trying to defeat it. Dekoninck et al. (2024b) show that several published detectors can be evaded cheaply, by contaminating a model in ways that leave the statistic they measure untouched. We take that as the correct frame for reading our own nulls. CleanScore measures a surface-form advantage and reports how large it could be; it is not a test of whether a model is contaminated, and Section 7.11 to Section 7.15 measure how much of a real, deliberately planted contamination it fails to register. A detector that can be evaded and an audit that bounds only part of what it is asked about are the same problem seen from two directions, and neither is repaired by reporting a verdict more confidently.

A second line of work rebuilds a benchmark rather than rewording it. GSM1k writes a fresh set of grade-school arithmetic problems matched to GSM8K in style and difficulty, and reports accuracy drops of up to about 13 points for some model families alongside near-zero gaps for others (Zhang, Da, et al., 2024). That design measures dependence on the benchmark's whole distribution: its problem types, its phrasing conventions, and its answer magnitudes. CleanScore measures something deliberately narrower, the advantage attached to the exact wording of the published item, holding every quantity and the answer fixed. The two questions are different and the answers need not agree. A model can be unaffected by rewording and still be helped by having trained on many problems of the same shape, which is why the null results in Section 7 are reported as a statement about surface form and not as a clean bill of health.

CleanScore overlaps with this literature in motivation but differs in its target and access assumptions. It estimates an observable score gap from correctness alone. It then treats residual form mismatch as an identification problem rather than assuming that every score drop is caused by prior exposure. The negative-control bank and transport curve are intended to make this assumption visible.

A third line of work is not about contamination at all but shares this paper's shape, and is the closest methodological neighbour to the design rather than to the question. SynthGuard-ReleaseBench audits whether a synthetic tabular dataset supports a named use, and it does so by locking the use, the panel of checks and the tolerances before any evaluation, reporting simultaneous finite-sample bounds on bounded loss gaps, requiring controls, and keeping what was measured separate from what may be claimed (Opoku and Banahene, 2026a). The subject matter is unrelated to benchmark contamination, but the four commitments are the ones made here: fix the analysis first, bound a bounded quantity without distributional assumptions, carry a control that isolates the nuisance, and refuse to convert the estimate into a verdict. Conformal prediction is the other large distribution-free family in machine learning, and its guarantees rest on exchangeability in the same way the exact test of Section 7.10 rests on exchangeability of a parent's forms; work on restoring those guarantees when exchangeability fails is therefore relevant to what an audit can promise when its own assumption is doubted (Opoku and Banahene, 2026b).

The confidence bounds used for finite-population sampling and the standard Wald interval come from established probability and survey-sampling tools (Hoeffding, 1963; Serfling, 1974; Cochran, 1977). Negative controls are also an established way to reveal bias that a primary comparison cannot separate on its own (Lipsitch et al., 2010). Their use is not claimed as a new theorem. The proposed contribution is the audit design that connects these tools to matched forms, private controls, and an explicit sensitivity analysis.

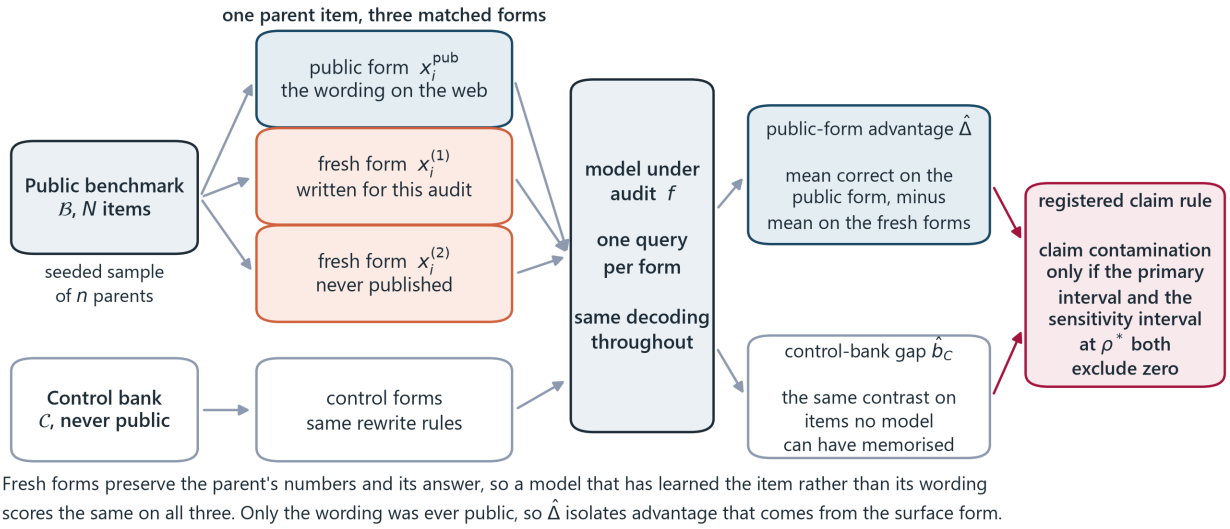


Figure 2: The matched-form audit. Each sampled parent item is put to the model in its public form and in two fresh forms that keep the same numbers, facts, and answer; the public-form advantage is the difference in mean correctness between them. A never-public control bank of items no model can have memorised is scored the same way and supplies the form-mismatch term. The registered claim rule requires both intervals to exclude zero, so ordinary form mismatch cannot on its own produce a contamination claim.

3 Audit setup

Figure 2 is the whole procedure on one page: where the forms come from, what the model is asked, and how the two estimands combine into the rule that decides whether a contamination claim may be made at all. This section defines the estimand and its intervals, Section 4 defines the control bank and the transport radius, and Section 7.1 records the settings actually used in the audit reported here.

3.1 Parent items and matched forms

Benchmark parent item i has one public original form and K fresh forms. All forms must have the same answer criterion and required reasoning. Let Y_{iO} be correctness on the original and Y_{iFk} correctness on fresh form k. Define

$$D_i = Y_{iO} - \frac{1}{K}\sum_{k=1}^{K} Y_{iFk}.$$

The value D_i lies between -1 and 1. A positive value means the model performed better on the public form for that parent item.

Fresh forms remain private until all target-model queries and scoring are complete. The protocol requires two checks before any model is queried. The first is mechanical and must be automated: answer preservation, preservation of every quantity the item depends on, and, for multiple-choice items, untouched options and key. The second is a reading by the audit's authors for missing information, added or removed reasoning steps, and unnatural phrasing. Both checks, and every item they send back for rewriting, are recorded in the audit log. Section 7.1 states exactly how the forms used in this paper were drafted and which of these checks had been completed at the time of writing.

3.2 Finite-benchmark target

For a benchmark containing N parent items, the target is

$$\Delta_N = \frac{1}{N}\sum_{i=1}^{N} D_i.$$

If every benchmark item is audited, this finite-benchmark mean has no item-sampling error. If only n parent items are sampled, the estimator is the sample mean or its stratified weighted version.

3.3 Dual uncertainty reporting

Because $D_i \in [-1, 1]$, a simple random sample without replacement gives the conservative half-width

$$h_H = \sqrt{\frac{2\log(2/\alpha)}{n}}.$$

For a stratified audit with benchmark weights W_h and sample sizes n_h, the conservative half-width becomes, by Proposition 2 in the mathematical core,

$$h_{H,\mathrm{str}} = \sqrt{2\log(2/\alpha)\sum_h \frac{W_h^2}{n_h}}.$$

These bounds are finite-sample and distribution-free, but they can be too wide for practical decisions. We therefore also report the standard finite-population variance estimate

$$\widehat{V} = \sum_h W_h^2(1 - f_h)\frac{s_h^2}{n_h}, \qquad f_h = \frac{n_h}{N_h},$$

and the large-sample interval

$$\widehat{\Delta}_{\mathrm{str}} \pm z_{1-\alpha/2}\sqrt{\widehat{V}}.$$

The two intervals answer the same sampling question but have different guarantees. Both are reported.

4 Negative-control sensitivity analysis

4.1 Why the public-form gap is not enough

Let b_B be the form gap that the same model would show on benchmark B without prior exposure. Let τ_B denote the exposure-related part. Then

$$\Delta_B = \tau_B + b_B.$$

The audit estimates Δ_B. It does not observe the counterfactual b_B. Human ratings alone cannot guarantee a bound on a model-specific accuracy difference.

4.2 Private control bank

The control bank is never public before evaluation. Each control parent has one public-like first form and later fresh forms created under the same protocol as the target audit. Let b_C be its average form gap. We use the transport assumption

$$|b_B - b_C| \le \rho.$$

The radius ρ is not fitted to obtain a desired conclusion. Results are shown over a grid chosen before the target audit is unblinded.

The shape of that guarantee is not peculiar to auditing. Whenever a method is calibrated on one population and applied to another, the honest statement is a guarantee that holds up to an explicit term for the mismatch between them, and the useful question becomes how large that term has to be before the conclusion changes. Conformal risk control for foundation models meets the same structure when the prompt stream drifts away from the calibration pool, and states its guarantee up to terms for distribution mismatch in the same way Proposition 3 states its guarantee up to ρ (Opoku and Banahene, 2026c). What differs is the source of the mismatch, prompts moving over time in that setting against a control bank standing in for a benchmark here, and what is done about it: that work reweights calibration towards the current stream, whereas an audit has no such recourse and must instead report the whole sensitivity curve.

If $I_\Delta = [L_\Delta, U_\Delta]$ covers Δ_B, and $I_C = [L_C, U_C]$ covers b_C, then the sensitivity interval is

$$[L_\Delta - U_C - \rho,\ U_\Delta - L_C + \rho].$$

When the component intervals have error probabilities α_Δ and α_C, the sensitivity interval has coverage at least $1 - \alpha_\Delta - \alpha_C$ under the transport assumption.

4.3 Placebo-label check

For a private parent with (K+1) forms, one form is assigned the pseudo-original role uniformly at random after responses are collected. If J is that label,

$$D^{\mathrm{pl}}(J) = Y_J - \frac{1}{K}\sum_{k \ne J} Y_k.$$

Averaging over all possible labels gives exactly zero. This provides a randomization check for code, scoring, and form-role handling. It does not show that a historically public original is exchangeable with a new form.

5 Controlled-exposure validation

5.1 Registered design

The validation trains a small open model under four training mixtures and asks whether the audit sees what was planted. The four conditions are: a clean control with unrelated instruction data only; a related-task control with arithmetic practice that shares no item with the audit bank; exact exposure, which adds the public forms of a random half of the 120 audit parent items to the related-task mixture; and paraphrase exposure, which adds fresh forms of the same half instead. Every condition uses 1,200 training records and three epochs. The two exposure arms use the same 60 parent-item identifiers. Each condition is trained under ten seeds, giving 40 model fits per family.

The analysis was locked before any condition was run (`PAPER10_ANALYSIS_LOCK.md`). The trained model run, not the question row, is the unit of analysis. For each seed and condition the parent-item gaps are averaged to give one public-form advantage; the two registered contrasts are the seed-paired differences

Table 1: Registered seed-paired contrasts in the controlled-exposure validation (ten seeds per family).

Family	Contrast	Mean	95% CI	Seeds positive	p (two-sided)
SmolLM2-360M	Exact exposure minus related-task	+0.020	[+0.006, +0.034]	9/10	0.010
SmolLM2-360M	Paraphrase exposure minus related-task	-0.002	[-0.009, +0.005]	3/10	0.61
Qwen2.5-0.5B	Exact exposure minus related-task	+0.043	[-0.018, +0.104]	7/10	0.14
Qwen2.5-0.5B	Paraphrase exposure minus related-task	-0.015	[-0.045, +0.016]	3/10	0.31

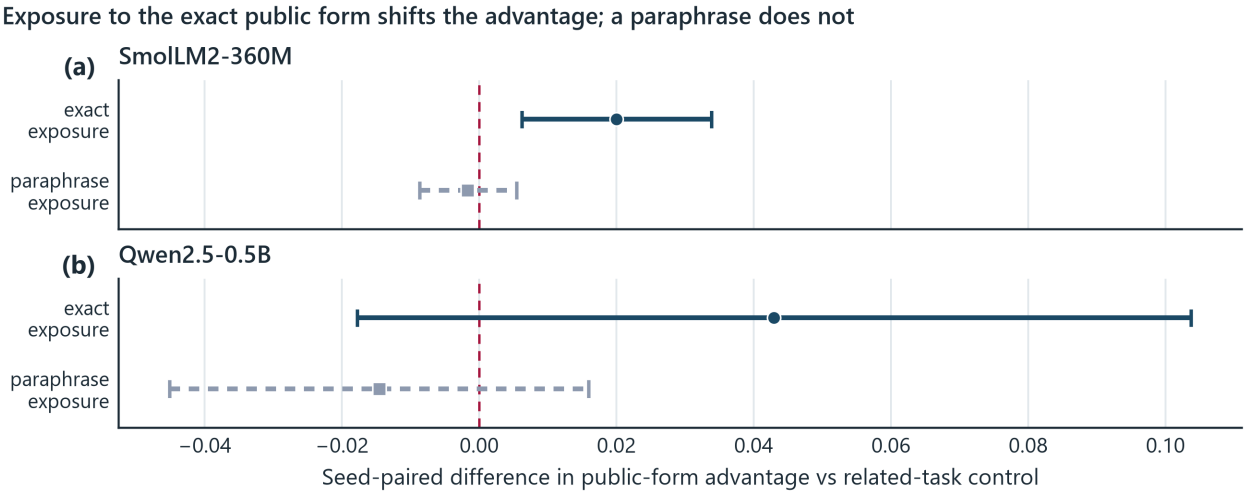


Figure 3: Registered seed-paired contrasts, ten seeds per family. Exact exposure to public forms shifts the public-form advantage upward in both families, decisively in SmolLM2-360M; exposure to fresh forms of the same items does not. The wide Qwen2.5-0.5B interval reflects that model's larger run-to-run variation, not a weaker effect.

of exact exposure and of paraphrase exposure against the related-task control. The lock also stated that exposure to one arithmetic item may help related unexposed items, so exposed-versus-unexposed comparisons inside an arm are diagnostics, not causal estimates.

The registered model was Qwen2.5-0.5B-Instruct. An independent replication in a second family, SmolLM2-360M-Instruct, used the identical design, seeds, and lock. All 80 fits completed on Kaggle T4 hardware with transformers 4.57; full per-run artifacts, exposure assignments, token counts, and driver logs are archived.

5.2 Results

Table 1 gives the registered contrasts.

Two patterns hold in both families. Training on the public forms raises the public-form advantage, and training on fresh forms of the same items does not. In SmolLM2-360M the exact-exposure contrast is decisive: nine of ten seeds are positive and the interval excludes zero. In Qwen2.5-0.5B the contrast has the same direction and a larger point estimate but a wider interval, because that model's run-to-run variation is about four times larger (clean-arm standard deviation across seeds 0.061 versus 0.016). We report Qwen2.5-0.5B as direction-consistent and underpowered, not as a success. Figure 3 puts all four contrasts on one axis.

The exposed-versus-unexposed diagnostic inside the exact-exposure arm is +0.018 in Qwen2.5-0.5B and -0.005 in SmolLM2-360M. The advantage spreads across the whole item bank rather than concentrating on the 60 planted parents. This is the spillover pattern the lock named in advance as the expected policy-level reading: exposure to public arithmetic items teaches the public surface form of the whole bank. It also means that an audit which samples unexposed items from a contaminated benchmark can still see part of the effect.

Table 2: Coverage and width of the two intervals in the finite-population calibration, 5,000 replications per design.

Design	Sample size	Hoeffding coverage	Hoeffding half-width	Wald coverage	Mean Wald half-width
Simple random sample	120	1.0000	0.2480	0.9476	0.0625
Simple random sample	400	1.0000	0.1358	0.9506	0.0321
Simple random sample	800	1.0000	0.0960	0.9522	0.0203
Simple random sample	1,200	1.0000	0.0784	0.9468	0.0144
Stratified sample	400	1.0000	0.1358	0.9488	0.0329

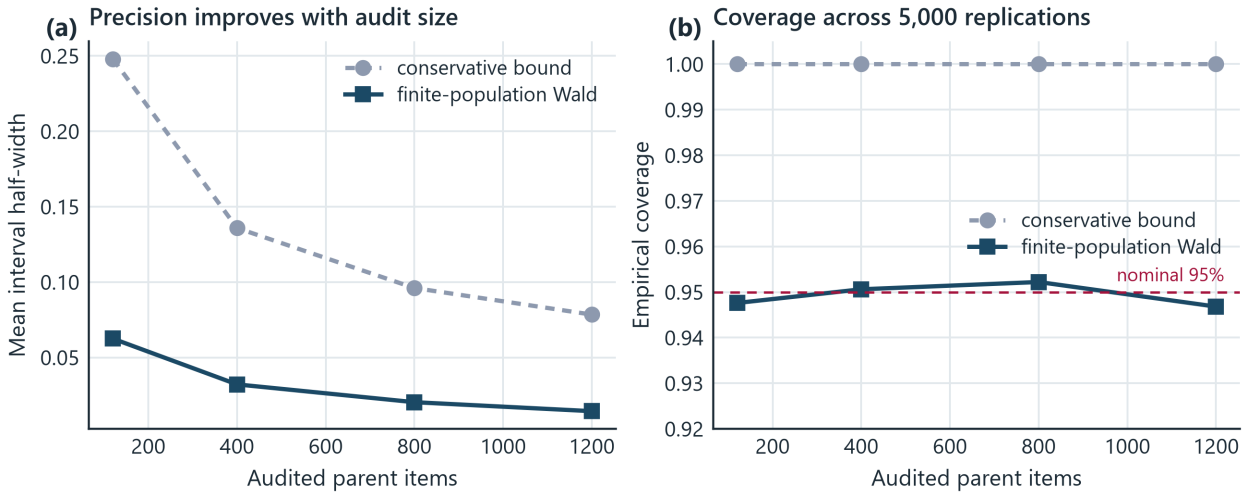


Figure 4: Interval precision and empirical coverage in the finite-population simulation. The conservative bound is valid but wide in these settings. The finite-population Wald interval is much sharper and stays close to the nominal 95% level.

