WHAT DOES "NO CONTAMINATION DETECTED" CERTIFY? DOSE-CALIBRATED DETECTION LIMITS FOR BENCHMARK CONTAMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model providers routinely certify their systems as uncontaminated by reporting a negative result from a statistical contamination detector. Such a report constrains nothing unless the detector's sensitivity is known, and the post-hoc likelihood-based detectors we study come with no dose-calibrated sensitivity curve that says what level of exposure a negative result can exclude. We supply one. Benchmark items are injected into continued pre-training of open models at controlled doses that vary duplication depth, leaked fraction, surface form, and model scale; we then measure the statistical power of 9 likelihood-based detectors over 36 dose cells and report, for each, the minimum detectable dose (MDD) at a calibrated 5% false-positive rate. A wide blind spot emerges. Most of the apparent memorization is not item specific: differencing against benchmark items the model never saw removes a median 87% of the raw effect. What survives is still invisible. In 8 of the 36 cells the item-specific gain is significantly positive while the best of 9 detectors reaches at most 0.39 power. The gap is not confined to likelihoods: on BIG-Bench Hard at 1.4B, averaged over 2 training seeds, four replicas inflate exact-match accuracy by $+8.6$ points (bootstrap $p = 0.031$, differenced against an uninjected holdout) while no per-item detector exceeds 0.39 power. How the audit is run matters as much as which detector is chosen: at four replicas Min-K% has 0.20 power per item, but the same scores aggregated over the benchmark reach 0.94 once a quarter of the test set is leaked. An observational arm then asks what exposure occurs without any injection. Among MMLU and Natural Questions items that appear at all in the Pile, the median n-gram count is 3, so naturally occurring overlap is concentrated in the same shallow-exposure regime where our calibration finds weak power. Replication under a second training seed preserves all 18 MDDs at 410M and the silent-cell set at 1.4B. We release CONTAM-LADDER, a suite of known-dose checkpoints and a harness, so that a negative audit can be reported together with the dose it excludes.

1 INTRODUCTION

Public leaderboards have become the currency of progress in language modeling, and benchmark contamination is their central threat. The leak need not be deliberate. Benchmarks are published on the open web, mirrored into dataset hubs, quoted in tutorials, and pasted into issue threads; web-scale crawls then sweep all of it back into the next pretraining corpus (Dodge et al., 2021; Elazar et al., 2024). The standard response is for a provider to run one of the available contamination detectors and report that nothing was found.

That practice rests on an unexamined premise. A negative result constrains the world only in proportion to the sensitivity of the instrument that produced it, and for the post-hoc likelihood-based detectors we study, published evaluations do not provide a dose-calibrated sensitivity curve stating what exposure a negative result excludes. Proactive watermarking does relate detection to controlled exposure (Sander et al., 2025), but requires acting before the leak. Existing work proposes new detectors, each validated on its own terms (Shi et al., 2024; Zhang et al., 2025; Oren et al., 2024;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Dekoninck et al., 2024b), attaches error control to a positive finding (Cheng et al., 2025; Zhang et al., 2026), or shows that a determined adversary can evade detection (Dekoninck et al., 2024a; Yao et al., 2024; Wang et al., 2026). None of these answers the question an auditor actually faces: if this detector reports nothing, what level of contamination has been ruled out?

Sensitivity and severity are distinct quantities, and nothing forces them to align. A benchmark could be absorbed strongly enough to change what the model does while staying below every detector's threshold. If such a regime exists, contamination auditing has a blind spot of unknown width, and an uncalibrated negative result leaves its own position relative to that blind spot unknown.

We measure that width by treating detection as measurement, following the logic of analytical chemistry: administer known quantities of an analyte, then record the smallest quantity the instrument reliably reports. Benchmark items are injected into continued pre-training at controlled doses, and detection power is measured as a function of dose. Dose has two dimensions that prior work conflates. *Depth* is how many times an item is duplicated; *breadth* is what fraction of the test set is leaked. For every detector we report the smallest depth at which power reaches 80% against a threshold calibrated to a 5% false-positive rate.

A calibration is informative only over the range it covers, so the controlled experiment is paired with an observational one. For benchmarks that predate a public pretraining corpus, the n-gram frequency of their items in that corpus records contamination that genuinely happened, with no injection on our part. The observational arm places natural exposure on a qualitatively comparable shallow-to-deep axis, which lets us ask whether naturally occurring overlap falls in the regime where controlled calibration finds weak sensitivity, without claiming the two exposure counts are the same quantity.

Contributions.

- We calibrate 9 likelihood-based detectors over 36 dose cells spanning two benchmarks, three model scales, two surface forms and 4 duplication depths, and report an MDD and an assurance bound for each detector in each setting.
- We separate item-specific memorization from answer-format learning, which a paired clean twin alone does not distinguish, and find the latter accounts for a median 87% of the apparent log-probability gain.
- We identify a silent zone covering 8 cells in which item-specific memorization is statistically evident while no detector exceeds 0.39 power.
- Using n-gram counts rather than injection, we show that natural exposure in a public pretraining corpus concentrates in the same shallow regime where those calibration curves are flat.
- We separate per-item from set-level auditing and oracle from realistic calibration, and quantify what each choice is worth: at four replicas the difference between the two auditing regimes is a factor of 4.6 in power.
- We release CONTAM-LADDER, a known-dose checkpoint suite and evaluation harness, so that detector papers can report calibrated sensitivity.

2 Related Work

Contamination detection. Detectors divide by the unit they score. Instance-level methods assign a suspicion score to a single example from its likelihood under the model: Min-K% (Shi et al., 2024) averages the least likely tokens, Min-K%++ normalizes each token by the model's own predictive moments (Zhang et al., 2025), and the zlib-ratio, reference-model and windowed perplexity baselines come from the memorization-auditing literature (Carlini et al., 2021; 2023). Set-level methods aggregate over a dataset: the exchangeability test of Oren et al. (2024), comparison against reference benchmarks (Dekoninck et al., 2024b), prompting probes that ask the model to complete withheld fields (Golchin & Surdeanu, 2024; Deng et al., 2024), and context interventions that read likelihoods under an altered prompt (Zawalski et al., 2026). We add no detector to this list. We measure the ones already on it.

Contamination effects. A parallel literature quantifies how contamination inflates reported scores (Magar & Schwartz, 2022; Palavalli et al., 2024; Roberts et al., 2023), and recent work shows that even a single replica of a test set measurably lowers loss (Schaeffer et al., 2026). That literature establishes a dose-to-effect relation. We establish the dose-to-detectability relation, and the gap between the two curves is what this paper is about.

Guarantees on the other error. Two recent lines attach formal error control to contamination work. DyePack plants backdoor samples before release, so that a model's response to them yields a provable bound on the false-positive rate of the resulting accusation (Cheng et al., 2025). FTD combines several detectors under an adaptive weighting and filters an evaluation set so that the false discovery rate among retained items stays below a chosen level (Zhang et al., 2026). Both control the error of a *positive* decision, the chance of wrongly accusing a clean model or wrongly keeping a contaminated item. What we calibrate is the complementary error, the error of a *negative* decision, which is a function of dose and which no FDR or FPR guarantee constrains. A procedure can hold its false discovery rate wherever the user likes and still miss every leak below sixteen replicas.

Evasion. Adversarial rephrasing (Dekoninck et al., 2024a), cross-lingual translation (Yao et al., 2024) and reinforcement-learning post-training (Wang et al., 2026) can all conceal contamination from detectors. Those are worst-case results about a motivated adversary. Our subject is the average case: ordinary, unintentional leakage, and the dose at which it becomes invisible with no evasion effort at all.

Membership inference. Instance-level detectors are membership inference attacks in disguise (Carlini et al., 2022), and membership inference on large models is known to perform near chance (Duan et al., 2024), partly because its evaluation sets suffer distribution shift between members and non-members (Das et al., 2025). Controlled injection removes that confound by construction, since our members and non-members are random splits of the same benchmark. Zarzecki et al. (2026) find the same failure mode across dataset-level auditing paradigms at realistic scales; our randomized design avoids it by construction, at the price of a synthetic exposure process. Aggregating weak per-example signals into a dataset-level decision, as dataset inference does (Maini et al., 2024), is the set-level counterpart we also evaluate.

Benchmark hygiene. Complementary defences act before release rather than after: withholding test data from plain text (Jacovi et al., 2023), embedding detectable backdoors (Cheng et al., 2025), and refreshing items continuously (White et al., 2025). Corpus-side audits document what web-scale corpora contain (Dodge et al., 2021; Elazar et al., 2024) and make n-gram frequency queryable at scale (Liu et al., 2024); we use the latter to anchor our dose axis to naturally occurring exposure.

Contemporaneous work. Shihab et al. (2026) independently argue that a non-rejection is uninterpretable without power calibration, and derive an information limit for a benchmark in which an unknown fraction of items was seen. Their axis is breadth, ours is depth; they contribute the theory we lack. We contribute released known-dose checkpoints, the scale and surface-form axes, and a measured separation between the dose-response curve of benchmark distortion and the dose-response curve of detectability.

3 METHOD

3.1 DOSE DESIGN

Each benchmark is partitioned into equal-sized slices by a deterministic hash, and slices are assigned to dose cells; one slice is always held out uncontaminated. A cell is defined by depth $r \in \{1, 4, 16, 64\}$, the number of independent document copies of each item, and by surface form, verbatim or paraphrased. Slice-to-cell assignment rotates across runs so that slice difficulty cannot be confounded with dose. Breadth is synthesized at analysis time by mixing a fraction of treated items with held-out items, which costs no additional training.

Because treated and untreated items are random splits of one benchmark, they are exchangeable by construction. Detector evaluations assembled post hoc from different sources have been shown to be confounded by exactly this kind of distribution shift (Das et al., 2025); here it cannot arise.

3.2 Injection

Each replica is rendered as an independent document using a rotating set of templates, so that replicas survive deduplication, and is inserted at a random offset in the token stream of a replay corpus. Since memorization grows with the number of duplicate copies (Carlini et al., 2023), depth is our primary dose axis. The token budget is fixed, so injected documents displace replay text rather than adding to it; they account for 1.94% of the 400M budget.

3.3 Paired clean twins

Every contaminated run is matched by a twin trained for the same number of optimizer steps, from the same seed, on the same replay stream, with no injection. Because injection displaces rather than extends, the twin sees 1.94% more of the replay corpus than its contaminated partner: the pair is matched on budget, not on identical text. That displaced fraction is small and is identical at every scale, which is what keeps the cross-scale comparison clean. The twin plays two roles. It supplies the reference model that ratio-based detectors require, and it defines an *oracle* null distribution. Real auditors have no twin, so we also report a *realistic* calibration in which the threshold is set on the audited model's own held-out slice.

3.4 Detectors

We evaluate 9 detectors that read token log-probabilities. Six are taken from the literature: Min-K% (Shi et al., 2024), Min-K%++ (Zhang et al., 2025), zlib-ratio, reference-model perplexity ratio, plain perplexity and windowed perplexity (Carlini et al., 2021; 2023). Three are constructed here. Max-K% averages the *most* likely tokens and serves as a mirror control for Min-K%. Entropy-ratio subtracts the model's own expected token surprisal from the observed log-probability. Z-score divides that residual by the predictive standard deviation. We include the set-level exchangeability test (Oren et al., 2024) and dataset inference (Maini et al., 2024) separately, since they score a benchmark rather than an item.

This suite is deliberately restricted to detectors that score an item from the likelihoods the model assigns to it as presented, which is the family an auditor can run against a served model. It excludes prompting-based probes that ask the model to reproduce withheld fields (Golchin & Surdeanu, 2024; Deng et al., 2024), behavioural detectors that do not reduce to a per-token likelihood, and interventional detectors, which read the same quantities under a manipulated context: CoDeC scores a dataset by how much in-context examples drawn from it shift confidence (Zawalski et al., 2026). None are covered by the sensitivity figures we report, and calibrating them on the same dose ladder is the obvious next application of this protocol.

3.5 Power, MDD, and assurance

Let $s(\cdot)$ be a detector score, oriented so that larger values are more suspicious, let C be a clean reference sample and let T_r be the treated slice at depth r. The threshold at nominal false-positive rate α is the empirical quantile $\tau_\alpha = \widehat{Q}_{1-\alpha}\big(\{s(x) : x \in C\}\big)$, and detection power is the resulting true-positive rate

$$\pi_r(\alpha) \;=\; \frac{1}{|T_r|} \sum_{x \in T_r} \mathbf{1}\big[s(x) > \tau_\alpha\big]. \tag{1}$$

Power at a fixed low false-positive rate, rather than AUROC, follows the evaluation practice argued for in the membership-inference literature (Carlini et al., 2022). Uncertainty is estimated by item bootstrap that resamples C and T_r jointly, so the interval carries both threshold and true-positive-rate variability.

Power should be non-decreasing in dose, so we smooth the observed curve by isotonic regression (Robertson et al., 1988), writing $\tilde{\pi}$ for the result. Two quantities follow. The *minimum de-*

tectable dose is

$$\mathrm{MDD} \;=\; \min\{r : \tilde{\pi}_r(0.05) \geq 0.8\}, \tag{2}$$

reported as >64 when no tested depth qualifies. The *assurance bound* summarizes what a negative reading licenses. Writing π_r^- for the lower 95% bootstrap bound on power,

$$A \;=\; \min\{r : \tilde{\pi}_r^-(0.05) \geq 0.95\}, \tag{3}$$

so that at doses at or above A the calibrated detector has an estimated miss probability below 5%, with a 95% lower confidence bound on the power behind it. Since A is a data-dependent minimum over rungs, that bound is a simultaneous band rather than pointwise, which moves A in 1 of 81 settings. We state A as an empirical threshold, not a confidence interval on the unknown dose: inverting a calibrated power curve into a statement about exposure is the step an independent analysis finds fragile (Shihab et al., 2026). MDD describes the instrument; A describes what a negative reading licenses under the calibration that produced it. Tests across the detector-by-cell grid are corrected for multiple comparisons (Benjamini & Hochberg, 1995).

3.6 Verification

Every run passes four automated gates before analysis: a document ledger check, a token-level search confirming that each injected item occurs in the materialized training stream exactly as many times as its nominal depth, a positional check on injection offsets, and a zero-injection check on every clean twin. The positional gate earned its place: an early version of our pipeline inserted every replica at a block boundary while the measurement code also scored from that boundary, which inflated apparent sensitivity until the gate caught it. Appendix E gives the full procedure and results.