The decision rule in the lock asked for a measurable exact-exposure signal and for controls that behave as expected before the design was applied to a real benchmark. Both conditions are met.

6 Design simulation

6.1 Setup

We generated realized finite populations of 2,400 bounded parent-level gaps and repeatedly sampled them without replacement. The first calibration used 5,000 replications for simple and stratified sampling. It compared the conservative Hoeffding interval with the finite-population Wald interval. A separate 3,000-replication stress grid varied the true and assumed transport radius.

6.2 Sampling coverage

Table 2 gives the first calibration.

The conservative interval covered every replication, but its width made it weak for moderate score gaps. The Wald interval stayed close to nominal coverage and was much narrower in these settings (Figure 4). This is a design result, not evidence about contamination in a real model.

6.3 Transport-radius failure analysis

The sensitivity result is only valid when the stated transport radius is large enough. This was visible in the stress test. With a true control-to-benchmark mismatch of 0.05 and 800 sampled items in each bank,

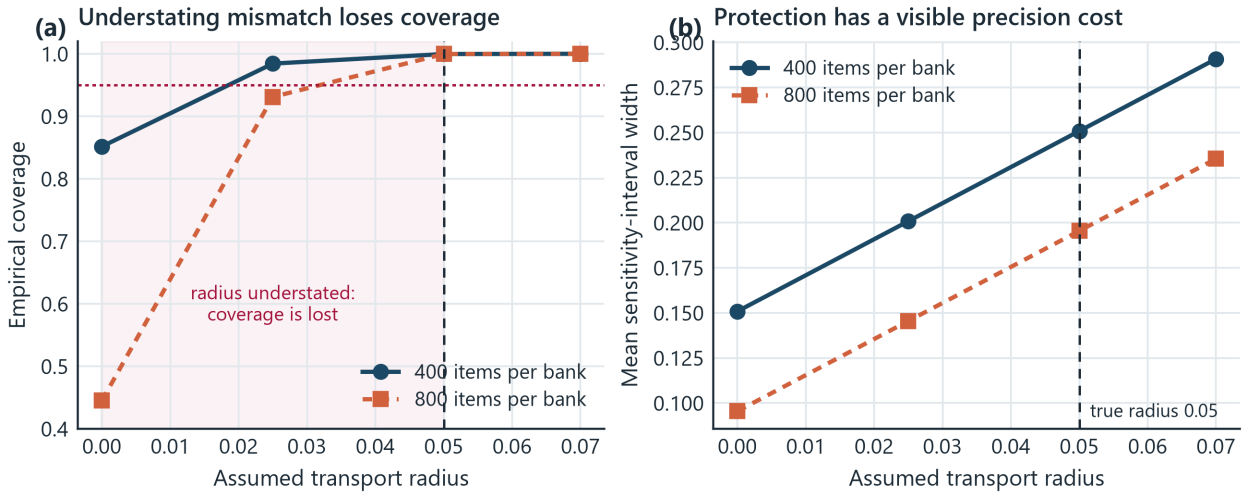


Figure 5: Sensitivity to the transport radius when the true control-to-benchmark mismatch is 0.05. Understating the radius can cause severe undercoverage. A radius that covers the true mismatch restores protection but widens the interval.

empirical coverage was 0.445 when the assumed radius was zero. Coverage rose to 0.931 when the assumed radius was 0.025 and to 0.9997 when the assumed radius matched the true value 0.05. The corresponding mean interval widths were about 0.095, 0.145, and 0.195. Figure 5 shows both sides of the trade at once: what an understated radius costs in coverage, and what honest protection costs in width.

This is not a defect that can be removed by wording. It is the central sensitivity tradeoff. A larger radius protects coverage but gives a wider conclusion.

6.4 Placebo identity

The exact placebo-label identity was checked over 500 randomly generated four-form outcome vectors. The largest numerical residual was below 6×10^{-17}.

6.5 Extended interval stress test

We next tested 36 simple-random-sampling settings formed from six bounded outcome shapes, two benchmark sizes, and audit fractions of 5%, 20%, and 50%. Each setting used 2,000 repeated audits. The shapes included sparse, near-null, highly variable two-point, low-variance, balanced, and right-skewed populations. Three additional designs used strongly imbalanced strata and different sample allocations.

The conservative interval had coverage from 0.9935 to 1.0000 across the 36 settings. Wald coverage ranged from 0.9145 to 0.9575. Six settings fell below 0.94 and two fell below 0.93. The worst case was a 50-item audit of a sparse 1,000-item population. The near-null population at the same audit size had coverage 0.9280. In the imbalanced-stratum experiment, Wald coverage was 0.9440 under proportional allocation, 0.9475 under equal allocation, and 0.9490 when the rare stratum was oversampled.

We then checked the main failure case without Monte Carlo error. For four discrete finite populations, we enumerated every possible sample composition and weighted it by its exact multivariate hypergeometric probability. In the sparse 1,000-item population with a 50-item audit, exact Wald coverage was 0.9127, close to the simulated value of 0.9145. Exact coverage rose to 0.9467 at 200 audited items and 0.9499 at 500. The calculation confirms that the small-sparse warning is not a simulation accident. The conservative interval had at least 0.9968 exact coverage in all enumerated settings. Figure 6 lays the 5,000-item settings out as a grid, shaded by distance from the nominal level rather than by coverage itself, because coverage above 0.95 means a wider interval and not a better one.

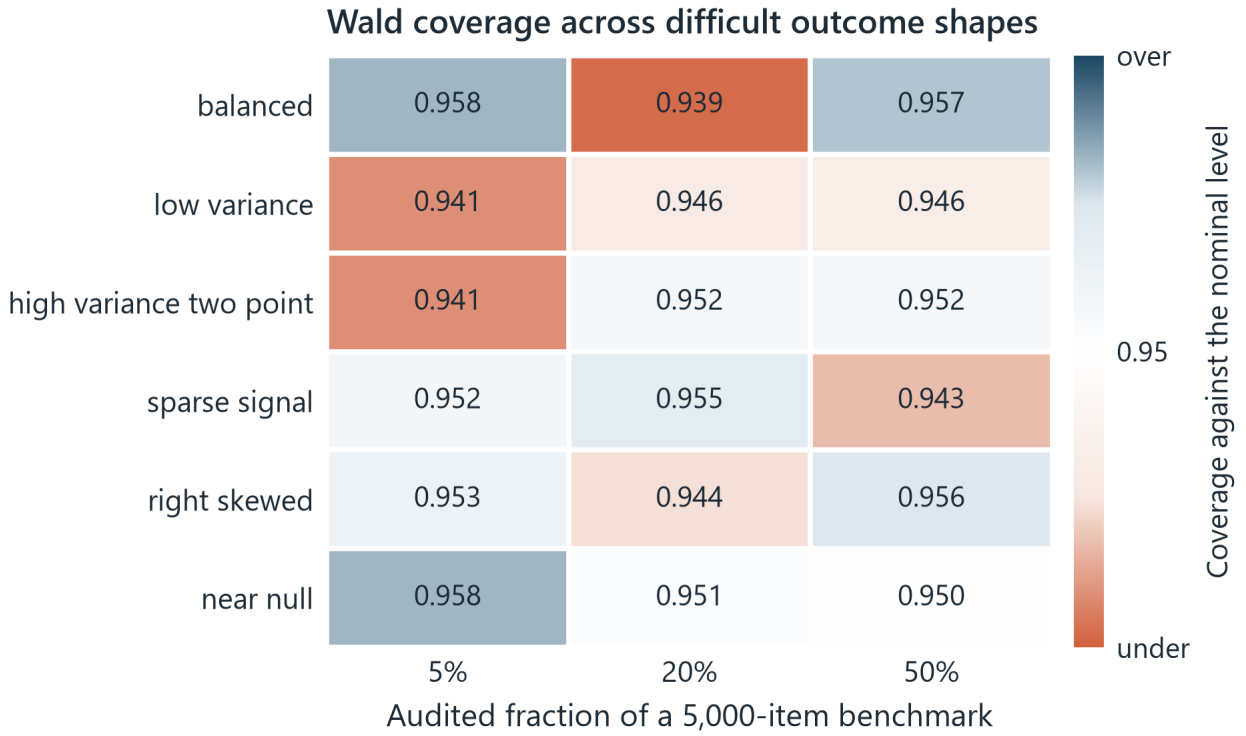


Figure 6: Wald coverage for the 5,000-item benchmark settings. Results are generally close to 95%, but the wider stress grid shows that this approximation is not uniformly reliable. Small sparse audits are the important failure case.

This changes the operating rule. The Wald interval is a secondary precision summary. When fewer than 200 parent items are audited, or when fewer than 30 sampled parent items have a nonzero gap, the distribution-free interval is the primary inferential result. Both intervals remain visible in every analysis. The thresholds are practical guardrails supported by the stress tests; they are not universal mathematical cutoffs.

7 Real-benchmark audit

7.1 Registered protocol

The audit protocol was locked before any audited model saw any audit item (`REAL_AUDIT_REGISTRATION.md`). Two real benchmarks were used. The primary benchmark is the GSM8K test split (N = 1,319). The secondary, held-out domain is the ARC-Challenge test split restricted to four-option items (N = 1,165). From each frame 200 parent items were drawn without replacement with a single seeded generator. For every parent, two fresh forms were written under fixed rules: for GSM8K, change the names, entities, and sentence structure but keep every numeric quantity, the logical structure, the units, and the gold answer; for ARC, rewrite the question stem only and keep the four options and the key verbatim.

Variants were drafted by a language model from a provider and family disjoint from every audited model, then checked by a deterministic script that verifies the multiset of numeric tokens, the gold answer, length bounds, and, for ARC, option integrity and a token-dissimilarity threshold against the original stem. Twelve ARC rewrites failed the dissimilarity check on the first pass and were rewritten before any model was queried. The final bank has 600 GSM8K forms and 600 ARC forms.

We are explicit about which of the two checks in Section 3.1 stands behind these numbers, and about the order in which they happened. The mechanical check was run on every form before any model was queried and is reproducible from the released bank. The authors' reading was completed after the results were

Table 3: Registered real-benchmark audit, all five models. $\hat{\Delta}$ is the public-form advantage over 200 sampled parents. The half-width is the primary interval: the finite-population Wald interval, or the Hoeffding bound (marked H) where the nonzero-gap guardrail applies. b_C is the control-bank form gap with its Wald half-width. No model-benchmark pair meets the registered claim rule.

Model	Benchmark	Public score	Fresh score	$\hat{\Delta}$	Half-width	Nonzero gaps	b_C (control)
OLMo-2-7B	GSM8K	0.720	0.705	+0.015	0.045	58	+0.017 ± 0.081
OLMo-2-7B	ARC	0.625	0.608	+0.018	0.040	40	+0.013 ± 0.102
SmolLM2-1.7B	GSM8K	0.475	0.430	+0.045	0.058	83	0.000 ± 0.119
SmolLM2-1.7B	ARC	0.485	0.470	+0.015	0.192 (H)	26	-0.088 ± 0.110
Phi-3-mini	GSM8K	0.820	0.803	+0.018	0.035	35	+0.008 ± 0.049
Phi-3-mini	ARC	0.595	0.598	-0.003	0.192 (H)	27	0.000 ± 0.105
Qwen2.5-1.5B	GSM8K	0.475	0.500	-0.025	0.055	80	+0.033 ± 0.093
Qwen2.5-1.5B	ARC	0.495	0.503	-0.008	0.029	32	+0.025 ± 0.093
Qwen2.5-7B	GSM8K	0.710	0.715	-0.005	0.047	59	-0.017 ± 0.052
Qwen2.5-7B	ARC	0.590	0.558	+0.033	0.036	38	+0.025 ± 0.086

produced, which is a real limitation of sequence and is recorded as one: it could not have changed which items were audited, only what is now known about them. It was a prioritised reading rather than a reading of all 1,500 forms, working from the ranked packet released with the paper, which renders the parents most likely to be defective together with every GSM8K rewrite sitting at or above the dissimilarity threshold. It found no form that changes its parent's answer, drops information the question needs, or adds a hint, and one defect in an ARC rewrite, described in Section 7.7 and left in the frozen bank. What the mechanical check guarantees is that a fresh GSM8K form carries its parent's quantities and answer and that a fresh ARC stem leaves the options and key untouched; what the reading adds is judgment about naturalness and about information a rewrite may have quietly dropped.

Two private control banks were composed from scratch and have never been public: 60 GSM8K-style word problems and 40 ARC-style science questions, each with a public-like first form and two fresh forms written under the same rules (300 forms). Because no audited model can have seen them, they estimate each model's ordinary form gap b_C under this writing protocol.

Five open-weight models with pinned revisions were registered: Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct, Phi-3-mini-4k-instruct, SmolLM2-1.7B-Instruct, and OLMo-2-1124-7B-Instruct. The last is included because its pretraining and post-training mixtures are public, which gives one model in the panel whose data provenance a reader can check independently of the audit.

GSM8K items were answered once each by greedy decoding through the model's chat template with a fixed instruction and a 320-token budget, and scored by exact numeric match after a registered parser. ARC items were scored without generation, by the character-normalized log probability of each option text given the stem; the raw sum was recorded as a secondary scoring.

The registered inference per model and benchmark is: the sample mean of the parent gaps D_i; the Hoeffding interval; the finite-population Wald interval; the guardrail of Section 6.5, which makes the distribution-free interval primary when fewer than 200 parents are audited or fewer than 30 sampled parents have a nonzero gap, of which only the second clause can bind here because every audit samples exactly 200 parents; a Wald interval for b_C from the control bank; and the Proposition 3 sensitivity interval for the exposure-related component over the grid ρ in 0, 0.01, 0.02, 0.05, 0.10, with $\rho^* = 0.02$ as the headline radius. A model-benchmark pair is reported as showing an exposure-consistent public-form advantage only if the primary interval for Δ excludes zero and the sensitivity interval at ρ^* excludes zero. The placebo-label identity is computed on every control parent.

7.2 Primary results

All five registered models completed under the registered settings. Table 3 reports the primary analysis.

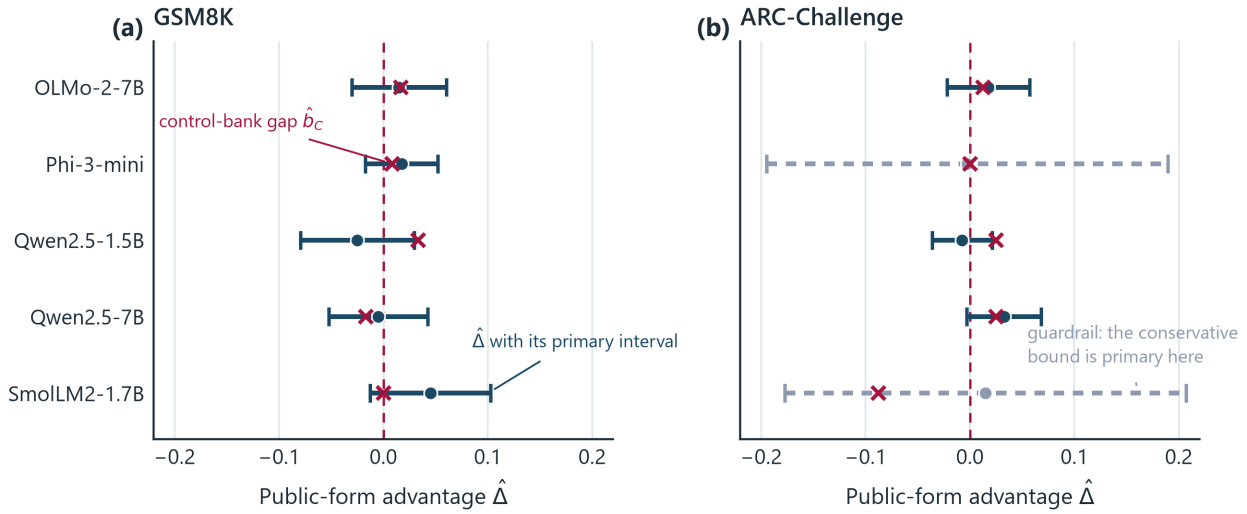


Figure 7: Public-form advantage with its primary interval for every model-benchmark pair, with the control-bank gap b_C marked separately. Dotted intervals are the two pairs where the registered nonzero-gap guardrail makes the conservative bound primary. Every interval covers zero.

Under the registered rule, no audited model shows an exposure-consistent public-form advantage on either benchmark. The ten point estimates are small and of both signs, and the largest, Qwen2.5-7B on ARC at +0.033, has a half-width of 0.036. The most precise GSM8K statement is for Phi-3-mini: +0.018 with half-width 0.035, so any surface-form inflation of its GSM8K score is below about five points. OLMo-2-7B, the one model whose training data are public, gives +0.015 and +0.018 with half-widths 0.045 and 0.040, so it sits in the same range as the models whose data are not documented. The placebo-label identity returned exactly zero for every model and domain, so the scoring pipeline and the pseudo-original role assignment introduce no asymmetry. Figure 7 puts all ten estimates on one axis with their control-bank gaps beside them.

On ARC the rewrites rarely change a log-probability scorer's choice: only 26 to 40 of 200 parents show any gap, which triggers the registered guardrail for two pairs. For the Wald-primary ARC pairs the intervals are narrow, with half-widths from 0.029 to 0.040. The sensitivity intervals for the exposure-related component τ at $\rho^* = 0.02$ include zero for every pair; their widths, from 0.23 to 0.44, are set by the control-bank intervals rather than by the audit intervals (Section 7.4).

7.3 Amendment 1: answer length and answer format

The saved outputs showed a problem with the registered generation budget. The two Qwen models hit the 320-token cap on 40.7% (1.5B) and 29.8% (7B) of GSM8K forms and wrote the registered answer phrase on only 2.7% and 8.8% of forms, using a boxed LaTeX format instead; OLMo-2-7B truncated on 25.3%, SmolLM2-1.7B on 14.5%, and Phi-3-mini on 11.3%. Truncation censors the outcome on exactly the items that need long reasoning, and it does so for both forms of a parent, which pulls the measured gap toward zero rather than adding symmetric noise. It also explains why the Qwen public scores in Table 3 sit well below their published GSM8K figures.

Before any rerun was started, Amendment 1 was logged in the audit trail: rerun the two Qwen models with a 768-token budget and a parser that also accepts a boxed answer, and report the result beside, never instead of, the registered primary. Table 4 gives the outcome.