4 Experimental Setup

Models are Pythia-160M, Pythia-410M and Pythia-1.4B (deduplicated) (Biderman et al., 2023), continued pre-trained for 400M tokens on a replay of the Pile (Gao et al., 2021). Every scale uses the same seed, the same injection file and the same effective batch size, so all three run for an identical 3052 optimizer steps under an identical schedule; since Pythia shares one tokenizer across the family, the training token stream is also identical across scales. Benchmarks are GSM8K (Cobbe et al., 2021) and BIG-Bench Hard (Suzgun et al., 2023). Both were released after the Pile cutoff, so the uncontaminated baseline is clean by construction, and both go into the same run, which makes the comparison between them a within-model comparison. Paraphrases come from an instruction-tuned model under a constraint that preserves all numerals; items that violate it are dropped from paraphrase analyses rather than silently reverting to verbatim.

5 Results

5.1 The silent zone

A contaminated model reads its injected answers more confidently, but most of that confidence is not about those answers: across the 36 cells the rise measured on the uninjected holdout accounts for a median 87% of the raw gap (range 47–101%). The injected documents teach the answer format, and the format transfers to items never injected. We therefore read memorization as a difference in differences against the holdout of the same benchmark; in 13 cells the surviving item-specific component is indistinguishable from zero.

Table 1 and Figure 2 in Appendix A show one pattern repeated across settings. At $r{=}1$ every detector sits at its nominal false-positive rate. At $r{=}4$ the best detector on the strongest setting reaches 0.33, while the item-specific log-probability gain is already $+0.27$ nats (95% CI $[+0.16, +0.38]$), against a format baseline of $+1.19$ nats on items the model never saw.

We call a cell *silent* when the best of the 9 detectors falls below 0.5 power while the item-specific gain is significantly positive. The 36 memorization tests behind that second condition are one family, corrected together by Benjamini–Hochberg at $q = 0.05$; this leaves 23 significant and does not change the count. 8 of 36 cells qualify, the smallest surviving gain is +0.13 nats, and the largest

Table 1: Minimum detectable dose: the smallest duplication depth at which a detector reaches 0.8 power against a threshold calibrated to a 5% false-positive rate. Columns are model scale × setting, where G-v and G-p are GSM8K verbatim and paraphrased and B-v is BBH verbatim. Entries of >64 mean the detector never reaches 0.8 power anywhere on the ladder. The full power-by-depth grid is in Appendix D.

	160M			410M			1.4B		
Detector	G-v	G-p	B-v	G-v	G-p	B-v	G-v	G-p	B-v
Min-K%	>64	>64	>64	16	64	>64	4	64	>64
Min-K%++	>64	>64	>64	16	64	64	4	16	16
zlib-ratio	>64	>64	>64	16	64	>64	4	64	>64
PPL-ratio	>64	>64	>64	64	>64	>64	4	>64	>64
Perplexity	>64	>64	>64	16	64	>64	4	64	>64
Max-K%	>64	>64	>64	>64	>64	>64	>64	>64	>64
Window-PPL	>64	>64	>64	64	64	>64	16	16	>64
Entropy-ratio	>64	>64	>64	16	64	>64	4	16	16
Z-score	>64	>64	>64	16	64	>64	4	16	16

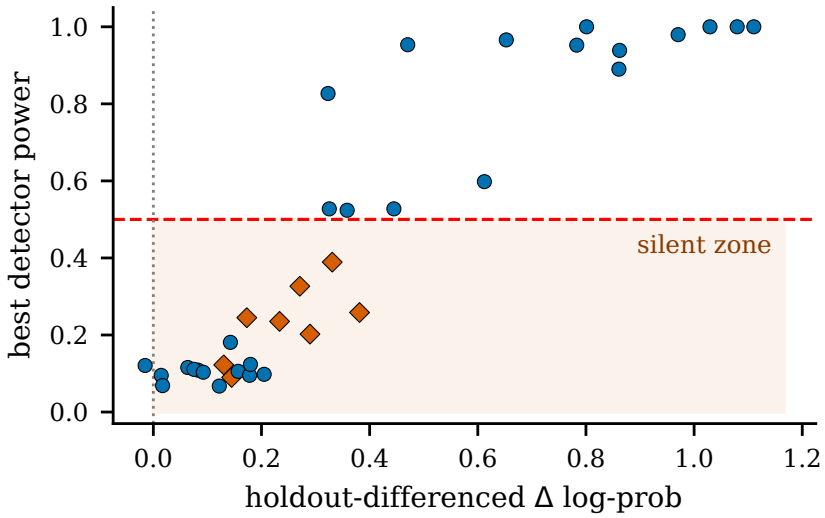


Figure 1: The separation this paper measures. Each point is a dose cell: the horizontal axis is item-specific memorization, the log-probability gap differenced against the uninjected holdout of the same benchmark so that answer-format learning is removed; the vertical axis is the power of the best of the 9 detectors. Orange diamonds are the 8 silent cells, those whose item-specific gain survives Benjamini–Hochberg correction across all 36 cells while no detector reaches 0.5 power; the shaded region marks that zone.

power reached anywhere inside the zone is 0.39. The zone thins as capacity grows but does not close: it covers 2 cells at 160M, 4 at 410M and 2 at 1.4B, and the surviving cells are the paraphrased and the harder-benchmark ones. At 1.4B what remains silent is paraphrased GSM8K at $r{=}1$ and BBH at $r{=}4$. Figure 1 plots every cell on the memorization-versus-detectability plane.

Does silence matter for reported scores? At 1.4B, averaged over 2 training seeds, four replicas raise BBH exact-match accuracy by $+8.6$ points (95% CI $[+0.7, +16.3]$, bootstrap $p = 0.031$) while the best of 9 detectors reaches only 0.39 power. The interval resamples items and conditions on the two trained pairs, so it is not a statement about the seed population; the per-seed estimates are $+10.2$ and $+6.9$ points. That is a cell in which a benchmark score is materially inflated and no instrument in the per-item suite we evaluate would report anything.

The same dose on GSM8K shows why a detector reading is not a proxy for harm. Four replicas there buy $+9.5$ points ($p = 0.001$), no less than the BBH gain, yet the best detector reaches 0.95 power. One model, one dose, one injection procedure, one detector suite: the damage is the same on

both benchmarks and the chance of noticing it differs by more than a factor of two. A clean detector report certifies the corpus that was searched, not the score that was reported.

The differencing behind these numbers is not optional. The contaminated run also gains 25.8 points on held-out BBH items it never saw, because the injected documents taught it the answer format, so we report a difference-in-differences against the uninjected holdout with a paired item bootstrap; the raw gap would overstate the effect threefold. At $r{=}1$ the same estimator returns $+1.0$ points with an interval spanning zero, so we claim no inflation there.

A contaminated run also differs from its twin twice over: it carries the benchmark text, and that text displaces an equal budget of replay tokens. A third 1.4B run separates them, identical but for injected documents that reproduce the template rotation and per-document token lengths while carrying no benchmark content. Displacement moves all benchmark items alike and contributes almost nothing that distinguishes injected items from held-out ones: differenced against the holdout its share is at most 0.14 nats across the 12 cells, and on accuracy it is indistinguishable from zero at every dose. Removing it anyway leaves $+7.6$ points on BBH and $+11.5$ on GSM8K. Nor is silence a calibration artifact: scored against that control, which never saw benchmark text, the 108 detector-by-cell rejection rates have median 0.061 against a nominal 0.05 (Appendix I).

5.2 Which tail carries the signal

Max-K%, which averages the *highest* token log-probabilities, stays at chance across the whole ladder (0.11 even at $r{=}64$), while Min-K% reaches 1.00. The two statistics differ only in which tail of the token distribution they read, so contamination evidently registers in the lower tail: the tokens a clean model finds hardest are the ones a contaminated model has memorized. Entropy-ratio, which normalizes each token by the model's own expected surprisal, attains 0.30 power at $r{=}4$ and matches the best published detector we tested. We report it because it locates the signal, not because the argument needs a better detector: what matters here is that even the strongest instrument fails at low dose.

5.3 Per-item versus set-level auditing

Depth and breadth are not interchangeable. Table 4 shows that at $r{=}4$ per-item power is 0.20, whereas the same scores aggregated into a set-level test detect the identical contamination with power 0.94 when a quarter of the benchmark is leaked, and 0.47 when only an eighth is. An audit should therefore ask whether *the benchmark* was seen, not whether *this item* was seen; the gain uses the same scores, aggregated differently.

5.4 Scale, benchmark, and surface form

Three comparisons bound the generality of Table 3.

Scale. Detectability climbs steeply with capacity, and on the easiest setting it climbs until the blind spot closes. On verbatim GSM8K the minimum detectable dose falls from >64 at 160M to 16 at 410M and to 4 at 1.4B, where 7 of 9 detectors reach 0.8 power at four replicas and the best reaches 0.95 there, against 0.33 at 410M. A reader who stopped at 410M would conclude that the blind spot is a small-model artifact.

That conclusion does not survive the rest of the grid. At the same 1.4B scale, paraphrased GSM8K still needs 16 to 64 replicas, and on BBH 6 of 9 detectors never reach 0.8 power anywhere on the ladder, with the best reaching 0.39 at $r{=}4$. Capacity buys sensitivity where the target text is verbatim and the benchmark is easy; it buys much less where the leak has been reworded or the benchmark is hard. The blind spot therefore does not shrink uniformly with scale. It retreats along one axis while staying open along the other two, which is why a sensitivity figure has to be reported per benchmark and per surface form rather than per detector.

Item-specific memorization does not shrink with capacity; it grows. At $r{=}4$ the holdout-differenced gain rises from $+0.01$ nats at 160M (95% CI $[-0.12, +0.15]$, indistinguishable from zero) to $+0.27$ at 410M and $+0.78$ at 1.4B. The smallest model learns the answer format and little else, which the raw twin contrast alone reports as strong memorization. Since detectors are typically validated on larger models, published sensitivity figures should not be extrapolated downward.

Benchmark. GSM8K and BBH were injected into the same run, so any difference between them is not a difference between models. They nonetheless yield different MDDs, which makes detector sensitivity a property of the benchmark-model pair rather than of the detector alone. A sensitivity figure measured on one benchmark does not transfer to another.

Surface form. Paraphrased injection raises MDD relative to verbatim injection at matched depth, consistent with reports that rephrasing suffices to evade detection (Dekoninck et al., 2024a; Yao et al., 2024). The difference is one of degree rather than kind. Our paraphrase condition is a faithful rewrite, not an adversarial one, and it already erodes detectability.

5.5 Assurance bounds

Inverting the power curves gives the claim an auditor is actually entitled to make. For most detector-setting combinations a negative result excludes only doses far above what ordinary web-scale leakage produces; on BBH, at every scale, it excludes nothing at all within the tested range. Only at 1.4B on verbatim GSM8K does a negative result start to exclude doses near the observed natural median. Table 2 in Appendix C reports the bound for every detector and setting.

5.6 Two channels: relational and textual memorization

Every detector above scores a *text*. What a contaminated corpus leaks is not only text: an injected item pairs a query with an answer, and the pairing is itself information. To separate the channels we build, for each held-out question, a matched triple: the true answer, an answer taken from a *different item of the same contaminated slice*, and an answer from the clean held-out slice. The first contrast holds textual familiarity fixed and varies only the pairing, isolating relational memorization; the second mixes the two. Figure 3 in Appendix B plots both against dose.

The relational gap grows with dose and is significant at the top of the ladder on both benchmarks ($+0.681$, 95% CI $[+0.569, +0.783]$ on GSM8K; $+0.493$ on BBH), so pairing information is absorbed as a channel of its own. The detectors we calibrate score item-level likelihood shifts and are not calibrated to isolate this relational channel. A leak carrying mostly relational content, as answer keys, relevance judgments and leaderboard tables do, would move the quantity that matters for evaluation validity while leaving little signal in the quantity detectors measure. Calibrating sensitivity on that channel is the natural extension of this protocol, which transfers to it unchanged.

5.7 Observational arm: which doses actually occur

The causal arm establishes what detectors can see at a given dose. Whether that matters depends on the doses the world produces. We therefore turn to a setting with no injection at all. Pythia is pretrained on the Pile (Gao et al., 2021; Biderman et al., 2023), and two benchmarks in our probe set, MMLU and Natural Questions, predate the Pile cutoff, so their n-gram frequency in that corpus is a record of contamination that genuinely happened. We query those frequencies with infini-gram (Liu et al., 2024) and score every item with the same detectors, using the base model with no continued pretraining.

Of 2700 probed items, 247 (9.1%) appear at least once in the Pile, and among those the median count is 3 (maximum 5203). Naturally occurring exposure is shallow and long-tailed, and it is concentrated in the same shallow-exposure regime in which our controlled calibration finds weak power.

The two arms measure related but not identical quantities. An n-gram count of three in a corpus is not the same object as three complete item replicas, and the observational probe uses MMLU and Natural Questions while the causal arm uses GSM8K and BBH. The claim is therefore about regimes rather than numerical identity: what the corpus supplies is a handful of partial exposures per affected item, and a handful of exposures is where the calibration curves are flat. Since our protocol renders each replica as a full independent document, three replicas is if anything the more generous condition, so reading natural exposure as at most comparable to our low-dose cells is the conservative direction.

Detector scores do rise with natural dose, and the association survives controlling for text length and the model's own expected surprisal. Min-K%, for instance, has a dose coefficient of $+0.300$

($t = 5.10$). Max-K% is the exception. Its raw rank correlation with dose is comparable to the others ($\rho = 0.298$ against $\rho = 0.377$ for Min-K%), but its controlled coefficient is -0.019 ($t = -1.03$), so its apparent association is explained by the confound that frequently occurring text is simply easier to predict. Max-K% is thus the only detector that fails in both arms, by two independent routes, which corroborates the tail asymmetry found under injection.

5.8 Robustness and replication

Conclusions do not depend on the calibration constants. Across the 45 detector-setting pairs re-analysed at nominal false-positive rates of 1%, 5% and 10%, tightening or loosening the rate moves an MDD by at most one rung of the ladder; varying the Min-K% tail fraction over 10–30% leaves the MDD unchanged in 10 of 10 cases; and dropping any single detector leaves the statement that none achieves 0.5 power at $r \leq 4$ intact in 36 of 45 leave-one-out combinations (Appendix F). Retraining both the 410M and the 1.4B pair end to end under a second seed leaves the MDD unchanged in 18 of 18 settings at 410M and 21 of 27 at 1.4B, with a largest single-cell power difference of 0.165; the set of silent cells at 1.4B is identical across the two seeds. A further 1.4B run injects only the BBH r=4 slice, so that dose exists without the others. Detection there falls to 0.20 from 0.39, the doses left uninjected sit at 0.08 or below, and the format transfer measured on the holdout drops from +25.8 to +11.9 points: isolating the dose deepens the blind spot rather than explaining it away.