Removing the budget raises Qwen2.5-7B's GSM8K score from 0.710 to 0.930, which is close to its published figure, and raises Qwen2.5-1.5B's from 0.475 to 0.695. The conclusion does not move: both gaps remain small and negative, and neither interval excludes zero. The null result in Table 3 is therefore not an artifact of the truncation that biased it toward zero. As a determinism check, the ARC results in the rerun are identical to

Table 4: Amendment 1: the two Qwen models on GSM8K with a 768-token budget and a boxed-answer parser. Truncation falls to zero. ARC is unaffected by both changes and was reproduced bit-for-bit.

Model	Setting	Public score	Fresh score	$\hat{\Delta}$	Half-width	Truncated
Qwen2.5-7B	Registered (320 tokens)	0.710	0.715	-0.005	0.047	29.8%
Qwen2.5-7B	Amendment 1 (768 tokens)	0.930	0.938	-0.008	0.192 (H)	0.0%
Qwen2.5-1.5B	Registered (320 tokens)	0.475	0.500	-0.025	0.055	40.7%
Qwen2.5-1.5B	Amendment 1 (768 tokens)	0.695	0.735	-0.040	0.054	0.0%

the primary run form by form, which is expected because ARC scoring involves neither generation nor the parser.

Two other things follow. First, published accuracy figures are not comparable to the public scores in Table 3, and an auditor who wants comparable absolute scores must give the model room to finish. Second, the registered instruction to answer in a fixed phrase was largely ignored by both Qwen models, so an audit protocol should either accept the formats models actually use or verify format compliance before scoring.

7.4 What the negative-control bank does and does not buy

The control banks behaved as designed. Control public-like forms and control fresh forms were answered at nearly the same rate by every model, so b_C is close to zero everywhere (Table 3, last column), and the sensitivity interval for τ is essentially the audit interval widened by the control interval and the radius. One detail of that combination matters to anyone reproducing it. Proposition 3 pays for two coverage events with a union bound, so the registered analysis splits the error budget evenly and builds both component intervals at $\alpha_\Delta = \alpha_C = 0.025$, giving the combined interval 95% coverage. Those components are wider than the $\alpha = 0.05$ half-widths reported in Table 3, which are the primary intervals for Δ on its own; a reader who substitutes the table's numbers into Proposition 3 will get an interval that is too narrow. But the control intervals are wide, from 0.049 to 0.119, because the banks hold 40 to 60 parents against 200 in the audit. The consequence is visible in the τ intervals: even where the audit interval has half-width 0.035, the sensitivity interval spans more than 0.2. The lesson is a design rule rather than a flaw in the idea: the control bank must be at least as large as the audit sample, and its size should be chosen from the audit's target precision instead of being treated as a small add-on. A bank of 200 control parents would bring the control half-width down to the audit's, roughly halving the width of the τ interval.

One inconsistency in the registered analysis is worth stating, because it matters for reuse even though it changes nothing here. The Δ component of the sensitivity interval is the finite-population Wald interval for every pair, including the two ARC pairs where the guardrail makes the distribution-free bound primary for Δ on its own. Those two pairs are exactly the sparse regime in which Section 6.5 found the Wald interval under-covering, so on them the sensitivity interval inherits the weakness the guardrail exists to avoid, and it does so in the direction that makes excluding zero easier rather than harder. Recomputing those two with the distribution-free component at the same $\alpha/2$ roughly doubles them, from a width of 0.355 to 0.711 for SmolLM2-1.7B and from 0.345 to 0.700 for Phi-3-mini. Neither excludes zero under either construction, so no reported conclusion depends on the choice; but an audit whose claim rule might actually fire should use the guardrail bound in both places, and we recommend registering it that way.

The same guardrail question can be asked of the control component, and it has an uncomfortable answer that we work through rather than leave implicit. The audit switches to a distribution-free interval below 200 parents because Wald under-covers there; the control banks hold 40 and 60 parents and are summarised with a Wald interval anyway. The banks are also composed rather than drawn from any frame, so it is fair to ask what population a sampling interval on them describes. Three readings are defensible and Table 5 reports all of them.

The guardrail reading is honest and nearly useless: a control half-width of 0.4 on a quantity bounded in [-1, 1] leaves a sensitivity interval that could not exclude zero for any plausible audit, which is a real cost of composing a bank of 40 items rather than 200.

Table 5: The sensitivity interval under three readings of the control bank, at the registered radius. *Wald* is the registered construction. *Hoeffding* applies the paper's own small-sample guardrail to the control bank as well. *Fixed* treats the bank as a fully measured object, so b_C is a constant and the transport radius carries everything.

Reading of the control bank	**Control half-width**	**Mean width of the τ interval**	**Pairs excluding zero**
Wald (registered)	0.057 to 0.136	0.335	0 of 10
Hoeffding (own guardrail)	0.382 to 0.468	0.982	0 of 10
Fixed bank, no interval	0	0.131	1 of 10

The fixed reading is the interesting one, because it is narrowest and it is wrong. Dropping the control interval makes one pair, SmolLM2-1.7B on ARC, exclude zero at [+0.05, +0.15], which under the claim rule would be a positive finding. That flag comes entirely from the control bank's measured form gap of -0.088 on 40 items, a number whose own interval comfortably contains zero. Treating it as a known constant converts control-bank noise into a contamination claim about a model. So the option of declaring the bank a fixed finite target, which removes the awkward superpopulation question, is not available: whatever the bank is a sample of, it is small enough that pretending it carries no uncertainty manufactures findings.

That leaves the registered construction, and we describe it as what it is. The Wald interval on b_C treats the bank as one realisation of the item-writing procedure that produced it. That population is informal, the interval is therefore a heuristic rather than a guarantee, and the transport radius is what carries the assumption across to the audited benchmark. Two of the three readings agree on all ten pairs, and the third disagrees for a reason that argues against itself. What the exercise establishes is narrower than a guarantee and worth having anyway: the nulls of Table 3 do not depend on the control interval being tight, and the one construction that would overturn a null does so by ignoring uncertainty rather than by measuring it.

7.5 A variance-adaptive bound for near-ceiling models

The audit exposed one place where the registered machinery answers honestly but almost uselessly. When a model is near ceiling on a benchmark, very few parents have any gap at all: Qwen2.5-7B under Amendment 1 has 16 nonzero gaps out of 200. The guardrail then makes the distribution-free Hoeffding bound primary, and at n = 200 that bound is ± 0.192 whatever the data look like, because it uses only the range of D_i.

A variance-adaptive distribution-free bound addresses this without giving up the distribution-free guarantee. For D_i in [-1, 1], the empirical Bernstein inequality (Maurer and Pontil, 2009) applied to the rescaled variable gives, with probability at least 1 - α,

$$|\widehat{\Delta} - \Delta_N| \leq \sqrt{\frac{2s^2 \log(4/\alpha)}{n}} + \frac{14 \log(4/\alpha)}{3(n-1)},$$

where s^2 is the sample variance of the parent gaps. The log term is $\log(4/\alpha)$ rather than $\log(2/\alpha)$ because the Maurer-Pontil result is one-sided: covering both tails costs $\alpha/2$ in each direction. Sampling without replacement is no less concentrated than sampling with replacement from the same finite population, so the bound transfers by the comparison used for Proposition 1. The derivation is Proposition 5 in the mathematical core.

Section 6 sets the standard that CleanScore does not report an interval whose coverage it has not simulated, so we hold this one to it. We reran the extended stress design of Section 6.5, the same six bounded outcome shapes, two benchmark sizes, and three audit fractions, with 4,000 repeated audits per setting, and recorded the coverage and mean width of all three intervals. Table 6 summarises them.

The Bernstein bound never under-covers: its minimum coverage over all 36 settings is 0.9998, and like the Hoeffding bound it is conservative. But it is not uniformly narrower, and the pattern of when it wins is the useful part. It is narrower than Hoeffding in 23 of the 36 settings, with the ratio of widths running from 0.40 to 2.17. The additive term 14 $\log(4/\alpha)$ / (3(n-1)) dominates at small audit sizes, so at n = 50 the Bernstein

Table 6: Coverage of the three intervals across the 36 stress settings (4,000 audits each), and the width of the Bernstein bound relative to the Hoeffding bound at the audit's own sample size, n = 200 in a 1,000-item benchmark.

Interval	Coverage range	Settings below 0.95	Width vs Hoeffding at n = 200
Hoeffding	0.9945 to 1.0000	0	1.00 by definition
Empirical Bernstein	0.9998 to 1.0000	0	0.77 to 1.62, depending on the shape
Finite-population Wald	0.9390 to 0.9563	24	not distribution-free

Table 7: Baseline detectors on the audited items. PaCoST flags a pair when the paired confidence test rejects at 0.05 with higher confidence on the public form.

Model	Benchmark	Public score	Fresh score	Confidence, public	Confidence, fresh	PaCoST p	PaCoST flags	CleanScore claim
OLMo-2-7B	GSM8K	0.720	0.705	-0.098	-0.096	0.28	no	no
OLMo-2-7B	ARC	0.625	0.608	-3.19	-3.17	0.71	no	no
SmolLM2-1.7B	GSM8K	0.475	0.430	-0.156	-0.156	0.83	no	no
SmolLM2-1.7B	ARC	0.485	0.470	-3.74	-3.67	0.18	no	no
Phi-3-mini	GSM8K	0.820	0.803	-0.119	-0.119	0.94	no	no
Phi-3-mini	ARC	0.595	0.598	-3.15	-3.20	0.017	**yes**	no
Qwen2.5-1.5B	GSM8K	0.475	0.500	-0.122	-0.122	0.99	no	no
Qwen2.5-1.5B	ARC	0.495	0.503	-2.82	-2.84	0.68	no	no
Qwen2.5-7B	GSM8K	0.710	0.715	-0.075	-0.076	0.43	no	no
Qwen2.5-7B	ARC	0.590	0.558	-6.09	-6.15	0.49	no	no

bound is wider in every shape. At the audit's own size of n = 200 it is 23% narrower on the low-variance shape, 12% narrower on the sparse and near-null shapes, level on the right-skewed shape, 12% wider on the balanced shape, and 62% wider on the high-variance two-point shape.

That pattern lines up with the guardrail. The guardrail fires when fewer than 30 of 200 sampled parents have a nonzero gap, which is precisely the concentrated, low-variance regime in which the Bernstein bound is narrower; the case where it loses, a two-point population splitting its mass between -1 and +1, is a regime the guardrail never routes to it. On the audited data the Bernstein bound is narrower than Hoeffding for eight of the ten model-benchmark pairs, giving half-widths from 0.148 to 0.197 against a constant 0.192. The two exceptions are the two GSM8K pairs with the most nonzero gaps, SmolLM2-1.7B and Qwen2.5-1.5B, and both are Wald-primary, so the guardrail would never route to the Bernstein bound there in any case. On both pairs where the guardrail does fire it is narrower. In the Amendment 1 run for Qwen2.5-7B, the near-ceiling case that triggers the guardrail, it gives 0.144 against 0.192, a 25% reduction. This is a property of these data and of the low-variance regime, not a general ordering of the two bounds.

We report this as an observation rather than as a result of the registered analysis, and we do not substitute it for the registered primary. Choosing the smaller of two intervals after seeing both would not retain the nominal level, so the two are reported side by side and the guardrail continues to use the bound that was registered. The recommendation for future audits is specific: register the empirical Bernstein bound as the distribution-free interval on the guardrail branch, where the low-variance condition that triggers the guardrail is the same condition that makes the bound narrower, and keep the Hoeffding bound elsewhere.

7.6 Baselines on the same items

Table 7 compares CleanScore with two detectors computed on exactly the same forms. The Clean-Eval-style statistic is the fresh-form score itself. The PaCoST-style statistic pairs, per parent, the model's mean per-token log probability on the public form with its mean over the fresh forms (the greedy continuation for GSM8K, the chosen option for ARC), and tests the paired difference.

The two detectors agree on nine of ten pairs. They disagree on Phi-3-mini and ARC: the model is measurably more confident on the public stems (paired t-test p = 0.017; Wilcoxon p = 0.010), yet its accuracy on public and fresh stems is the same, -0.003 ± 0.028. Both readings are correct. The model recognizes the surface of the public stems, and that recognition does not change which option it picks. A confidence-based detector reports the first fact; CleanScore reports the second. For someone reading a leaderboard, only the second one changes what the score means. This separation of a confidence shift from a score shift is, we think, the most

Table 8: Parents on which at least two of the five models show the pattern, against a permutation null with 20,000 draws.

Benchmark	Pattern	Observed	Null mean	p
GSM8K	rewrites harder	4	3.37	0.45
GSM8K	rewrites easier	2	3.29	0.88
ARC	rewrites harder	3	0.70	0.02
ARC	rewrites easier	0	0.30	1.00

useful thing the comparison shows, and it is exactly the distinction a verdict-style detector cannot make on its own.

Multiplicity belongs in this comparison and we had left it out. Ten model-benchmark pairs are tested and no correction was applied to any of them. For the audit's own results the omission is harmless in the direction that matters, because correcting for ten tests can only make ten nulls harder to overturn, and none of them was close. It is not harmless for the detector's single flag: at $p = 0.017$ the Phi-3-mini result does not survive a Bonferroni threshold of 0.005 across ten pairs, and under Benjamini-Hochberg it is the smallest of ten p-values against a threshold of 0.005 and does not survive either. So the flag should be read as what a detector reports before correction, not as an established confidence shift, and the point the comparison makes stands on the structure of the disagreement rather than on that p-value. Any audit that reports across a panel should fix a correction in its registration; ours did not, and the omission is recorded here rather than repaired after the fact.

7.7 Do rewrite failures cluster on particular items?

The audit rests on the fresh forms being faithful, and the authors' reading came after the results rather than before them. That makes it worth asking what the results themselves say about the items, independently of anyone's reading. A rewrite that is harder than its original, or that drops information the question needs, should trip more than one model. Two independent models failing the same parent is evidence about the item; one model failing it is evidence about the model.

For each parent we counted the models that answered the public form correctly and both fresh forms incorrectly, and tested the count of parents with at least two such models against a null that permutes each model's failures across parents. The permutation keeps how often each model fails and destroys any alignment between models. We ran the mirror-image test as well, for parents where the public form was answered incorrectly and both rewrites correctly, which is what a rewrite that leaked a hint would look like. Table 8 gives both tests.

On GSM8K, the benchmark that carries the paper's primary claim, failures do not cluster: four parents show the pattern where chance predicts 3.4, and the mirror test is if anything below chance. Whatever variation the rewrites introduce is not concentrated on particular items, which is the pattern a bank of systematically defective rewrites would not produce.

On ARC three parents cluster where chance predicts 0.7. All three rewrites preserve the information, the options, and the key, and none adds a hint, and what the three share is a short or weakly cued stem, including one whose original is the seven-word fragment "The suspension system of a truck includes the". Under length-normalized option scoring, a stem that carries little information leaves the choice to be settled largely by the option text, so small changes in wording can move the argmax without any change in what the question asks. That is a property of the scoring protocol, and it is a further reason to treat the generation-and-exact-match arm on GSM8K as the more informative of the two.

It is not the whole story for one of the three, and we correct an earlier reading of our own here. The second rewrite of that seven-word parent begins "Included in the suspension system of a truck is the", where every option is a compound plural, so the completion is ungrammatical whichever option is chosen and the stem should read "are the". Being wrong for all four options about equally, it is unlikely to favour any particular one, but an ungrammatical stem makes that fresh form noisier than its original for a reason that has nothing

Table 9: Sensitivity of the GSM8K audit to rewrite depth. The third column drops the 21 parents whose closest fresh form meets or exceeds the ARC dissimilarity threshold. The last column is the correlation, across all 200 parents, between a parent's rewrite similarity and its measured gap.

Model	**$\hat{\Delta}$, all 200**	**$\hat{\Delta}$, deep rewrites only**	**Shift**	**r (similarity, gap)**	**p**
OLMo-2-7B	+0.015	+0.028	+0.013	+0.029	0.69
Phi-3-mini	+0.018	+0.017	-0.001	+0.023	0.75
Qwen2.5-1.5B	-0.025	-0.020	+0.005	-0.035	0.62
Qwen2.5-7B	-0.005	+0.003	+0.008	-0.072	0.31
SmolLM2-1.7B	+0.045	+0.050	+0.005	-0.004	0.96

to do with contamination. So one of the three clustered parents does contain a defect in the item and not only a sensitivity of the scorer. The item is left as it is, for the reason given at the end of this section, and the three parents are listed in the released review packet so a reader can judge them directly.

This diagnostic does not replace the authors' reading, and we do not present it as doing so. It bounds one specific failure mode, systematic defects concentrated on particular parents, and it is what let the reading be prioritised rather than exhaustive: the parents the models disagreed about are the ones a reader most needs to see.

A second failure mode is worth bounding in the same way, and disclosing it requires admitting an asymmetry in the registered verifier. ARC fresh stems are held to a token-dissimilarity rule: a Jaccard overlap of 0.80 or more against the original is a failure, and twelve first-pass rewrites were rejected under it. GSM8K fresh forms were checked for exact identity to the original and never for near-identity, so that threshold was never applied to them. It should have been. Applying it after the fact, 21 of the 200 GSM8K parents have at least one fresh form at or above 0.80, the closest reaching 0.895; on ARC, where the rule was enforced, no parent reaches 0.80 and the closest is 0.789.

This matters in a specific direction. A fresh form that is nearly a copy of its original should be answered nearly like the original whether or not the model has seen either, so its parent's gap is pulled toward zero, and enough such parents would manufacture a null. Table 9 tests that directly by removing every parent above the threshold and recomputing.

The largest shift is 0.013, against half-widths of 0.035 to 0.058, so no conclusion in Table 3 depends on the shallow rewrites. Nor is there a relationship between rewrite depth and the measured gap over the full range: the correlation is within noise for every model on both benchmarks, pooled r = -0.014 on GSM8K and -0.039 on ARC. The 21 near-copy parents in fact show a mean gap of -0.043 against +0.016 for the rest, which is the opposite of the direction the concern predicts and is not significant (Welch p = 0.18).