6 Discussion

What a negative result means. In the dose regime the observational arm shows to be typical, "we ran a contamination check and found nothing" carries close to no information unless a sensitivity statement travels with it. Audits should report an assurance bound, not a binary verdict.

What follows for practice. Aggregate first, calibrate the null on the audited model itself, and report the bound rather than the verdict. Appendix K states these recommendations, and the corresponding advice for benchmark authors, in full.

7 Limitations

The causal arm uses continued pre-training rather than training from scratch, and the largest model we contaminate is 1.4B parameters. Detectability rises steeply on the easiest setting and the blind spot there closes by 1.4B, so our result is not that detectors fail everywhere but that where they succeed is narrow and predictable. Whether the cells still silent at 1.4B also close at frontier scale is beyond our design, and the trend inside our range does not extrapolate cleanly because it differs by benchmark and surface form. The observational arm addresses external validity from the opposite side but is correlational, and the two arms index exposure differently, by n-gram count against whole-document replicas and on different benchmarks, so together they support a statement about regimes rather than a numerical identity (§5.7).

Further limits bear on interpretation. Every threshold we report is calibrated on a reference slice assumed uncontaminated, which a long-public benchmark cannot guarantee; a contaminated reference inflates the null and understates power. Inflation inside the silent zone is demonstrated at one scale on two benchmarks, and the paraphrase condition is available for one of them. Verification of the training stream uses deterministic replay rather than the materialized dataset, following a storage incident; replay reproduces the token stream by construction and the twins pass a zero-injection check over the full budget (Appendix E). Finally, all conclusions concern detectors that read token likelihoods without intervening on the context; interventional and gradient-based methods are out of scope.

8 Conclusion

We calibrated contamination detectors as measurement instruments and found a dose regime in which a benchmark score is measurably inflated yet no likelihood-based detector we evaluate can see it, with natural exposure sitting in that same regime. Scale closes this blind spot for verbatim

leaks of easy benchmarks and leaves it open elsewhere. Our protocol and released checkpoints are meant to make "we found no contamination" a claim with a number attached.

REPRODUCIBILITY STATEMENT

Every number in this paper is generated by a scripted pipeline; none is transcribed by hand. The dose design, slice assignment and injection renderer are deterministic given a seed (§3), so the training token stream can be reproduced exactly, and we use that property as a verification gate rather than only as a convenience (Appendix E). The code and artifacts required to regenerate every number, table, and figure in this paper will be released upon publication: the slice files that define each dose cell, the injected corpora, the extracted per-token log-probabilities, the detector implementations, and the analysis code that turns them into the power curves, MDDs and assurance bounds reported here. The known-dose checkpoints for every contaminated run and its paired clean twin will also be released upon publication; they are needed to re-extract likelihoods or to calibrate a new detector, not to verify any result stated here. The observational arm depends on a third party n-gram service (Liu et al., 2024); we release the retrieved counts so that arm can be re-analysed without re-querying. Base models and corpora are public (Biderman et al., 2023; Gao et al., 2021). Hardware requirements are modest: every run in this paper fits on one 24GB consumer GPU.

ETHICS STATEMENT

This work studies how well contamination audits detect leakage, which is information that could in principle help someone hide contamination. We judge the disclosure balance to favour publication. The evasion result is already established in the literature we cite, our injection uses no adversarial transformation, and the parties who most need calibrated sensitivity figures are auditors, benchmark maintainers and reviewers rather than anyone trying to game a leaderboard. The released checkpoints are deliberately contaminated small research models trained on public benchmarks and a public corpus; they are research instruments and are not intended or suitable for deployment. No human subjects, private data or personally identifying information are involved.

AI USE STATEMENT

Generative AI tools were used as research, coding, and writing assistants, including brainstorming and refinement of research ideas, development and debugging of experimental and analysis code, manuscript drafting and editing, and consistency checking. All experimental decisions and final scientific judgments were made by the authors. All experiments were executed by the authors, and all data, numerical results, statistical analyses, citations, and claims were independently verified by the authors, who take full responsibility for the content of the paper.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2304.01373.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021. arXiv:2012.07805.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, 2022. arXiv:2112.03570.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2202.07646.

Yize Cheng, Wenxiao Wang, Mazda Moayeri, and Soheil Feizi. Dyepack: Provably flagging test set contamination in LLMs using backdoors. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv:2505.23001.

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

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *IEEE Deep Learning Security and Privacy Workshop (DLSP)*, 2025. arXiv:2406.16201.

Jasper Dekoninck, Mark Niklas Müller, Maximilian Baader, Marc Fischer, and Martin Vechev. Evading data contamination detection for language models is (too) easy. In *arXiv preprint*, 2024a. arXiv:2402.02823.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. Constat: Performance-based contamination detection in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. arXiv:2405.16281.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. arXiv:2311.09783.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. arXiv:2104.08758.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2402.07841.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight; arXiv:2310.20707.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight; arXiv:2308.08493.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.10160.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *Conference on Language Modeling (COLM)*, 2024. Spotlight; arXiv:2401.17377.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2203.08242.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.06443.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models. In *International Conference on Learning Representations (ICLR)*, 2024. Outstanding Paper Award; arXiv:2310.17623.

Medha Palavalli, Amanda Bertsch, and Matthew R. Gormley. A taxonomy for data contamination in large language models. In *Workshop on Data Contamination (CONDA)*, 2024. arXiv:2407.08716.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. Data contamination through the lens of time. *arXiv preprint arXiv:2310.10628*, 2023.

Tim Robertson, F. T. Wright, and Richard L. Dykstra. *Order Restricted Statistical Inference*. Wiley, 1988.

Tom Sander, Pierre Fernandez, Saeed Mahloujifar, Alain Durmus, and Chuan Guo. Detecting benchmark contamination through watermarking, 2025. arXiv:2502.17259.

Rylan Schaeffer, Joshua Kazdan, Baber Abbasi, Ken Ziyu Liu, Brando Miranda, Ahmed Ahmed, Fazl Berez, Abhay Puri, and Stella Biderman. Quantifying the effect of test set contamination on generative evaluations. *arXiv preprint arXiv:2601.04301*, 2026.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.16789.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. When is benchmark contamination detectable? information limits and power-calibrated audits, 2026. arXiv:2608.07914.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2023. arXiv:2210.09261.

Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.02386.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. Spotlight; arXiv:2406.19314.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. Data contamination can cross language barriers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2406.13236.

Wojciech Zarzecki, Jan Dubiński, and Sebastian Cygert. The reliability gap in benchmark auditing: Distribution shift and scale as failure modes of contamination detection. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2026. arXiv:2606.03305.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in llms via in-context learning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.27055.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. In *International Conference on Learning Representations (ICLR)*, 2025. Spotlight; arXiv:2404.02936.

Zheng Zhang, Qi Liu, Siyuan Liang, Ning Li, Zirui Hu, Weibo Gao, Rui Li, Zhenya Huang, Leszek Rutkowski, Baosheng Yu, and Dacheng Tao. Controllable contamination detection for reliable LLM evaluation with statistical guarantees. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 30122–30143, 2026. doi: 10.18653/v1/2026.acl-long.1390.