Reading the 21 shows the defect is systematic rather than scattered, which is worth more than the count. Every one of them is the parent's *first* fresh form, and not one is the second; the median near-copy changes 11% of its words, and the most extreme changes one word in nineteen, turning "A bag has a 5% discount" into "A jacket has a 5% discount". The two variants were drafted by the same rules but the first pass came out as a light substitution of names and objects while the second restructured the sentence, and on these 21 parents the light pass was too light. That also explains why dropping them moves so little: a parent's gap is averaged over both of its fresh forms, so a parent with one near-copy and one genuine rewrite still contributes real signal, diluted rather than absent.

We report this as a defect in the verifier and in the drafting procedure rather than in the bank, and we did not modify the bank. It was frozen before any model was queried, and editing items after seeing the results is exactly the move the registration exists to prevent; the honest response to finding a flaw at this stage is to measure what it cost, which is what this section does. The verifier now computes the GSM8K similarity and reports the 21 items, and an audit built on this code from scratch should apply the threshold to both benchmarks before freezing, and should check that a parent's two rewrites are of comparable depth rather than only that each is not identical to the original.

Table 10: Budgeted-audit curves (mean over models and 2,000 subsamples). "Exact order" is recovery of the full five-model ordering; "pairs correct" is the share of the ten model pairs ordered as the full sample orders them.

Benchmark	Audited items	Mean half-width	Exact order	Pairs correct
GSM8K	25	0.142	0.39	0.92
GSM8K	50	0.101	0.50	0.95
GSM8K	100	0.070	0.62	0.97
GSM8K	150	0.056	0.73	0.98
ARC	25	0.093	0.15	0.78
ARC	50	0.067	0.25	0.86
ARC	100	0.047	0.49	0.94
ARC	150	0.038	0.67	0.97

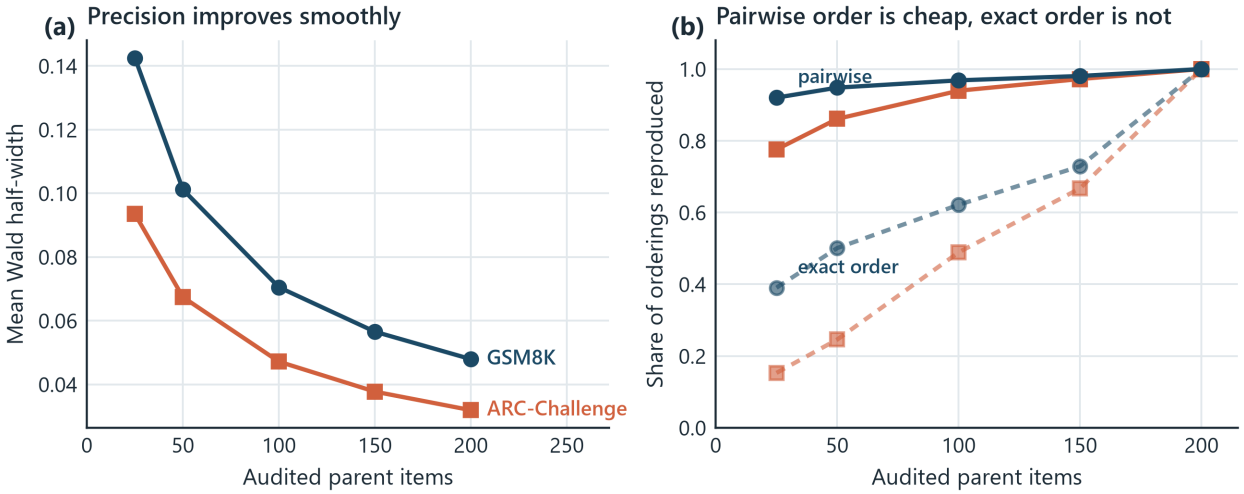


Figure 8: Budgeted audits. Precision improves smoothly with audit size, and pairwise ordering is recovered cheaply, but exact recovery of the five-model ordering lags well behind because two model pairs differ by less than the audit's own half-width.

7.8 Budgeted audits

Contribution 6 asked how small an audit can be. Using the 200 audited parents as a finite population, we drew 2,000 subsamples without replacement at each budget and recorded three things: the finite-population Wald half-width, whether the subsample reproduced the full-sample ordering of all five models by fresh-form score, and the share of the ten model pairs it ordered correctly. Table 10 gives the curves and Figure 8 plots them.

The two ranking columns tell different stories, and the difference is the practical point. Pairwise ordering is cheap: 100 audited GSM8K items order 97% of model pairs as the full audit does, and 100 ARC items order 94%. Recovering the exact ordering of all five models is much harder, 0.62 and 0.49 at the same budget, and it stays below one until the audit is complete.

The reason is visible in the scores rather than in the method. On GSM8K, OLMo-2-7B and Qwen2.5-7B differ by 0.010 in fresh-form score; on ARC, Phi-3-mini and OLMo-2-7B differ by 0.010. Both gaps are smaller than the audit's own half-width at every budget below the full sample, so no honest audit of that size can order those pairs reliably, and one swapped adjacent pair destroys the exact ordering. An auditor should therefore read the budget table as follows: a few dozen items suffice to separate models that differ by more than the half-width in that row, and no feasible audit separates models that differ by less. Reporting the

Table 11: Detection rate for a planted public-form advantage, 4,000 resampled audits per cell, pooling the five audited models within each benchmark.

Benchmark	Audited items	delta = 0.02	delta = 0.05	delta = 0.10	delta = 0.20
GSM8K	50	0.00	0.00	0.00	0.00
GSM8K	100	0.00	0.00	0.00	0.04
GSM8K	200	0.22	0.56	0.95	1.00
GSM8K	400	0.43	0.89	1.00	1.00
GSM8K	800	0.82	1.00	1.00	1.00
ARC	50	0.00	0.00	0.00	0.00
ARC	100	0.00	0.00	0.00	0.01
ARC	200	0.36	0.84	1.00	1.00
ARC	400	0.74	1.00	1.00	1.00
ARC	800	0.99	1.00	1.00	1.00

audit-adjusted score with its interval, rather than a bare rank, makes that limit visible to the reader instead of hiding it inside a leaderboard position.

7.9 What the audit could have detected

A null result is worth only as much as the instrument's ability to have found something, and Section 7.2 reports ten of them. This section asks what a CleanScore audit at the registered size can see, using the variance actually observed in the audited banks rather than an assumed model.

Two quantities. The minimum detectable effect is the primary half-width: the audit declares an advantage when the interval excludes zero, so a true advantage smaller than the half-width will usually not be declared. The detection rate is the probability of declaring an advantage of a stated size, obtained by resampling the observed parent gaps and planting the advantage in a random share of parents that currently show no gap. Planting this way keeps the real bank's variance and its heavy concentration at zero, and it moves items in the direction that exposure to a public form actually moves them, from wrong-on-every-form or right-on-every-form to right-on-the-public-form-only. Table 11 gives the resulting detection rates.

Three things follow, and the first two qualify this paper's own results.

The audit reported here detects a ten-point advantage almost always and a five-point advantage about half the time on GSM8K. At n = 200 the minimum detectable effect runs from 0.035 for Phi-3-mini, the model with the fewest nonzero gaps, to 0.058 for SmolLM2-1.7B, the model with the most; on ARC, from 0.029 to 0.040 for the three pairs where the guardrail does not apply. So the nulls in Table 3 rule out a large surface-form advantage and do not rule out a small one. The correct reading of Phi-3-mini's GSM8K result is that an inflation of ten points would almost certainly have been seen and was not, while an inflation of two or three points would probably have been missed. We have stated the bound this way throughout, and the detection rates say what the bound is worth.

Second, and more sharply: **below 200 audited parents the registered guardrail leaves the audit with no power at all.** At n = 100 the guardrail forces the distribution-free interval, whose half-width is 0.272 whatever the data show, and no advantage in a plausible range clears it; the detection rate is zero for every effect size up to ten points, and 4% at twenty. This is not a failure of the estimator but the price of the guardrail, which exists because Section 6.5 found the Wald interval under-covering in exactly the small, sparse regime. The variance-adaptive bound of Section 7.5 does not rescue it either: at n = 100 its additive term alone is 0.174, so it gives 0.255 against Hoeffding's 0.272. Distribution-free inference on a bounded outcome simply cannot say much from a hundred paired items.

The practical guidance is therefore concrete. An auditor who can only afford tens of items should not expect a contamination audit of this kind to conclude anything, and should not read a wide distribution-free interval as reassurance. Two hundred parents is roughly the point at which the audit becomes informative about ten-point effects; four hundred buys about ninety percent detection at five points on GSM8K and

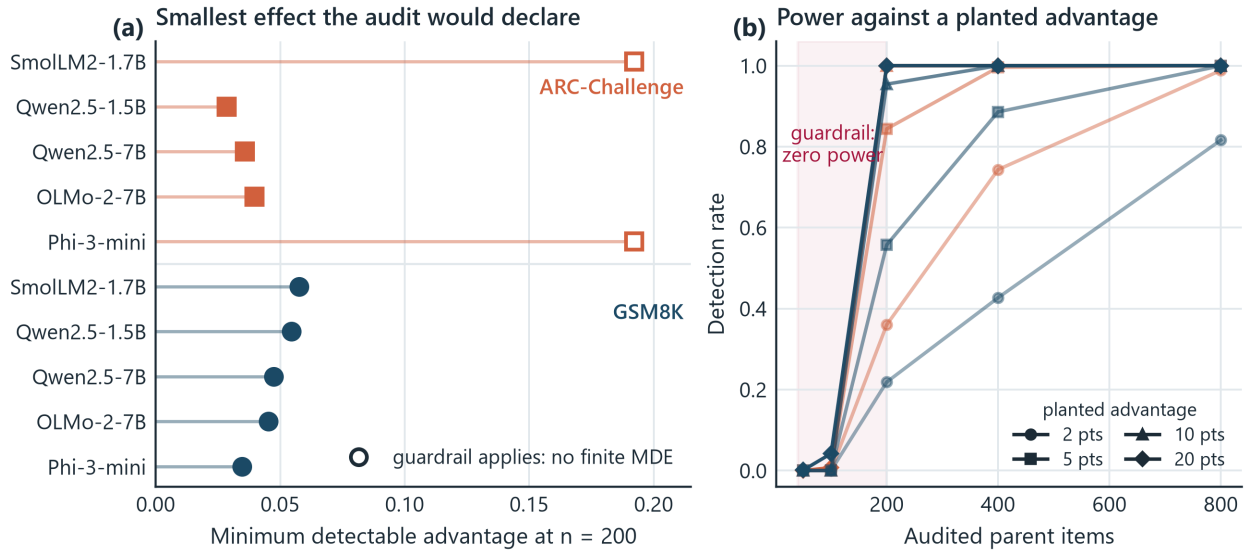


Figure 9: Left: the primary half-width, which is the smallest advantage the audit would declare, for each model and benchmark at n = 200; open markers are the two pairs where the guardrail applies. Right: detection rate against audit size for planted advantages of two, five, ten, and twenty points. The flat zero region below 200 items is the guardrail.

near-certainty on ARC; eight hundred reaches two-point resolution. Those numbers are specific to bounded correctness outcomes with the concentration these banks show, and the script that produces them is released so an auditor can recompute them for their own benchmark from a pilot sample. Figure 9 shows both quantities together: the smallest advantage each pair could declare, and the detection curves those thresholds come from.

7.10 An exact test that works where the interval cannot

Section 7.9 leaves a gap that is worth closing rather than only reporting. The interval on the mean gap has no power below about two hundred parents, and the reason is structural: a bounded-difference bound uses only the range of D_i, so the many parents answered identically on every form, which carry no information about a public-form advantage, widen the bound exactly as much as the informative ones. For detection rather than estimation the audit can do better, and the ingredient is already in the paper.

Proposition 4 says that if the pseudo-original role is assigned at random among a parent's forms, the expected gap is exactly zero. Proposition 6 takes the whole reference distribution rather than only its mean. Under the sharp null that the public form is exchangeable with its fresh forms within each parent, relabelling which form plays the original role, independently in each parent, generates the exact null distribution of the mean gap. The resulting p-value is valid at any sample size, for any joint distribution of correctness, with no appeal to asymptotics. It is the same randomization idea that already underpins the placebo check, used for inference instead of only for diagnosis. Table 12 compares the two procedures at the sizes where they differ, and Figure 10 adds the test's simulated level.

The change is confined to where it was needed. At two hundred parents and above the two agree closely, because there the interval is already doing its job. Below that the test recovers most of what the guardrail gives up: at one hundred parents a ten-point advantage goes from undetectable to 77% detection on GSM8K and 95% on ARC, and even fifty parents support a real chance of catching a large effect. An auditor with a small budget is no longer choosing between an invalid interval and a vacuous one.

Applied to the audited data, the test agrees with the intervals: it rejects for none of the ten model-benchmark pairs, with p-values from 0.16 to 1.00. It is now computed as a standard column of the analysis script, so any

Table 12: Detection rate of the exact randomization test against the interval of Section 7.9, same planted advantages, same resampled banks, 2,000 audits per cell. The test's simulated type-I error across all sizes and both benchmarks runs from 0.029 to 0.052 against a nominal 0.05.

Benchmark	Audited items	delta	Interval	Exact test
GSM8K	50	0.10	0.00	0.43
GSM8K	50	0.20	0.00	0.92
GSM8K	100	0.05	0.00	0.32
GSM8K	100	0.10	0.00	0.77
GSM8K	200	0.05	0.56	0.61
GSM8K	200	0.10	0.95	0.97
ARC	50	0.10	0.00	0.64
ARC	100	0.05	0.00	0.52
ARC	100	0.10	0.00	0.95
ARC	200	0.05	0.84	0.86

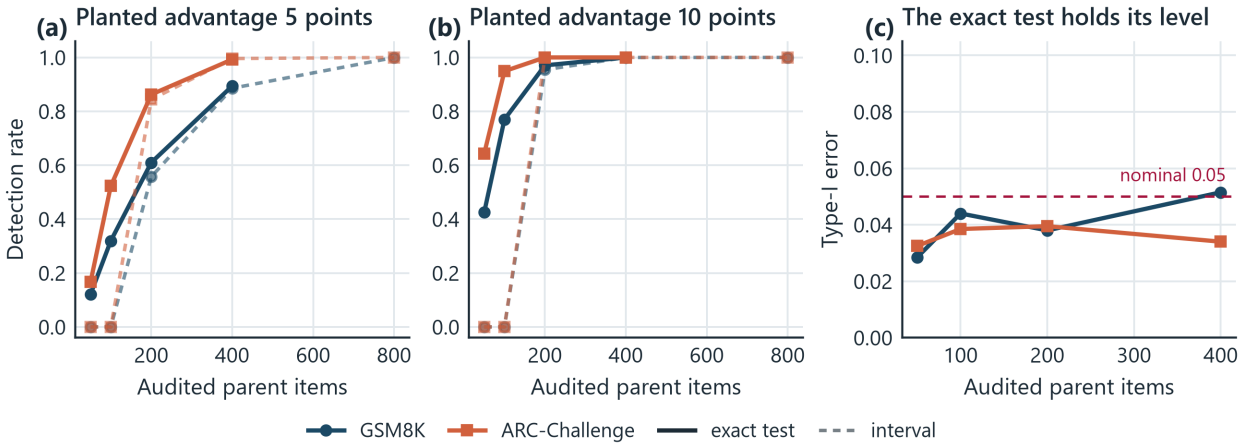


Figure 10: Left and centre: detection rate against audit size for a five-point and a ten-point planted advantage, exact test against interval. Right: simulated type-I error of the exact test, obtained by relabelling the real correctness vectors so the null holds by construction; the test sits at or below the nominal level at every size.

reuse of this code reports detection alongside the intervals without extra work. That agreement matters more than it might appear. The nulls of Table 3 could have been an artefact of an underpowered procedure, and Section 7.9 shows that worry is not idle at these sample sizes. A more powerful test applied to the same outcomes still finds nothing, which makes the nulls harder to explain away as insensitivity of the instrument.

What the test is exact for needs stating carefully, because the sharp null is a joint one. It says a parent's forms are exchangeable for this model, which is true only when there is neither contamination nor ordinary form mismatch. The audit measures the second of those separately, on the control banks, and Table 3 shows it is small but not zero: the form gap runs from -0.088 to +0.033 across the ten pairs. So the null the test is exact for is not quite true even for a model that has never seen the benchmark, and the type-I simulation in Figure 10 cannot see this, because it relabels the real correctness vectors and thereby imposes exchangeability by construction.

Table 13 measures the cost. It simulates audits of two hundred parents with no contamination whatever and a fresh-form penalty of a given size, and reports how often the test rejects at the nominal level.

The first row is the exactness the paper claims and finds. The last row is the same test, still with nothing to detect, rejecting on more than two thirds of audits. **A rejection from this test therefore cannot be read as evidence of a public-form advantage**, and we withdraw that reading, which an earlier version of this section recommended. What the test can do is one-directional. A non-rejection is a statement about the

Table 13: What ordinary form mismatch does to the exact test. No contamination is present in any row; every rejection is caused by the fresh forms being slightly harder than the public one. The largest form gap is the largest measured on any control bank in this audit.

Form gap in the data	**Rejection rate at the 0.05 level**
0.000 (exchangeable)	0.039
0.020	0.107
0.050	0.315
0.088 (largest measured)	0.689

Table 14: Where the contamination effect went. Accuracy before and after exposure, on the hundred leaked parents and the hundred untouched ones. The measurable gap is the difference between the public and fresh columns, which is what any paraphrase audit observes.

Parents	**Form**	**Before**	**After**	**Change**
Exposed (n = 100)	public, trained on	0.460	0.610	**+0.150**
Exposed (n = 100)	fresh, never trained on	0.485	0.570	**+0.085**
Exposed (n = 100)	*measurable gap*	-0.025	+0.040	**+0.065**
Unexposed (n = 100)	public	0.490	0.400	-0.090
Unexposed (n = 100)	fresh	0.515	0.425	-0.090
Unexposed (n = 100)	*measurable gap*	-0.025	-0.025	**+0.000**

joint null, so it bounds contamination and form mismatch together, and it does so exactly and at any sample size. All ten audited pairs are non-rejections. The nulls of Table 3 therefore survive a procedure with far more power than the interval, which is the use we make of the test and the only use it supports.

The recommendation that follows is narrower than the one we started with. An audit should use the test to strengthen a null and never to announce a positive; the intervals bound how large an advantage could be; and the sensitivity analysis of Section 7.4 is what separates exposure from ordinary form mismatch, which the test by construction cannot. The registered analysis of this paper is unchanged, and the test is reported alongside it rather than in place of any part of it.

7.11 A positive control on the real bank: most contamination is invisible to a paraphrase audit

Sections 7.9 and 7.10 say what the audit could detect, computed from the variance of the observed gaps. This section asks the same question by experiment: contaminate a model with the audited items, then run the identical audit on it. The protocol was registered before training, and the exposed model is Qwen2.5-1.5B-Instruct, already in the audited panel so its clean audit is on record in Table 3.

A random hundred of the two hundred audited GSM8K parents were drawn as exposed. The model was fine-tuned on the public forms of those hundred items with their reference solutions, mixed into nine hundred unrelated arithmetic records, for three epochs. Fresh forms never entered training, and a check confirmed that no fresh-form string appears anywhere in the mixture. The remaining hundred parents were left unexposed and serve as a within-model control for the effects of fine-tuning itself. Table 14 gives the four accuracies before and after.

Two things follow. The first is a check on the design, the second is the result, and both use the unexposed hundred as the control group they were created to be.

The paired design absorbs what it should. Fine-tuning damaged the model's general arithmetic: on the untouched parents, accuracy fell by nine points on the public form and by nine points on the fresh forms, and the measurable gap moved by 0.000. The other never-contaminated sets behave the same way. The ARC audit, on a benchmark the contamination never touched, moved +0.023; the private GSM8K control bank moved +0.067; the ARC control bank moved +0.013. None differs significantly from zero, with p between 0.27 and 1.00. The intervals are wide and the exact 0.000 on the unexposed half is partly luck, so the claim we

Table 15: Difference-in-differences estimates for the effect of leaking an item's public wording, with bootstrap intervals over parents.

Quantity	Exposed	Unexposed	Difference in differences	95% CI
Change in public-form accuracy	+0.150	-0.090	**+0.240**	[+0.080, +0.400]
Change in fresh-form accuracy	+0.085	-0.090	**+0.175**	[+0.035, +0.315]
Change in the measurable gap	+0.065	+0.000	**+0.065**	—

Table 16: Four contaminated models from four families. Difference-in-differences against the unexposed control parents of the same model, with bootstrap intervals. The visible gap is what a paraphrase audit observes.

Model	Family	DiD public	DiD fresh	Visible gap
Qwen2.5-1.5B	Qwen	+0.240 [+0.080, +0.400]	+0.175 [+0.035, +0.315]	+0.065
SmolLM2-1.7B	SmolLM2	+0.100 [-0.050, +0.250]	+0.040 [-0.060, +0.140]	+0.060
OLMo-2-1B	OLMo	+0.500 [+0.340, +0.660]	+0.350 [+0.230, +0.470]	+0.150
StableLM-2-1.6B	StableLM	+0.200 [+0.050, +0.350]	+0.230 [+0.115, +0.340]	-0.030
Pooled		**+0.260 [+0.183, +0.335]**	**+0.199 [+0.140, +0.258]**	+0.061

make is the one the data supports: nine points of non-specific damage to the model produced no detectable movement in the paired difference anywhere it was not contaminated.

Because the fine-tuning moved everything downward, the raw before-and-after numbers on the exposed half understate the contamination. The unexposed half tells us what would have happened to the exposed half without a leak, so the effect of contamination is a difference in differences. Table 15 gives it.

Leaking an item raised accuracy on its memorised wording by twenty-four points relative to control, and it raised accuracy on paraphrases of that item, which the model never saw, by seventeen and a half points. Both intervals exclude zero; the transferred component is significant, which it was not before the general fine-tuning damage was controlled for. Only the residual six and a half points appears as a public-form advantage, which is what any audit comparing an original against a rewrite can observe.

The transferred share is therefore estimated at 73%. That ratio remains imprecise, with a bootstrap interval of [19%, 168%], because it divides one noisy quantity by another; what is now firm is that its lower bound sits well above zero, that 96% of resamples put the share above a quarter and 79% above a half. We report the two components separately for that reason: they are individually significant and interpretable, and the ratio is a summary rather than the measurement.

We also record an analysis we tried and rejected, because it is an easy mistake for anyone repeating this work. Conditioning on the items the model newly answered correctly on the leaked wording gives an apparently sharp estimate: for those twenty-five items, fresh-form accuracy rose by +0.400 with a 95% interval of [+0.242, +0.558]. Running the identical calculation on the unexposed half, where no leak occurred, gives +0.250 on twelve items. Most of the apparent effect is selection: conditioning on an item that improved on one form selects items that improved on correlated forms. The difference between the two, +0.150, is not significant at $p = 0.28$. The difference-in-differences estimator above avoids this because it conditions on assignment rather than on outcome.

Three more models, each registered before it was run. Study 1 rests on one model, so we repeated it on SmolLM2-1.7B-Instruct, and then, in a separately registered Study 3, on OLMo-2-1B-Instruct and StableLM-2-1.6B-Chat. The two Study 3 models are not in the audited panel, so each run audits the clean model first, trains, and audits the exposed model, all in one session. Every study uses the same exposed parents, the same mixture, the same seed, the same three epochs and the same analysis. Study 3's registration fixed the falsifier in advance: two or more models with a fresh-form effect at or below zero alongside a clearly positive public-form effect would mean the transfer is not general. Table 16 gives all four studies with the pooled estimate.

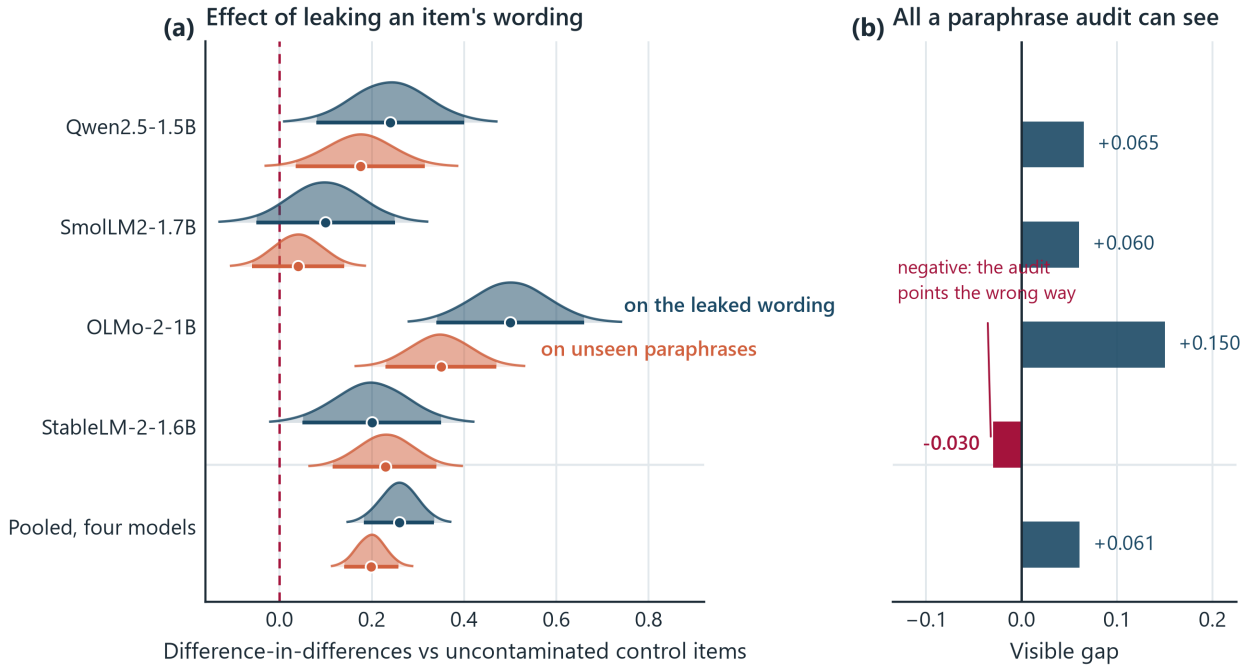


Figure 11: Left: for each model, the effect of leaking an item on the wording that was trained on and on paraphrases the model never saw, as differences in differences against uncontaminated control items, with bootstrap intervals. Right: the visible gap, the difference between the two, which is all a paraphrase audit observes. For StableLM-2-1.6B it is negative.

All four models show a positive transfer to paraphrases they never saw, and three of the four exclude zero on their own; SmolLM2, which absorbed the leak least, does not. Pooled across the four, the effect on the memorised wording is +0.260 and the effect on unseen paraphrases is +0.199, with every bootstrap resample positive. The transferred share has a pooled interval of [52%, 110%], and every resample puts it above a quarter. The registered falsifier did not fire: no model has a fresh-form effect at or below zero.

The transferred share is now bounded well away from zero, which is the substantive gain from four studies over one. The majority of what a model gains from seeing an item does not stay attached to the wording it saw. It is not bounded away from unity: the interval runs to 110%, so the data are consistent with the entire effect transferring and none of it being visible to a paraphrase audit. That is the direction that matters for an auditor, and it is why we report the two components rather than lean on the ratio.

Figure 11 draws the same four studies, and two features of them need saying plainly rather than glossing.

The magnitude of the leak's effect varies a great deal across models, from ten points to fifty on the memorised wording, and the models that were damaged most by the fine-tuning itself show the largest difference-in-differences. OLMo-2-1B lost thirty-four points on the untouched control parents, so a large part of its +0.500 is training protecting the items it saw from a degradation that hit everything else. That is still item-specific knowledge behaving in an item-specific way, and its transfer to paraphrases is measured the same way, but it is a different regime from Qwen's nine-point drift and the two should not be read as measurements of the same underlying quantity.

The visible gap is not stable across models. Studies 1 and 2 gave +0.065 and +0.060 and we noted the coincidence while declining to call it a pattern; with two more models it reads +0.150 and -0.030, and there is no pattern to call. The StableLM result is the sharper one, and it is the clearest single illustration of this section's point: leaking a hundred items raised that model's accuracy on them by twenty points relative to control, and a paraphrase audit of it would observe a gap of -0.030, pointing the wrong way. On that model

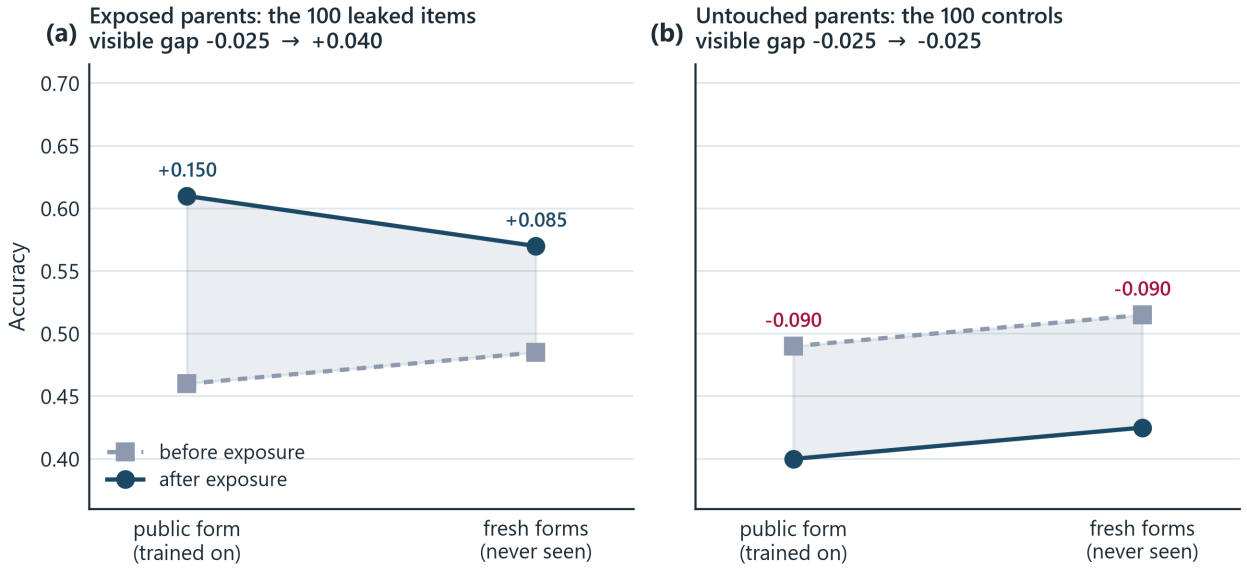


Figure 12: The positive control. Left: accuracy before and after exposure on the leaked parents, for the trained public wording and for paraphrases never shown to the model; the shaded band is the portion a paraphrase audit can see. Right: the same for the untouched parents, where fine-tuning moved both forms by the same amount and the measurable gap did not move.

the audit would report no evidence of contamination at all, correctly by its own definition, while a fifth of the leaked items had been learned.

That mechanism explains the shape of this paper's other findings. A model can gain twenty-four points on the items it memorised and still present a six-point public-form gap, so the nulls of Table 3 bound surface-form inflation and say considerably less about total contamination than the phrase "contamination audit" suggests. It is also consistent with the benchmark-rebuilding literature, where writing entirely new problems finds effects that paraphrase studies do not.

The registered claim rule did not fire, and we report that as the outcome it is. Averaged over the whole bank, where only half the parents were contaminated, the audit's estimate moved from -0.025 to +0.008, a shift of +0.033 against a minimum detectable effect of 0.057 for this model at two hundred parents; the exact test gives $p = 0.45$. Restricted to the exposed hundred, the gap of +0.040 is not significant either, $p = 0.23$. Both non-detections are what Section 7.9 predicts, and that is the narrow thing this experiment establishes about the instrument: it behaved as its measured sensitivity says it should. Figure 12 shows where the effect went in Study 1, the exposed hundred beside the untouched hundred.

7.12 A second positive control that failed, and what a failure is worth

The transfer result of Section 7.11 rests on one benchmark. GSM8K is scored by letting the model write an answer and checking the number it arrives at. ARC-Challenge is scored a different way: the model writes nothing, and each of the four options is ranked by how likely its text is as a continuation of the question. There is no reason those two scoring paths have to behave alike, so we registered a second positive control to find out, and locked the protocol before anything was trained.

The design needed one extra piece. The released ARC fresh forms rewrite the question stem and leave the four options word for word, so a model that has memorised nothing but an option list would score the same on the public form and on both fresh forms. To break that, we wrote a third form for each exposed item, a deep paraphrase that rewords the stem and the options together while keeping what each option refers to and where the correct one sits. Twelve items carry options that cannot be reworded at all, such as balanced

Table 17: The failed manipulation. Accuracy on the ARC items that were leaked, beside the ones that were not, before and after fine-tuning. The difference in differences on the trained wording is the manipulation check, and the registration required it to be at least +0.05.

Parents	Form	Before	After	Difference in differences	95% CI
Exposed (n = 100)	public, trained on	0.460	0.280	**-0.220**	[-0.350, -0.090]
Exposed (n = 100)	fresh, never trained on	0.420	0.375	-0.090	[-0.205, +0.025]
Exposed (n = 93)	deep paraphrase	0.387	0.301	-0.131	[-0.231, -0.032]
Unexposed (n = 100)	public	0.400	0.440	—	—
Unexposed (n = 100)	fresh	0.400	0.445	—	—

equations, bare quantities, and proper nouns, and those were declared out of the deep-paraphrase arm before any item was drafted.

The study failed. Fine-tuning on the leaked items did not raise accuracy on the leaked wording; it lowered it, by eighteen points, while the untouched items in the same model went up by four. Table 17 gives the numbers.

By the rule fixed in advance, that ends the study. A manipulation check that comes in at -0.220 against a floor of +0.05 means no contamination was successfully planted, and a study that planted no contamination cannot say whether contamination transfers. We report it as a failed manipulation and draw nothing from it about ARC in either direction. In particular it is not evidence that contamination fails to transfer on a ranking benchmark; that question is still open, and Section 8 leaves it open. The transfer shares the registration asked for are not reported, because a share of a negative effect is not a share of anything, and the analysis code withholds them rather than printing a ratio that would read like a healthy result to anyone who skipped the check above it.

What makes the failure worth a section is that its cause is not a mystery. The exposure record was the question, the four options, and the line `Answer:  A.`, which is what a leaked multiple-choice item looks like when it is written down. The scorer never shows the model an option list at all: it presents the question, appends `Answer:`, and measures how likely each option's text is as what comes next. So fine-tuning τght the model to put a single letter after `Answer:`, and the scorer asked it for a sentence. What the training actually installed was a pull toward whichever option had been listed first, which is the right answer only a quarter of the time.

That pull is visible and it is specific. On the items it was trained on, the model picked the first-listed option 68% of the time, against a true rate of 25%. On the same items with the stem rewritten and the options left alone it was 62% and 60%. On the deep paraphrases, where the memorised option strings are gone, it fell to 33%, close to chance. On the hundred items the model never saw, it stayed at 19% throughout. Figure 13 shows the whole pattern, and the ordering across those four columns is the argument: the pull tracks how much of the memorised option text survives the rewrite, which is what a memorised string looks like and not what knowing an answer looks like.

One deviation from the registration belongs here rather than in a footnote. The protocol made a human reading of every deep-paraphrase item a precondition for running at all, on the ground that a script can check that an option was reworded but not that it still means what it meant. The run went ahead before that reading was finished. It does not touch the outcome, because the manipulation check is computed from the public forms, which are the original ARC items and not anything written for this study, and because the deep-paraphrase accuracies contribute nothing once that check has failed. The mechanical checks all ran beforehand. We record it because a registration that is only honoured when it is convenient is not a registration.

Two things follow, and the second is the more useful of them.