A THE SILENT ZONE ON THE MEMORIZATION-DETECTABILITY PLANE

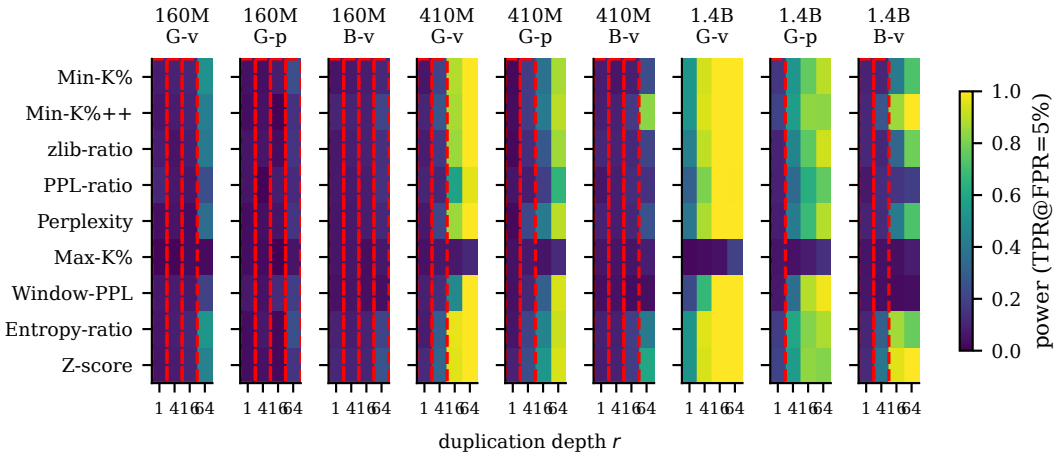


Figure 2: Detection power across dose cells. Dashed red columns mark doses at which no detector exceeds 0.5 power.

B RELATIONAL VERSUS TEXTUAL MEMORIZATION

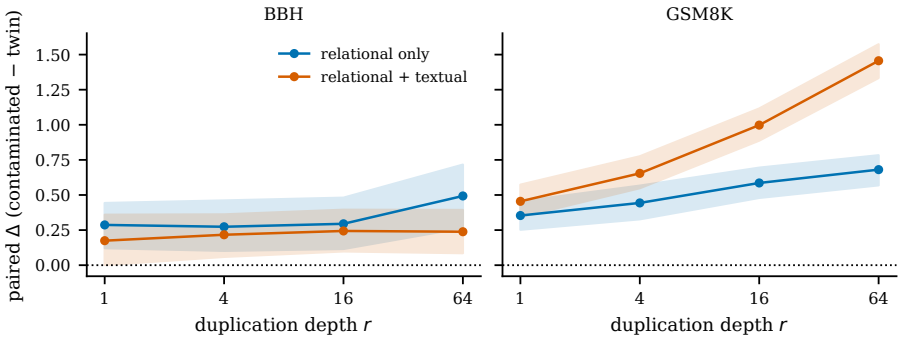


Figure 3: Paired preference gap (contaminated minus clean twin) against dose, for the two contrasts defined in §5.6. The relational-only curve uses distractors whose text the model has also seen during training, so any remaining gap must come from the memorized query-answer pairing rather than from textual familiarity.

C ASSURANCE BOUNDS

The lower band behind A is simultaneous over the 4 rungs. Each of 2,000 bootstrap replicates resamples the calibration slice and the treated slice jointly, so one replicate yields a power estimate

at every rung. Deviations from the point estimates are studentized by the per-rung bootstrap standard error, and we take the 5% quantile of the across-rung *minimum* of those studentized deviations as a single critical value c; the band is then $\hat{\pi}_r + c\,\widehat{\mathrm{se}}_r$, which holds at all rungs at once with probability 0.95. Isotonic smoothing is applied to the band, not to the individual replicates, and these replicates are drawn with a separate seed from those behind the per-cell intervals in Table 1, so adding the band leaves those intervals untouched. The realized critical value has median -1.98 against -1.96 for a pointwise bound, which is why the substitution moves A in only 1 of 81 settings.

Table 2: Dose excluded by a negative result at 95% confidence, that is, the smallest depth at which the lower bound on power reaches 0.95. Dashes mark detector-setting pairs whose negative result excludes nothing within the tested range.

	160M		410M		1.4B	
Detector	GSM8K	BBH	GSM8K	BBH	GSM8K	BBH
Min-K%	—	—	64	—	16	—
Min-K%++	—	—	64	—	16	64
zlib-ratio	—	—	64	—	16	—
PPL-ratio	—	—	—	—	16	—
Perplexity	—	—	64	—	16	—
Max-K%	—	—	—	—	—	—
Window-PPL	—	—	64	—	16	—
Entropy-ratio	—	—	64	—	16	—
Z-score	—	—	64	—	16	—

D FULL POWER-BY-DEPTH GRID

Table 1 in the main text reports the minimum detectable dose, which is the summary an auditor needs. Table 3 below gives the underlying quantity: detection power at every depth in every setting.

Table 3: Detection power (TPR at a threshold calibrated to FPR=5%) by duplication depth. Bold marks power ≥ 0.8. MDD is the first depth at which the isotonically regressed curve reaches 0.8.

Setting	Detector	r=1	r=4	r=16	r=64	MDD
160M/gsm8k/verb.	Min-K%	0.07	0.10	0.11	0.51	>64
	Min-K%++	0.06	0.08	0.10	0.43	>64
	zlib-ratio	0.07	0.05	0.07	0.41	>64
	PPL-ratio	0.12	0.07	0.06	0.24	>64
	Perplexity	0.04	0.05	0.03	0.36	>64
	Max-K%	0.01	0.01	0.03	0.03	>64
	Window-PPL	0.07	0.07	0.07	0.20	>64
	Entropy-ratio	0.06	0.09	0.11	0.53	>64
	Z-score	0.05	0.07	0.10	0.45	>64
160M/gsm8k/para.	Min-K%	0.04	0.04	0.08	0.24	>64
	Min-K%++	0.03	0.05	0.01	0.14	>64
	zlib-ratio	0.06	0.06	0.03	0.15	>64
	PPL-ratio	0.06	0.01	0.08	0.12	>64
	Perplexity	0.03	0.04	0.05	0.15	>64
	Max-K%	0.03	0.03	0.00	0.05	>64
	Window-PPL	0.07	0.05	0.12	0.18	>64
	Entropy-ratio	0.03	0.07	0.02	0.24	>64
	Z-score	0.03	0.05	0.02	0.15	>64
160M/bbh/verb.	Min-K%	0.06	0.07	0.07	0.11	>64
	Min-K%++	0.10	0.10	0.09	0.21	>64
	zlib-ratio	0.07	0.08	0.09	0.13	>64
	PPL-ratio	0.07	0.07	0.07	0.14	>64
	Perplexity	0.06	0.07	0.07	0.12	>64