The narrow one is about this benchmark. Section 8 already notes that the ARC arm of our bank cannot see memorisation attached to the options, because its fresh forms hold the options fixed by construction. The deep-paraphrase bank was built to look into that blind spot, and although the study around it failed, the

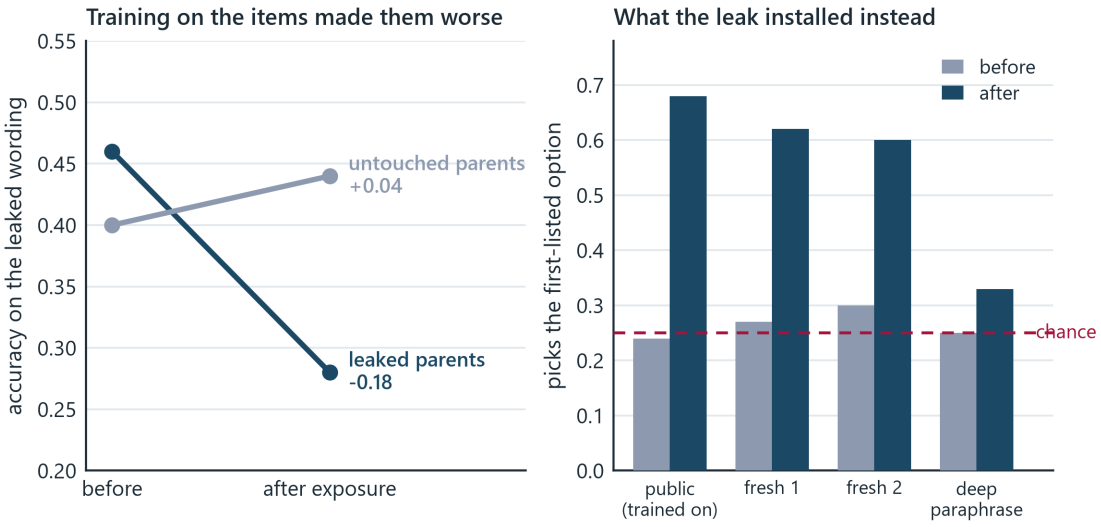


Figure 13: The failed manipulation and its cause. Left: training on an item lowered accuracy on that item while untouched items rose, which is the opposite of contamination. Right: how often the scorer's preferred option is the one listed first, on the exposed items only. Before fine-tuning the model sits at chance on every form. After it, the pull toward the first option is strongest on the trained wording, nearly as strong on the released fresh forms, which keep the same option strings, and almost gone on the deep paraphrases, which reword them.

bank did its job: it is the only one of the three form types on which the artefact largely disappears. We release it with the rest.

The broad one is about positive controls. A positive control is only a control if the thing it plants is a thing the measuring instrument can register. Ours planted a letter and measured a sentence, and the mismatch was not visible in the training loss, which fell from 1.01 to 0.16 exactly as it would have if everything were working. Nothing in the fitting told us the study had failed; only the manipulation check did, and only because it had been written down beforehand with a number attached. We had also declared in advance that a failed manipulation would be reported and would not be retried, so this section exists instead of a second run with a friendlier exposure format. That constraint is the point. An experiment that is allowed to be repeated until it works is not a positive control, and the version of this section that reported a success after three attempts would have been worth less than this one.

7.13 The same leak, written so the scorer can read it

Section 7.12 diagnosed a failure rather than merely reporting one, and a diagnosis that is never tested is a story. So we registered one more study, changed one thing, and ran it once.

The change is the exposure record. Study 4 leaked each item as a question, an option list, and the line `Answer: A.`; Study 5 leaks it as the question and the text of the correct option, with no list and no letter. That is the form a scraped question-and-answer pair takes, and it is the form the scorer reads. Everything else is held where Study 4 had it: the same model, the same hundred exposed parents from the same seed, the same untouched hundred, the same thousand-record mixture, the same three epochs, the same learning rate, the same optimizer. Two studies that differ in one thing, so the difference between them means something. Table 18 gives the result.

The registration says plainly what this costs. A leak written to match the scoring prompt makes a large effect on the trained wording nearly certain, so the manipulation check here is a precondition and not a finding. It also confronts the obvious objection, which is that Study 4's stopping rule said no further studies on ARC whatever the outcome. The defence we offer is narrow: the fault was in the instrument rather than in the

Table 18: The corrected leak. Accuracy on the ARC items that were leaked, beside the hundred that were not, before and after fine-tuning. The rightmost column is the difference in differences against the untouched parents.

Parents	Form	Before	After	Difference in differences	95% CI
Exposed (n = 100)	public, trained on	0.460	1.000	**+0.490**	[+0.360, +0.620]
Exposed (n = 100)	fresh, never trained on	0.420	0.980	**+0.510**	[+0.395, +0.625]
Exposed (n = 93)	deep paraphrase	0.376	0.602	+0.176	[+0.039, +0.315]
Unexposed (n = 100)	public	0.400	0.450	—	—
Unexposed (n = 100)	fresh	0.400	0.450	—	—

result, it is provable from the two scripts without running either, Study 4 stays in this paper at full length and unchanged, and every interpretation below was fixed before any number existed. A reader who does not accept that should treat Section 7.12 as the result and this section as an addendum to discount.

The diagnosis holds. Changing nothing but the shape of the leaked line moved the manipulation check from -0.220 to +0.490. Study 4 did not fail because ARC resists contamination; it failed because what it planted and what it measured were different things.

The result that matters is in the fourth column of the first two rows. Leaking an item raised accuracy on the leaked wording by 49 points and on the rewritten wording by 51 points, so **the gap an auditor would actually compute is -0.020, with an interval from -0.110 to +0.070.** Forty-nine points of deliberate, known contamination sit in this model, and the audit that Section 7 runs on ARC sees none of it. This is the paper's central claim on a second benchmark and a second scoring path, and it is the strongest single demonstration of it we have.

It is also, in this case, mechanical, and we said so before running rather than after. ARC fresh forms rewrite the stem and keep the four options word for word. A model that has memorised nothing except the string of the right answer will therefore score on a fresh form exactly as it scores on the public one, because the string it memorised is sitting in both. The near-perfect accuracies say that is close to what happened: the exposed model answered every one of the hundred trained items correctly and 98 of a hundred rewritten ones. So the -0.020 gap is a real measurement of a real blindness, and the blindness follows from how the bank is built rather than from any deep claim about what models learn. Any ARC-style audit whose rewrites hold the options fixed inherits it.

That leaves the deep-paraphrase arm, which is the one built to see past the string, and here the registration constrains us. The counterfactual for the deep paraphrases is the control group's change on fresh forms, and the registration required that assumption to be checked against the twenty control parents that carry a deep paraphrase, with a tolerance of 0.05. It came in at -0.100, so by the rule fixed in advance **the deep-paraphrase arm is descriptive only, and no confirmation of Section 7.11 is drawn from it.** Descriptively, accuracy on the reworded items rose from 0.376 to 0.602 while the untouched parents did not move, which is about a third of the effect surviving a rewrite that destroys every memorised string. We report the number and decline to build on it.

One flaw in that check is ours and worth stating. Twenty parents cannot resolve a 0.05 tolerance: the observed -0.100 carries a bootstrap interval from -0.350 to +0.150, so the check cannot distinguish a real violation from noise, and it would have fired about as readily on a bank where the assumption held perfectly. The rule was fixed in advance and we follow it, but a registration that writes an underpowered gate has arranged for its own most interesting arm to be discarded on a coin flip, and the next version of this experiment should size that check rather than assume it.

Figure 14 puts the two studies beside each other and shows where the corrected leak's advantage landed. Two further cautions belong beside the table. Accuracy on the trained items reaches exactly 1.000, so both the public and fresh differences in differences are censored from above; their signs are safe and their sizes are not, and the transfer shares built from them are unreliable enough that we do not report them. And the leak

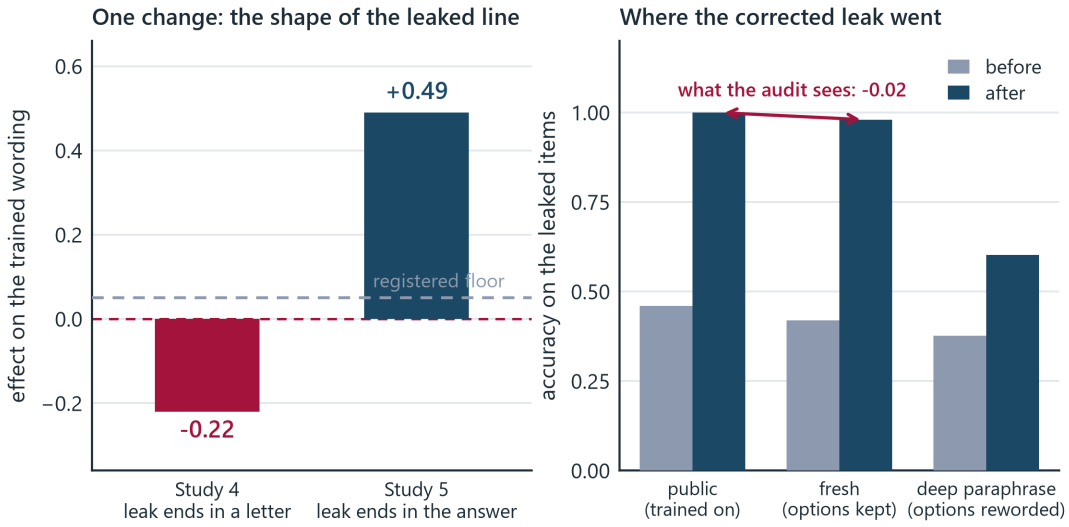


Figure 14: Study 4 and Study 5 side by side. Left: the difference in differences on the trained wording under the two exposure formats, from the same model, the same items and the same seed. The only difference is whether the leaked line ended in a letter or in the correct option's text. Right: accuracy on the exposed parents before and after the corrected leak, for the trained wording, the released fresh forms that keep the options verbatim, and the deep paraphrases that reword them. The gap between the first two is what an audit sees.

here mirrors the evaluation prompt closely, which is favourable to detection by construction and is the reason the manipulation check is not offered as evidence of anything.

What this settles and what it does not. It settles that a contaminated model can carry a very large, entirely real advantage on ARC while the paraphrase audit of Section 7 reports nothing, which is the reading we attach to the ARC nulls from here on. It does not settle whether contamination on ARC transfers to genuinely new wording, because the only arm that could speak to that was demoted by its own pre-declared check. Section 8 records both.

7.14 Nothing left to assume: contamination reaches wording the model never saw

Section 7.13 left one question open and named the reason. The deep-paraphrase arm existed for all 100 exposed parents but for only 25 of the 100 control parents, so the control group had no deep-paraphrase change of its own and the analysis borrowed its fresh-form change instead. That borrowing needed an assumption, the assumption needed a check, and the check ran on twenty items, which cannot resolve the tolerance it was given. The arm was demoted and the question stayed open.

The fix is not a larger check. It is to write the missing items, so that there is nothing to assume. Deep paraphrases were drafted for the 75 control parents that lacked one, under the six rules now in force including rule 6, and the deep-paraphrase estimate becomes what the other two already were: the exposed group's change minus the control group's change on the same form type. Every audited ARC parent now carries all three forms.

Everything else is held at Study 5's values, so the public and fresh arms are a rerun. They reproduced to four decimal places, which confirms the pipeline is deterministic and is not evidence about sampling variability; an identical seed on identical data should give an identical answer, and it did. Table 19 gives all three arms.

The answer is in the third row. Leaking an item raised accuracy by 24 points on a rewrite that changes the question stem and every one of the four options, against control parents that moved by -1 point on the same kind of rewrite. The interval excludes zero, and nothing in the estimate rests on an assumption: it is one

Table 19: The complete deep-paraphrase bank. Accuracy on the leaked parents beside the untouched ones, for all three form types, before and after the same fine-tuning as Study 5. The deep-paraphrase row is now a direct comparison rather than a substituted one.

Parents	Form	Before	After	Difference in differences	95% CI
Exposed (n = 100)	public, trained on	0.460	1.000	**+0.490**	[+0.360, +0.620]
Exposed (n = 100)	fresh, options kept	0.420	0.980	**+0.510**	[+0.395, +0.625]
Exposed (n = 93)	deep paraphrase	0.376	0.602	**+0.238**	[+0.089, +0.387]
Unexposed (n = 100)	public	0.400	0.450	—	—
Unexposed (n = 100)	fresh	0.400	0.450	—	—
Unexposed (n = 83)	deep paraphrase	0.410	0.398	—	—

group's change minus another group's change on the same form type, the same quantity the first two rows report.

So contamination on ARC does reach wording the model never saw. No option string in a deep paraphrase appears anywhere in training, and the exposed model still gains 24 points on those items. That is not a memorised string being recognised; it is something about the item that survived being rewritten. The transfer finding of Section 7.11, measured on GSM8K under generation and exact match, holds on a second benchmark under a scoring path that ranks fixed option strings, and the generalisation the abstract makes is supported on both.

Three things sit alongside that and none of them is comfortable.

The first vindicates Section 7.13's demotion. Study 5's assumption check fired on twenty parents and we criticised it for being unable to resolve its own tolerance. It was underpowered and it was also correct: measured now on 83 control parents, fine-tuning moved the fresh forms up by 5 points and the deep paraphrases down by 1, an asymmetry of about -0.062 in the same direction as the -0.100 the small check reported. The substitution really was invalid, and it was biasing the estimate downward, since the borrowed counterfactual gave +0.176 where the direct comparison gives +0.238. A gate that cannot see clearly still stopped us from publishing a number that was wrong by a quarter of its own size.

The second is the ceiling, and it is why we report the difference in differences rather than a share. Accuracy on the trained items is exactly 1.000, so the public arm is censored: the effect there is at least 49 points and the data cannot say how much more. The ratio of the deep-paraphrase effect to the public effect comes out at 0.49, but a censored denominator makes that an overestimate of the fraction, and we give it no weight. What the data support is the sign and the size of the deep-paraphrase effect itself.

The third is that the audit still sees none of this. The gap between the public form and the released fresh forms is -0.020, exactly as in Study 5, while 49 points of contamination sit in the model and roughly half of the effect survives a total rewrite. On ARC the released bank holds the four options fixed, so the memorised answer string is present in the rewrite as well as the original and the difference cancels. The blindness is a property of how the bank is built, and Figure 15 puts the three arms side by side so that the distance between what is there and what is visible can be read directly.

One number here moved after the fact and the change is stated rather than absorbed. Section 7.16 reruns this experiment under three further training seeds, and the seed used here produced the largest deep-paraphrase estimate of the four. The +0.238 above is the registered single-seed result and the top of a range running down to +0.179, with a four-seed mean of +0.204. The sign, the ordering of the three arms and the conclusion are unchanged on every seed; the magnitude is about six points lower than this section alone would suggest.

Read together with Section 7.11, the position is this. On GSM8K, most of a contamination effect survives paraphrase and is therefore invisible to a paraphrase audit. On ARC, all of it survives the released paraphrase, because that paraphrase preserves the option strings, and about half survives a rewrite that preserves nothing. Two benchmarks, two scoring paths, two different mechanisms, and the same consequence for the auditor: a null on the visible gap bounds surface-form inflation and bounds contamination much more weakly than the word audit suggests.

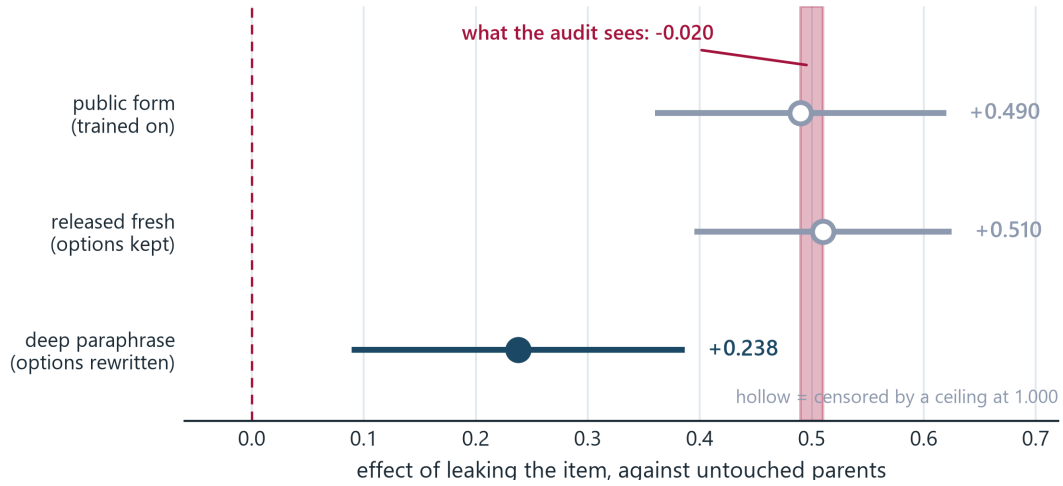


Figure 15: The three arms of Study 6 with bootstrap intervals over parents. The public and fresh-form effects are censored by a ceiling at 1.000 and are shown hollow for that reason. The deep-paraphrase effect is not censored, rests on a direct comparison against control parents that carry the same form type, and excludes zero. The bracket marks the quantity a paraphrase audit of this benchmark actually reports.

Table 20: The same statistic with the leak removed. No audited item's wording is in the training mixture, so every entry in the last column is an empirical null for the whole procedure, and the planted effects from Study 6 are shown beside them for scale.

Form family	Accuracy on leaked-half parents	Difference in differences, no leak	95% CI	Study 6, with the leak
Public form	0.460 to 0.460	-0.010	[-0.120, +0.100]	**+0.490**
Released fresh	0.420 to 0.435	+0.020	[-0.060, +0.100]	**+0.510**
Deep paraphrase	0.376 to 0.344	-0.008	[-0.147, +0.128]	**+0.238**

7.15 A negative control on the procedure itself

Every interval on a positive-control effect in this paper is a bootstrap over parents, which captures which items were drawn and nothing else. It does not see optimisation variability, it does not see whatever else a fine-tuning run does to a model besides implanting the leak, and it does not see interference between the parents that were trained on and the parents that were not. A reader is entitled to ask how much of a reported effect is the leak and how much is the fine-tuning.

The direct answer is a second training seed, and we do not have one. Every registration in this paper rules out additional seeds by name, and Section 8 says plainly that this is a cost of how we wrote them rather than a defence of not running one. What can be measured without a second seed is the other half of the question: whether the procedure produces an effect when there is nothing to find.

So we ran it with nothing to find. The mixture is the same size, the schedule, the optimiser and the seed are Study 6's, and all 1,000 training records are background items: no audited parent's wording appears anywhere, and the script asserts it before training starts. The clean audit is not rerun, because it is deterministic and Study 6 already produced it. Everything downstream is identical, so the difference in differences is computed exactly as before on a model that has been fine-tuned and has learned nothing about the benchmark. Table 20 reports it beside the planted effects.

The largest of the three is 0.020, well inside the 0.05 the registration fixed as the threshold for calling the procedure clean, and all three intervals contain zero. Fine-tuning on a thousand background science questions for three epochs moves this statistic by about two points in the worst case, against planted effects

Table 21: All four training seeds, including Study 6's. Each row is a complete rerun of Section 7.14 differing only in the trainer seed and the order of the training records.

Training seed	Public form	Released fresh	Deep paraphrase
20260824 (the registered run, Section 7.14)	+0.490	+0.510	**+0.238**
20260828	+0.460	+0.510	+0.179
20260829	+0.500	+0.495	+0.204
20260830	+0.480	+0.505	+0.195
Mean	**+0.483**	**+0.505**	**+0.204**
Spread	0.040	0.015	0.059

of forty-nine, fifty-one and twenty-four. The positive-control estimates are between twelve and twenty-five times the empirical null.

Two things this does and does not settle. It does settle that the procedure is not manufacturing its own results: an effect of the size Sections 7.13 and 7.14 report cannot be produced by the fine-tuning alone, because we ran the fine-tuning alone and it was not. It does not settle that a second seed with a real leak would reproduce +0.490, because the run that would answer that is the one the registrations forbid. A negative control bounds the false-positive behaviour of a procedure; it says nothing about how variable the procedure is when there is something to find.

We note one asymmetry in what this can show. The null run is a single run too, so its own value carries the same unmeasured seed variability as the positive controls do. What makes it informative anyway is the direction of the comparison: a single draw from the null landing within two points of zero is evidence that the null is near zero, whereas a single draw landing near fifty points would have been evidence that the whole design was broken. The test could have failed and did not.

7.16 Four training seeds, and where the registered run sits among them

Section 7.15 bounded what the procedure does when there is nothing to find. It could not bound how variable the procedure is when there is something to find, because that needs a second training seed and every registration in this paper forbids one. We ran three anyway, as an authors' override recorded in the registration, in the audit log and here: the rules were written to stop an experiment being repeated until it produced a convenient answer, and a replication that can only weaken a published estimate is not that. What replaces the rule is a set of commitments fixed before the runs, of which the binding one is that every seed is reported and none may be dropped for any reason.

The same hundred parents are leaked in every run and the same thousand records are trained on. Only their order and the trainer seed change, so what varies is the optimisation and nothing else; which items are leaked is what the parent bootstrap already varies, and moving both at once would have confounded them. The clean audit is not rerun, being deterministic. Table 21 gives every run.

All four manipulations pass, and the deep-paraphrase effect is positive on all four. By the rule fixed in advance, the public-form spread of 0.040 sits inside the 0.10 tolerance, so optimisation variability is small relative to the planted effect and Sections 7.13 and 7.14 stand.

The part worth stating plainly is where the registered run falls. **Study 6's seed produced the largest deep-paraphrase estimate of the four.** Its +0.238 is the top of a range running down to +0.179, and the four-seed mean is +0.204. Nothing was selected: that seed was fixed before any of this and the other three were added afterwards. But a reader comparing the single-seed figure in Section 7.14 against the mean here is entitled to know which end of the distribution the registered run landed on, and the answer is the favourable end. From here the deep-paraphrase effect is quoted as approximately 20 points with a four-seed range of 18 to 24, and Section 7.14's +0.238 is retained as the registered single-seed result rather than as the headline.

The ordering of the three arms is stable across every seed and is what the paper's argument actually rests on: the public and fresh effects are large and nearly equal, so the gap an audit computes stays near zero on all

four runs, and the deep-paraphrase effect is smaller than both but clearly positive on all four. The conclusion of Section 7.14, that contamination on ARC reaches wording the model never saw, does not depend on which seed was run. Its magnitude does, by about six points.

Two limits remain. The deep-paraphrase arm is the most seed-sensitive of the three, with a spread nearly four times the public form's, which is worth knowing for anyone sizing a replication of this experiment. And four seeds of one model at one dose on one benchmark bound optimisation variability in that setting only; nothing here speaks to variability across models, and Section 7.11's four-model pooled estimate remains single-seed per model.

8 Limitations

CleanScore cannot prove that a named model trained on a named item, and the audit in Section 7 does not prove that the audited models did not. The estimand is the surface-form advantage: fresh forms keep every number, fact, and answer and change only the wording. A model can depend on the training distribution of a benchmark, the pattern of problem types and answer magnitudes, without depending on the exact wording of any item. Studies that write entirely new problems in the style of GSM8K measure that broader dependence and do find it for some model families (Zhang, Da, et al., 2024); our null results on surface-form advantage are consistent with such distribution-level effects being present in the models we audited.

Fresh forms were drafted by a language model under the written rules in the registration and passed a deterministic check on every form. The authors' reading, the second check the protocol asks for, was completed after the results were produced rather than before, and was prioritised rather than exhaustive: it covered the ranked packet released with the paper, not all 1,500 forms. It found no rewrite that changes an answer, drops needed information, or adds a hint, and one ungrammatical ARC rewrite that was reported and left in the frozen bank. A reading that follows the results cannot have shaped them, but neither can it rule out a defect among the forms it did not reach. The drafting model belongs to a different provider and family than every audited model, but a shared style across variants is possible and is not something the mechanical check can detect. The private control banks were composed by the same process. A reader who discounts LLM-drafted items should read the audit as conditional on the released bank, which is published in full so that the forms can be judged directly.

The deep-paraphrase bank that Sections 7.14 and 7.16 rest on is in a weaker state than that, and the difference is worth setting out rather than leaving inside the sentence above. It holds 200 items. All 200 have been screened item by item against their public originals by a language model, in two passes whose findings are released with the paper. The first pass covered the 125 items drafted for Study 4 and produced two things: three items needing a decision, and one systematic fault, options rewritten into their own definitions, which produced rule 6 and thirteen repairs. The second pass covered the 75 items added for Study 6 and found no case where a rewording made a distractor correct or a key wrong.

No human has read any of the 200. The reading that Study 4's registration made a precondition has not been done. A machine screening is not a substitute for it and we do not present it as one: the same model family drafted these items, so it shares whatever blind spot would have put a fault in them, and an item that reads correctly to its drafter is exactly the item a drafter-shaped check will pass.

The three flagged items were left in the bank rather than repaired. Studies 6 and 8 had already run on it, and removing items after seeing results is what makes a bank untrustworthy; the same choice was made for the 21 near-copy GSM8K forms in Section 7.7. What can be measured is what those items are worth, and dropping all three from the deep-paraphrase arm moves the estimate by between +0.003 and +0.005 on the four training seeds, in the direction of a larger effect, leaving it positive on every one. So the conclusion does not rest on our judgement about any of them.

What this does and does not touch. The mechanical checks cover the rules a script can enforce: that every option was reworded, that the key stayed at its letter, that no quantity or named entity moved, that the four options remain distinct, and that no reworded option borrows a content word from its own stem. What they cannot cover is rule 2, that rewording an option must not make a distractor correct, and that is a semantic condition on which the deep-paraphrase accuracies depend directly. A defect there would move

the deep-paraphrase arm and nothing else, since the public and released fresh forms are untouched by it. Because the estimate is a difference in differences between exposed and control parents, a defect would have to fall unevenly between the two halves to bias it, and the drafting was blind to that split in the sense that a control parent is never trained on; but items were not drafted in a randomised order and we cannot claim the blinding was procedural. Anyone weighing Section 7.14's result should weigh it as resting on a bank that is mechanically verified, partly screened by machine, and unread by a person.

The registered 320-token budget truncated a large share of the Qwen models' GSM8K reasoning, so the public scores in Table 3 are not comparable to published accuracy figures. Amendment 1 shows the gap conclusion survives removing that budget, but the rerun covers only the two Qwen models; the other three were rerun at neither a longer budget nor the amended parser, and their truncation rates, which reach 25.3% for OLMo-2-7B, are lower than the Qwen models' but a long way from zero.

The positive control of Section 7.11 rests on four models between 1B and 1.7B parameters, one fine-tuning regime, and a hundred exposed items each. Three of the four are individually significant and the pooled estimate is strong, but the models are small, the leak was deliberate and concentrated, and the effect sizes vary fourfold across them. One model, OLMo-2-1B, lost thirty-four points on control items, so its estimate reflects protection from catastrophic forgetting as much as gain; we report it with the others and flag the difference in regime. Nothing here establishes the transferred share under whatever process contaminates models in the wild. The transferred share is a ratio of two noisy quantities, and pooling four models brings its interval to [52%, 110%]: bounded well away from zero, but still loose enough at the top that the point estimate of 77% should be read as an estimate rather than a measurement. What the experiment supports is that the share is substantial and positive; pinning down its upper end would need a positive control several times larger than a hundred exposed items. Three earlier attempts at that experiment failed to implant any contamination and are reported in the audit log; they contribute nothing to the result and are excluded from it.

Every interval on a positive-control effect in this paper is a bootstrap over parents, and that is narrower than it sounds. It captures which items were drawn and nothing else. It does not capture optimisation variability, because each model was fine-tuned once; it does not capture whatever else a particular run does to a model besides implanting the leak; and it does not capture interference between the parents that were trained on and the parents that were not. A second training seed would bound the first of those, and we did not run one. The registrations forbid it: Studies 1, 2, 3 and 6 each rule out additional seeds by name, and Study 6 declares itself the last study on ARC whatever it finds. Those rules exist so that an experiment cannot be repeated until it produces a convenient answer, and having invoked them we keep them rather than reinterpret a fourth one. The cost is real and belongs here rather than in a defence: a stopping rule written to prevent outcome shopping also prevents replication, and a registration that wants both should say in advance how many seeds it will run and commit to reporting all of them. Ours did not, and that is a lesson for the next one.

What can be bounded without a second seed is whether the procedure invents an effect when there is none, and Section 7.15 measures it.

The ARC arm has a blind spot that follows from its rewrite rule rather than from any defect in the items. ARC fresh forms rewrite the question stem and leave the four options and the answer key untouched, which holds for all 200 of the 200 ARC parents. A model that has learned nothing except which option a given option set belongs with will therefore score the same on the public form and on both fresh forms, and the gap the audit measures will be zero. On ARC, in other words, the estimand can see memorisation attached to the stem and is blind by construction to memorisation attached to the options. The GSM8K arm does not have this problem, because its fresh forms change the surface of the whole item while holding the quantities and the answer fixed, which is one more reason to read the generation-and-exact-match arm as the more informative of the two. It also means the transfer result of Section 7.11, which was measured on GSM8K, has not been tested against a scoring path where the discriminating content is held fixed. We registered a second positive control to test exactly that, and Section 7.12 reports why it did not answer the question: the leak was written in the form a leaked multiple-choice item takes, which teaches a model to produce a letter, while the benchmark is scored by ranking the option text, so nothing the training installed was something the scorer could reward. The manipulation check caught it and the registration's stopping rule ended the study there.

Section 7.13 then does make the leak and the scoring path speak the same language, and the result splits in two. A leak written in the form the scorer reads plants 49 points on ARC and the gap an audit computes is -0.020, so the blindness this paper is about is present on both benchmarks. But on ARC that blindness follows from the bank rather than from the model: the fresh forms keep the four options word for word, so a memorised answer string is sitting in the rewrite as well as the original, and the near-perfect accuracies say that is close to what happened. The one arm that could have separated memorised string from transferred knowledge, the deep paraphrase, failed the assumption check the registration attached to it and is reported descriptively only. Section 7.14 then closes that by writing the missing items, so that all 200 parents carry a deep paraphrase and the arm becomes a direct comparison with nothing assumed. Twenty-four points of the effect survive a rewrite of the stem and all four options, against control items that moved by -1 point, so contamination on ARC does reach wording the model never saw. Two limits stay attached to that. Accuracy on the trained items is exactly 1.000, so the public arm is censored and no ratio built on it is trustworthy in size. And the whole ARC result rests on one model of 1.6B parameters under one fine-tuning procedure, so it establishes that the transfer exists on this scoring path rather than how large it is in general.

The control banks are small relative to the audit sample, and they set the width of the sensitivity intervals. The transport radius remains an assumption: a radius that is too small can undercover, as the design simulations show, and the registered $\rho^* = 0.02$ is a judgment, reported alongside the full grid.

The audit is powered for large effects and not for small ones: at the registered size it finds a ten-point advantage almost always and a five-point advantage about half the time on GSM8K, so the nulls bound large inflation and leave small inflation open. The exact test of Section 7.10 improves detection but does not change this, since at two hundred parents the two procedures agree; its gain is at smaller audit sizes. It is also exact only for a joint null that covers contamination and form mismatch together, and Table 13 shows it rejecting on most uncontaminated audits once the form gap reaches the largest value measured here, so it is usable for strengthening a null and not for declaring a positive. The audit covers five open models between 1.5B and 7B parameters on two benchmarks. It says nothing about frontier or closed models, and a panel of five cannot support claims about open models in general. Controlled exposure with 0.5B and 360M models trained on 1,200 records for three epochs may not reproduce the behavior of frontier systems. The Qwen2.5-0.5B validation is direction-consistent but underpowered. Exposure spills across related items, so within-arm exposed-versus-unexposed comparisons are diagnostics only.

9 Reproducibility

Every analysis in this paper was fixed before the corresponding data existed: the controlled-exposure lock (July 2026), the real-audit registration (21 August 2026), and Amendment 1 (logged after the primary results and before any rerun). The registration, the sampling script and seed, the frozen parent lists, the item verification script, the audit log with all deviations, the item banks, the model revisions, the per-form outputs with log probabilities, the driver logs, the coverage simulation for Proposition 5, and the exact Kaggle notebooks are archived with the paper and released as a single archive with a digest for every file. The seven deterministic software checks on the interval code remain in place. Every number in the tables of this paper is checked against the archived analysis outputs by a script that is part of the repository and is rerun whenever a table changes; it caught two stale values during writing, both recorded in the audit log. The fresh forms and control banks were kept private until scoring completed. The audit log states which item checks had been completed at the time these results were produced.

10 Discussion

Benchmark contamination is partly a data-access problem and partly an identification problem. Black-box output differences can show that a public score is fragile, but they cannot by themselves say why. CleanScore is built around that boundary. It reports an observable gap, a private-control estimate, and the range of conclusions a stated transport radius allows.

Three things in the experiments stand out. First, the controlled-exposure validation found what it was built to find and nothing it was built to ignore: planted public forms raised the public-form advantage in both

families, planted fresh forms did not, and the effect spread across the bank in the way the lock predicted. Second, the real audit, applied to five open models on two real benchmarks, found no surface-form inflation under a rule fixed in advance; the tightest bound is about five points, the result survives a pre-declared rerun that removes an answer-length artifact, and the confidence-based detector's single flag on a pair with a zero accuracy gap shows why confidence and score must be reported separately. Third, the audit's precision is limited by the negative-control bank rather than by the audit sample; a variance-adaptive distribution-free bound, whose coverage we simulated to the same standard as every other interval in the paper, repairs the near-ceiling regime where the registered bound is uninformative, and wins precisely because the condition that triggers the guardrail is the condition that makes it narrower; and an audit of 100 items orders 97% of model pairs correctly while no feasible audit separates models that differ by less than its own half-width. Each of these three is a design lesson that only running the audit could have produced.

The detection-rate calibration in Section 7.9 is what makes the nulls interpretable, and it cuts both ways. A ten-point surface-form advantage would almost certainly have been seen in these audits and was not. A two-point advantage would probably have been missed. It also shows that an audit smaller than about two hundred paired items, under a guardrail chosen to protect coverage, can detect nothing at all; anyone running a small paraphrase audit and reading a wide distribution-free interval as reassurance is reading noise. That gap turned out to be closable rather than merely reportable. Conditioning on the observed correctness vectors instead of bounding their range gives a test that is exact at any size for the joint null of exchangeable forms and detects a ten-point advantage in three quarters of hundred-item audits, and applying it to the audited data leaves the nulls standing. Because that null is joint, the gain is one-directional: it makes a null harder to dismiss and cannot be turned around into a positive finding.

The null results are not a claim that the audited models are clean. They are a statement that, for these benchmarks and these models, rewriting the surface of a question does not change the answer, within stated bounds. The positive control puts a number on how much that leaves out. When we contaminated four models ourselves and could therefore see the ground truth, most of what a model gained from training on an item was still there once the item was reworded: pooled, +0.199 on unseen paraphrases against +0.260 on the wording actually trained on, so between about half and all of the effect never reaches the audit. On one of the four the visible gap was negative while a fifth of the leaked items had been learned, which is the sharpest form of the point. A paraphrase-based detector, ours included, measures only the part of contamination that stays attached to the surface, and in three of our four models, and in the pooled estimate, that was the smaller part. Whether those models depend on the benchmark's distribution is a different question, and one this estimand was not designed to answer.

11 Conclusion

CleanScore provides a black-box design for auditing public benchmark scores with matched forms, private negative controls, and explicit sensitivity bounds, and this paper tests the design rather than describing it. A registered controlled-exposure experiment in two model families shows the audit detects planted exposure to public forms and stays quiet under exposure to fresh forms. A registered audit of five open models on GSM8K and ARC-Challenge, with 1,500 verified forms and private control banks, finds no exposure-consistent public-form advantage under a claim rule fixed in advance, survives a pre-declared robustness rerun, separates a confidence shift from a score shift on real data, and shows that the control bank is the binding constraint on precision. The method's purpose is not to turn a gap into an accusation. It is to say what the data show, what depends on an assumption, and how much that assumption matters.

References

Cheng, Y., Wang, W., Moayeri, M., & Feizi, S. (2025). DyePack: Provably flagging test set contamination in large language models using backdoors. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 15356-15373. https://doi.org/10.18653/v1/2025.emnlp-main.776

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.

Dekoninck, J., Muller, M. N., & Vechev, M. (2024a). ConStat: Performance-based contamination detection in large language models. *Advances in Neural Information Processing Systems, 37*.