Continued on next page

Table 3 – continued from previous page

Setting	Detector	r=1	r=4	r=16	r=64	MDD
	Max-K%	0.06	0.06	0.04	0.06	>64
	Window-PPL	0.05	0.05	0.04	0.04	>64
	Entropy-ratio	0.10	0.08	0.10	0.21	>64
	Z-score	0.11	0.10	0.10	0.26	>64
410M/gsm8k/verb.	Min-K%	0.06	0.20	**0.88**	**1.00**	16
	Min-K%++	0.11	0.27	**0.87**	**1.00**	16
	zlib-ratio	0.08	0.16	**0.85**	**1.00**	16
	PPL-ratio	0.10	0.11	0.58	**0.96**	64
	Perplexity	0.07	0.21	**0.85**	**1.00**	16
	Max-K%	0.05	0.07	0.07	0.11	>64
	Window-PPL	0.09	0.13	0.47	**1.00**	64
	Entropy-ratio	0.08	0.30	**0.97**	**1.00**	16
	Z-score	0.12	0.33	**0.95**	**1.00**	16
410M/gsm8k/para.	Min-K%	0.02	0.18	0.37	**0.87**	64
	Min-K%++	0.03	0.18	0.42	**0.90**	64
	zlib-ratio	0.02	0.13	0.27	**0.86**	64
	PPL-ratio	0.06	0.08	0.19	0.65	>64
	Perplexity	0.02	0.22	0.42	**0.90**	64
	Max-K%	0.04	0.07	0.05	0.12	>64
	Window-PPL	0.04	0.11	0.31	**0.94**	64
	Entropy-ratio	0.06	0.21	0.45	**0.92**	64
	Z-score	0.09	0.24	0.53	**0.94**	64
410M/bbh/verb.	Min-K%	0.08	0.06	0.09	0.24	>64
	Min-K%++	0.09	0.09	0.17	**0.83**	64
	zlib-ratio	0.11	0.10	0.11	0.21	>64
	PPL-ratio	0.06	0.08	0.13	0.14	>64
	Perplexity	0.09	0.08	0.11	0.25	>64
	Max-K%	0.05	0.05	0.05	0.09	>64
	Window-PPL	0.04	0.05	0.03	0.04	>64
	Entropy-ratio	0.09	0.09	0.18	0.41	>64
	Z-score	0.08	0.09	0.17	0.60	>64
1.4B/gsm8k/verb.	Min-K%	0.52	**0.93**	**1.00**	**1.00**	4
	Min-K%++	0.52	**0.95**	**1.00**	**1.00**	4
	zlib-ratio	0.44	**0.90**	**1.00**	**1.00**	4
	PPL-ratio	0.31	**0.80**	**1.00**	**0.99**	4
	Perplexity	0.39	**0.88**	**1.00**	**1.00**	4
	Max-K%	0.02	0.03	0.05	0.20	>64
	Window-PPL	0.22	0.67	**1.00**	**1.00**	16
	Entropy-ratio	0.52	**0.95**	**1.00**	**1.00**	4
	Z-score	0.52	**0.94**	**1.00**	**1.00**	4
1.4B/gsm8k/para.	Min-K%	0.16	0.52	0.76	**0.90**	64
	Min-K%++	0.20	0.56	**0.84**	**0.83**	16
	zlib-ratio	0.10	0.45	0.75	**0.93**	64
	PPL-ratio	0.09	0.34	0.63	0.76	>64
	Perplexity	0.08	0.41	0.68	**0.89**	64
	Max-K%	0.06	0.05	0.08	0.14	>64
	Window-PPL	0.07	0.36	**0.89**	**0.98**	16
	Entropy-ratio	0.18	0.60	**0.81**	**0.88**	16
	Z-score	0.20	0.55	**0.84**	**0.82**	16
1.4B/bbh/verb.	Min-K%	0.10	0.11	0.41	0.72	>64
	Min-K%++	0.11	0.27	**0.85**	**1.00**	16
	zlib-ratio	0.11	0.11	0.33	0.78	>64
	PPL-ratio	0.07	0.13	0.15	0.19	>64
	Perplexity	0.10	0.12	0.41	0.72	>64
	Max-K%	0.06	0.05	0.04	0.08	>64
	Window-PPL	0.06	0.04	0.03	0.04	>64
	Entropy-ratio	0.11	0.30	**0.86**	0.77	16
	Z-score	0.12	0.39	**0.95**	**1.00**	16

E VERIFICATION OF THE TRAINING STREAM

Contamination experiments are only as trustworthy as the evidence that the intended dose actually reached the model, so every run passes four gates before any detector is scored.

Ledger. Each dose cell must contain exactly slice size $\times$ depth rendered documents. A mismatch indicates a rendering or assignment fault and blocks analysis.

Occurrence count. A token-level search over the materialized training stream confirms that each injected item appears as many times as its nominal depth. Because injection rotates over five templates while the probe uses one, the expected number of probe hits is $\lceil r/5 \rceil$ rather than r; using r here produces a spurious "data is corrupt" verdict.

Position. Injection offsets must be uniform within blocks. This gate exists because an early version of the pipeline placed every replica at a block boundary while the extraction code also began scoring at that boundary, so the detectors were partly reading position rather than memorization. Apparent sensitivity was inflated until the gate caught it. Offsets are now drawn as $\text{pos} \sim \text{Uniform}(\text{seq_len}, |\text{buffer}|)$.

Zero injection. Every clean twin is checked to contain no injected item at any depth, over the full 400M budget. Twins pass this gate.

Two probes out of the sixteen we sample return more hits than the formula predicts, both drawn from the same BBH subtask (`logical_deduction_three_objects`, one at r=1 and one at r=64 with 25 hits against 13 expected). Items in that subtask share a highly templated stem, so a 40-token probe prefix matches injected documents belonging to *other* items of the same subtask. This is a limitation of probe specificity rather than evidence of a wrong dose: the direction of the error is extra matches, never missing ones, and the ledger and positional gates pass for those cells. The remaining probes match the expected count exactly.

For all runs the stream was re-materialized by deterministic replay rather than read from the original cache, following a storage incident. Replay reproduces the token stream by construction from the same seed and generator, and because Pythia shares one tokenizer across the family, a single replay verifies the stream at every model scale.

F FULL ROBUSTNESS TABLES

The sweeps cover the 5 verbatim scale-benchmark settings for which a paired clean twin exists (160M/GSM8K, 410M/GSM8K, 410M/BBH, 1.4B/GSM8K, 1.4B/BBH), giving 45 detector-setting pairs. Paraphrase cells are excluded because they share a twin with their verbatim counterpart, so including them would double-count the same null. Within that set the MDD is identical at 5% and 10% in 43 of 45 pairs, and identical across all three levels in 33; pairs that differ between 5% and 10% move in the direction of easier detection at the looser rate. Tightening to 1% moves an affected MDD up by at most one rung and never down, which is the expected direction and confirms that no conclusion rests on the choice of 5%. The Min-K% tail fraction is swept over $\{10\%, 20\%, 30\%\}$ and leave-one-out is run over all 9 detectors in the same settings.

G OBSERVATIONAL ARM: NATURAL DOSE

H SET-LEVEL AUDITING ACROSS LEAKED FRACTIONS

I SHAM-REPLACEMENT CONTROL ARM

The contaminated run differs from its twin along two axes at once: it contains benchmark text, and that text displaces replay tokens. The sham run holds the second axis fixed while removing the first. Its injected corpus has the same document count, the same five-template rotation with the same per-template counts, and per-document token lengths matched to the real injection; the filler is ordinary Pile prose and contains no benchmark content, which we verified with 350 probes covering every slice. Length matching is exact to 752 tokens out of 7,686,685 displaced (0.0098%), the residual