Dekoninck, J., Muller, M. N., Baader, M., Fischer, M., & Vechev, M. (2024b). Evading data contamination detection for language models is (too) easy. *arXiv preprint* arXiv:2402.02823. https://doi.org/10.48550/arXiv.2402.02823

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association, 58*(301), 13-30. https://doi.org/10.1080/01621459.1963.10500830

Oren, Y., Meister, N., Chatterji, N., Ladhak, F., & Hashimoto, T. B. (2024). Proving test set contamination in black box language models. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2310.17623

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology, 21*(3), 383-388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Maurer, A., & Pontil, M. (2009). Empirical Bernstein bounds and sample-variance penalization. *Proceedings of the 22nd Conference on Learning Theory*. https://doi.org/10.48550/arXiv.0907.3740

Opoku, J., & Banahene, D. (2026a). SynthGuard-ReleaseBench: Locked-audit evidence for synthetic tabular data releases. *arXiv preprint* arXiv:2608.14753. https://doi.org/10.48550/arXiv.2608.14753

Opoku, J., & Banahene, D. (2026b). Drift-aware spectral conformal prediction for non-exchangeable streaming data. *arXiv preprint* arXiv:2606.15953. https://doi.org/10.48550/arXiv.2606.15953

Opoku, J., & Banahene, D. (2026c). PromptShift-CRC: Drift-aware conformal risk control for foundation models under prompt and domain shift. *arXiv preprint* arXiv:2606.15964. https://doi.org/10.48550/arXiv.2606.15964

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics, 2*(1), 39-48. https://doi.org/10.1214/aos/1176342611

Tu, S., Zhu, K., Bai, Y., Yao, Z., Hou, L., & Li, J. (2024). DICE: Detecting in-distribution contamination in large language model fine-tuning for mathematical reasoning. arXiv:2406.04197. https://doi.org/10.48550/arXiv.2406.04197

Zhang, H., Lin, Y., & Wan, X. (2024). PaCoST: Paired confidence significance testing for benchmark contamination detection in large language models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1794-1809. https://doi.org/10.18653/v1/2024.findings-emnlp.97

Zhao, Q., Huang, Y., Lv, T., Cui, L., Sun, Q., Mao, S., Zhang, X., Xin, Y., Yin, Q., Li, S., & Wei, F. (2024). MMLU-CF: A contamination-free multi-task language understanding benchmark. arXiv:2412.15194. https://doi.org/10.48550/arXiv.2412.15194

Zhu, W., Hao, H., He, Z., Song, Y.-Z., Yueyang, J., Zhang, Z., Liu, Q., Liang, S., Li, N., Hu, Z., Gao, W., Li, R., Huang, Z., Rutkowski, L., Yu, B., & Tao, D. (2026). Controllable contamination detection for reliable LLM evaluation with statistical guarantees. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*.

Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., & Stoica, I. (2023). Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint* arXiv:2311.04850. https://doi.org/10.48550/arXiv.2311.04850

Zhang, Y., Hu, H., Wei, Y., Wang, R., & Lu, H. (2024). CLEAN-EVAL: Clean evaluation on contaminated large language models. *Findings of the Association for Computational Linguistics: NAACL 2024*, 835-847. https://doi.org/10.18653/v1/2024.findings-naacl.53

A Mathematical core

A.1 Observable target

Fix a model, prompt, decoding rule, and scoring rule. Benchmark parent item i has one public original form and K independently written fresh forms. Let Y_{iO} be correctness on the original form and let $Y_{iF1}, \ldots, Y_{iFK}$ be correctness on the fresh forms. Each outcome lies in $0, 1$. Define the parent-level public-form advantage

$$D_i = Y_{iO} - \frac{1}{K} \sum_{k=1}^{K} Y_{iFk}.$$

Thus $D_i \in [-1, 1]$. For a finite benchmark containing N parent items, the audit target is

$$\Delta_N = \frac{1}{N} \sum_{i=1}^{N} D_i.$$

This is an observable score gap. It is not, by itself, a contamination probability or proof that the model saw the benchmark during training. If decoding is stochastic, all repetitions for one parent item remain inside the same parent cluster. They must not be treated as independent benchmark items.

A.2 Conservative interval for a sampled audit

Suppose n parent items are selected uniformly without replacement from the N-item benchmark. Let

$$\widehat{\Delta} = \frac{1}{n} \sum_{i \in S} D_i.$$

Proposition 1. *For any $\alpha \in (0, 1)$,*

$$\Pr\left(|\widehat{\Delta} - \Delta_N| \leq \sqrt{\frac{2 \log(2/\alpha)}{n}} \right) \geq 1 - \alpha.$$

Proof. The sampled values lie in an interval of width two. Sampling without replacement is no less concentrated than independent sampling with replacement from the same finite population. Applying Hoeffding's inequality to the with-replacement comparison gives

$$\Pr(|\widehat{\Delta} - \Delta_N| \geq \varepsilon) \leq 2 \exp(-n\varepsilon^2/2).$$

Solving for ε proves the claim. The interval is conservative. If every benchmark item is audited, there is no item-sampling uncertainty for the finite-benchmark mean, although model-output randomness may still remain.

A.3 Stratified audit

Real audits should sample separately by domain and difficulty. Let stratum h contain N_h items, let $W_h = N_h/N$, and sample n_h items without replacement. Define

$$\widehat{\Delta}_{\text{str}} = \sum_{h=1}^{H} W_h \widehat{\Delta}_h.$$

Proposition 2. *If sampling is independent across strata, then*

$$\Pr\left(|\widehat{\Delta}_{\text{str}} - \Delta_N| \le \sqrt{2\log(2/\alpha)\sum_{h=1}^{H}\frac{W_h^2}{n_h}}\right) \ge 1-\alpha.$$

Proof. Each sampled value in stratum h enters the weighted estimator with coefficient W_h/n_h. Its weighted range has width $2W_h/n_h$. Hoeffding's inequality for the weighted sum therefore has total squared range

$$\sum_{h=1}^{H} n_h\left(\frac{2W_h}{n_h}\right)^2 = 4\sum_{h=1}^{H}\frac{W_h^2}{n_h}.$$

The same comparison applies to sampling without replacement inside each stratum. Substitution into the two-sided Hoeffding bound proves the result. This proposition gives a transparent budget rule: more items should be assigned to strata with large benchmark weight and large observed variation. A variance-adaptive allocation may be used as a secondary design, but it must be fixed using pilot data rather than the final audit outcomes.

A.4 Practical design-based precision interval

The distribution-free interval above can be very wide. The main paper therefore reports a second, standard survey-sampling interval. Let s_h^2 be the sample variance of the D_i values in stratum h, and let $f_h = n_h/N_h$. Then

$$\widehat{V} = \sum_{h=1}^{H} W_h^2(1-f_h)\frac{s_h^2}{n_h}$$

is unbiased for the design variance of the stratified sample mean under independent simple random sampling within strata. Under the usual finite-population central limit conditions,

$$\widehat{\Delta}_{\text{str}} \pm z_{1-\alpha/2}\sqrt{\widehat{V}}$$

has asymptotic $1-\alpha$ coverage. This interval is a standard precision tool, not a new theorem. CleanScore reports it beside the conservative bound and checks its coverage in simulation. The two intervals must not be described as having the same guarantee. The extended simulation shows that the approximation can under-cover when the audit is small and most parent-level gaps are zero. An independent exact calculation enumerated all possible sample compositions under four discrete finite populations. In the sparse case with population size 1,000 and sample size 50, exact Wald coverage was 0.9127; it rose to 0.9467 at sample size 200 and 0.9499 at sample size 500. This confirms that the small-sample failure is not Monte Carlo noise. CleanScore therefore uses the following registered guardrail: if the audit contains fewer than 200 parent items, or fewer than 30 sampled parent items have a nonzero gap, the distribution-free interval is the primary inferential result and the Wald interval is descriptive. These thresholds are operating rules supported by stress tests, not universal mathematical cutoffs.

A.5 Why exposure is not identified from a score gap

Let b_B be the original-versus-fresh gap that the same model would have shown on benchmark B without prior exposure. Let τ_B be the exposure-related part of the gap. We define

$$\Delta_B = \tau_B + b_B.$$

The decomposition is a definition, not an identification result. The observed audit identifies Δ_B, but it does not identify τ_B unless something is learned or assumed about the model-specific counterfactual gap b_B. Human judgments of question difficulty alone do not guarantee a bound on a model's accuracy gap.

A.6 Negative-control calibration with a transport radius

Build a private control bank C that was never public. For every control parent item, create one public-like first form and K later fresh forms using the same writing and review protocol as the audited benchmark. Because the whole bank remains private until querying is complete, its average form gap b_C can be estimated without known item exposure. To use this control for benchmark B, state a transport assumption

$$|b_B - b_C| \le \rho,$$

where $\rho \ge 0$ is reported across a sensitivity grid.

Proposition 3. *Let $I_\Delta = [L_\Delta, U_\Delta]$ cover Δ_B with probability at least $1 - \alpha_\Delta$, and let $I_C = [L_C, U_C]$ cover b_C with probability at least $1 - \alpha_C$. Under the stated transport assumption,*

$$I_\tau(\rho) = [L_\Delta - U_C - \rho,\ U_\Delta - L_C + \rho]$$

covers τ_B with probability at least $1 - \alpha_\Delta - \alpha_C$.

Proof. On the event that both confidence intervals cover, $b_B \in [L_C - \rho, U_C + \rho]$. Since $\tau_B = \Delta_B - b_B$, subtracting the widest allowed interval for b_B from the interval for Δ_B gives the stated endpoints. The union bound shows that the joint coverage event has probability at least $1 - \alpha_\Delta - \alpha_C$. The value of ρ cannot be hidden or selected because it produces a favorable conclusion. The paper must show the full sensitivity curve and identify the smallest ρ at which the practical conclusion changes.

A.7 Exact placebo-label check

The private control bank also supports a design check that requires no model assumption. Suppose a control parent has (K+1) fresh forms with fixed scored outcomes $Y_1, \dots, Y_{K+1}$. After all responses are collected, choose one form index J uniformly at random and label it the pseudo-original. Define

$$D^{\mathrm{pl}}(J) = Y_J - \frac{1}{K} \sum_{k \ne J} Y_k.$$

Proposition 4. *Conditional on the observed outcomes,*

$$\mathbb{E}_J[D^{\mathrm{pl}}(J) \mid Y_1, \dots, Y_{K+1}] = 0.$$

Proof. Let $T = \sum_{j=1}^{K+1} Y_j$. Averaging over all possible pseudo-original labels gives

$$\frac{1}{K+1} \sum_{j=1}^{K+1} \left(Y_j - \frac{T - Y_j}{K} \right) = 0.$$

This identity yields an exact randomization check for scoring, coding, and form-role asymmetry under random labels. It does not prove that a historically public original form is exchangeable with newly written forms, so it complements rather than replaces the transport sensitivity analysis.

A.8 Controlled-exposure validation target

For training seed s, let $\widehat{\Delta}_{s,c}$ be the average parent-level public-form advantage after training under condition c. The registered policy-level effects are

$$\Theta_{\mathrm{exact}} = \mathbb{E}_s[\widehat{\Delta}_{s,\mathrm{exact}} - \widehat{\Delta}_{s,\mathrm{related}}]$$

and

$$\Theta_{\text{para}} = \mathbb{E}_s[\widehat{\Delta}_{s,\text{paraphrase}} - \widehat{\Delta}_{s,\text{related}}].$$

These are effects of training-mixture policies, not item-level direct effects. An exposed training item can improve performance on related unexposed items, so exposed-versus-unexposed item summaries are diagnostic unless a credible no-interference assumption is added.

A.9 Scope of the mathematics

The finite-sample bounds concern the observable public-form advantage under the stated sampling design. The negative-control result concerns a sensitivity set under an explicit transport radius. Neither result proves that a proprietary model trained on a named benchmark. That boundary is part of the method, not a weakness to conceal.

A.10 Variance-adaptive distribution-free interval (post hoc)

The bound in Proposition 1 uses only the range of D_i. When a model is near ceiling on a benchmark, almost every parent has $D_i = 0$ and the bound is far wider than the data warrant. The empirical Bernstein inequality replaces the range with the sample variance while remaining distribution-free.

Proposition 5. *Let s^2 be the sample variance of $D_1, \ldots, D_n$ with $D_i \in [-1, 1]$, sampled uniformly without replacement from the N-item benchmark. For any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*

$$|\widehat{\Delta} - \Delta_N| \le \sqrt{\frac{2s^2 \log(4/\alpha)}{n}} + \frac{14 \log(4/\alpha)}{3(n-1)}.$$

Proof. Set $Y_i = (D_i + 1)/2 \in [0, 1]$. Theorem 4 of Maurer and Pontil (2009) is one-sided: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\mathbb{E}Y - \bar{Y} \le \sqrt{\frac{2V_n \log(2/\delta)}{n}} + \frac{7 \log(2/\delta)}{3(n-1)},$$

where V_n is the sample variance of the Y_i. Applying the same theorem to $1 - Y_i$, which also lies in $[0, 1]$ and has the same sample variance, bounds the opposite tail. Taking $\delta = \alpha/2$ in each direction and applying the union bound gives a two-sided statement at level α in which $\log(2/\delta)$ becomes $\log(4/\alpha)$. Since $D_i = 2Y_i - 1$ we have $\widehat{\Delta} - \Delta_N = 2(\bar{Y} - \mathbb{E}Y)$ and $V_n = s^2/4$. Substituting and multiplying by two gives the stated bound. As in Proposition 1, sampling without replacement from a finite population is no less concentrated than sampling with replacement from it, so the with-replacement bound applies. An earlier version of this proposition wrote the two-sided statement with $\log(2/\alpha)$, which the cited one-sided theorem does not supply. The correction widens the bound by about 9% in the variance term and 19% in the additive term. Every number reported for this interval in the paper is computed from the corrected constant.

A.11 Use

This proposition was added after the real-benchmark audit was analysed, and it is reported as a separate interval rather than substituted for the registered one. Selecting the narrower of two intervals that each hold at level $1 - \alpha$ does not itself hold at level $1 - \alpha$, so the two are always reported side by side. Its coverage was simulated to the same standard as every other interval in the paper, on the six bounded outcome shapes and the finite-population sampling design of the extended stress grid, with 4,000 repeated audits in each of 36 settings. Empirical coverage ranges from 0.9992 to 1.0000 and never falls below the nominal level. The bound is not uniformly narrower than Proposition 1: it is narrower in 25 of the 36 settings, with width ratios from 0.36 to 1.91, because the additive term $14 \log(2/\alpha)/(3(n-1))$ dominates when n is small. At $n = 200$ it is 33% narrower on a low-variance population and 45% wider on a two-point population splitting its mass

between −1 and +1. The practical consequence is specific rather than general. The registered guardrail routes to a distribution-free bound exactly when fewer than 30 of the sampled parents have a nonzero gap, that is, in the concentrated regime where Proposition 5 is narrower; the regime where it loses is one the guardrail never selects. Future audits should therefore register Proposition 5 as the distribution-free interval on the guardrail branch and retain Proposition 1 elsewhere.

A.12 Exact randomization test for a public-form advantage

Proposition 5 widens what a distribution-free interval can say, but it does not change the fact that an interval on the mean of D_i uses only the range of D_i and ignores that most parents contribute exactly zero. Section 7.9 shows the cost: below roughly two hundred audited parents the interval has no power at any plausible effect size. For detection, as opposed to estimation, the audit can do much better, and the ingredient is already in Proposition 4.

A.13 Setup

Parent i has (K+1) scored forms $Y_{i0}, \ldots, Y_{iK} \in \{0, 1\}$, where index 0 is the public form. Let J_i be the index of the form playing the original role and define

$$D_i(J_i) = Y_{iJ_i} - \frac{1}{K} \sum_{k \neq J_i} Y_{ik}, \qquad \bar{D}(J) = \frac{1}{n} \sum_{i=1}^{n} D_i(J_i).$$

The observed statistic is $\bar{D}(0)$, the audit's estimate.

A.14 Null hypothesis

H_0: within each parent, the public form is exchangeable with that parent's fresh forms. This is the sharp null that the published wording confers no advantage of any kind, whether from prior exposure or from ordinary form mismatch. It is a detection null, not an identification statement; separating exposure from form mismatch remains the job of the transport analysis.

Proposition 6. *Draw $J_1, \ldots, J_n$ independently and uniformly from $\{0, \ldots, K\}$ and let*

$$p = \Pr_J\big(|\bar{D}(J)| \geq |\bar{D}(0)| \;\big|\; Y\big)\,.$$

Under H_0, $\Pr(p \leq \alpha) \leq \alpha$ for every $\alpha \in (0, 1)$, every sample size n, and every joint distribution of the correctness vectors.

Proof. Condition on the multiset of outcomes within each parent. Under H_0 the observed labelling of forms within parent i is exchangeable, so the observed $J_i = 0$ has the same conditional law as a uniform draw from $\{0, \ldots, K\}$, independently across parents. The observed statistic is therefore an exchangeable draw from the same conditional distribution that defines p. A p-value computed as the tail probability of a statistic under the distribution its observed value is drawn from is stochastically no smaller than uniform, which is the stated bound. Finiteness of the reference set makes this exact rather than asymptotic; with Monte Carlo evaluation over M draws, using $(1 + \#\{|\bar{D}(J^{(m)})| \geq |\bar{D}(0)|\})/(1 + M)$ preserves validity. Proposition 4 is the first moment of the same reference distribution: it states that $\mathbb{E}_J[D_i(J_i)] = 0$, which is why the reference distribution is centred at zero.

A.15 What this buys and what it does not

The test conditions on the observed correctness vectors, so parents on which every form was answered identically contribute no variability and do not dilute the signal, whereas they inflate the range-based bound. Section 7.10 measures the consequence: at one hundred audited parents the interval detects nothing at any

effect size up to ten points, while the exact test detects a ten-point advantage in 77% of audits on GSM8K and 95% on ARC, at a simulated type-I error between 0.029 and 0.052. The test reports whether an advantage is present, not how large it is. The intervals of Section 3.3 and Proposition 5 remain the way to bound magnitude, and the sensitivity interval of Proposition 3 remains the way to ask how much of any advantage could be exposure. An audit should report the test for detection and the intervals for magnitude; neither replaces the other.