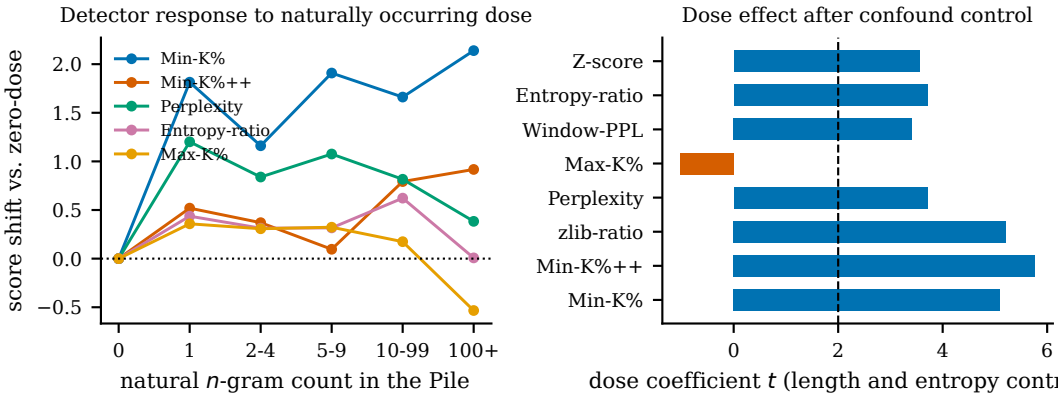


Figure 4: Observational arm. Left: detector scores shift with naturally occurring n-gram frequency in the Pile. Right: dose coefficients after controlling for text length and the model's own expected surprisal. Max-K% is the only detector whose apparent association does not survive.

Table 4: Set-level power as a function of leaked fraction, against per-item power at the same depth (Min-K%).

Depth	per-item	set-level, leaked fraction			
		12%	25%	50%	100%
r=1	0.06	0.24	0.45	0.94	1.00
r=4	0.20	0.47	0.94	1.00	1.00
r=16	0.88	1.00	1.00	1.00	1.00
r=64	1.00	1.00	1.00	1.00	1.00

coming from a decode-truncate-reencode round trip that is not bijective under byte-pair encoding. Every other hyperparameter, including the seed, is identical to the contaminated run.

Writing c, s and t for the contaminated, sham and twin models, $c - t$ is the quantity reported in the main text, $s - t$ isolates displacement, and $c - s$ isolates benchmark content. The three are additive by construction, so the decomposition involves no modelling assumption. Table 5 gives all 12 cells. On raw answer log-probability displacement is negative throughout, a uniform downward shift on benchmark text; once differenced against the holdout it is at most 0.14 nats, so it carries almost nothing that separates injected items from held-out ones. On accuracy its estimate is indistinguishable from zero at every dose. That is why we treat $c - t$ as the primary estimand and $c - s$ as the conservative reading rather than the other way round.

J Training and evaluation configuration

All three scales share every setting below; only the base checkpoint changes.

The micro-batch differs at 1.4B only to fit 24GB of device memory; the effective batch, and therefore the optimizer trajectory, is unchanged. Paraphrases are generated with Qwen2.5-1.5B-Instruct under greedy decoding, with a prompt that requires every numeral in the source question to be reproduced as a digit. Generations that violate the constraint are dropped rather than reverted to verbatim, which removes 457 of 1319 GSM8K items from the paraphrase cells. The drop rate is high, and it is not random: items with many numerals are the ones most likely to be dropped, so the paraphrase cells are mildly biased toward simpler problems. This affects the paraphrase MDDs only, and in the direction of making paraphrased contamination look easier to detect than it would be on the full set. Accuracy is greedy exact match with a per-benchmark parser: the final numeral for GSM8K, and the option letter or normalized label string for BBH.

Table 5: Sham-replacement decomposition at 1.4B. $c - t$ is the contaminated-minus-twin contrast reported in the main text, $s - t$ is displacement alone, $c - s$ is benchmark content alone. Log-probability columns are nats on answer tokens; accuracy columns are points of exact match, each differenced against the uninjected holdout of the same benchmark.

	Δ log-prob (nats)			Δ accuracy (points)			
Setting	$c-t$	$s-t$	$c-s$	$c-t$	$s-t$	$c-s$	p
GSM8K verb. d=1	+1.346	-0.087	+1.432	+2.0	+1.4	+0.7	0.904
GSM8K verb. d=4	+1.770	-0.079	+1.850	+11.6	+0.0	+11.5	0.007
GSM8K verb. d=16	+2.098	-0.078	+2.176	+74.1	+1.4	+72.8	0.000
GSM8K verb. d=64	+2.067	-0.084	+2.151	+95.2	+2.7	+92.5	0.000
GSM8K para. d=1	+1.277	-0.093	+1.370	-0.9	+3.6	-4.5	0.127
GSM8K para. d=4	+1.600	-0.079	+1.679	+2.8	-0.6	+3.4	0.334
GSM8K para. d=16	+1.848	-0.079	+1.927	+17.7	+2.5	+15.3	0.003
GSM8K para. d=64	+1.958	-0.078	+2.035	+28.6	+5.5	+23.1	0.000
BBH verb. d=1	+1.803	-0.637	+2.440	+1.0	+0.0	+0.9	0.827
BBH verb. d=4	+1.954	-0.689	+2.643	+10.2	+2.6	+7.6	0.060
BBH verb. d=16	+2.094	-0.737	+2.831	+29.2	+0.3	+28.9	0.000
BBH verb. d=64	+2.424	-0.552	+2.977	+46.8	-1.8	+48.6	0.000

Table 6: Continued pre-training configuration, identical across scales.

Base models	Pythia-160M / 410M / 1.4B, deduped (Biderman et al., 2023)
Replay corpus	Pile, uncopyrighted subset (Gao et al., 2021)
Token budget	400M (195,312 sequences)
Sequence length	2048
Optimizer	AdamW, learning rate 10^{-4}
Schedule	cosine decay, warmup ratio 0.02
Optimizer steps	3052
Effective batch	64 sequences (131,072 tokens)
Micro-batch × accum.	8×8 (160M, 410M), 4×16 (1.4B)
Precision	bfloat16 with gradient checkpointing
Seed	0 (second seed 1 for the replication in §5.8)
Injected fraction	1.94% of the token budget

K Recommendations in full

For auditors. Aggregate first. Set-level tests reach usable power at doses where per-item scoring is at chance, at no cost beyond a different summary of the same scores. Calibrate empirically: every threshold here was set on a null derived from the audited model itself, and the nominal p-value of at least one published test proved sensitive to an implementation detail (Appendix L). Report the bound rather than the verdict: a negative result is publishable as "under this calibration, a dose of A replicas or more would have been missed with probability below 5%", which a reader can check.

For benchmark authors. Keeping items off the open web matters more than any post hoc detector, since the doses that accumulate naturally sit in the blind spot, which favours withholding plaintext (Jacovi et al., 2023), embedded backdoors (Cheng et al., 2025) and continuous refresh (White et al., 2025). A withheld slice also supplies the uncontaminated reference that calibration needs.

L Empirical null for the exchangeability test

While calibrating the exchangeability test (Oren et al., 2024) we found its nominal p-value highly sensitive to tokenization convention: concatenating documents as text before tokenization, rather than concatenating pre-tokenized documents, changes the sequence log-likelihood by hundreds of nats and the p-value threefold. Under an empirical null from randomized canonical orders it is well calibrated, which is further reason to calibrate detectors empirically rather than trust them analytically. The test of Oren et al. (2024) compares the log-likelihood of a benchmark in its canonical order against permuted orders. We found the nominal p-value to depend on a detail that is easy to

get wrong: whether documents are concatenated as text and then tokenized, or tokenized individually and then concatenated. The two conventions differ in sequence log-likelihood by hundreds of nats and in p-value by roughly a factor of three, because the tokenizer merges across document boundaries in the first case. We therefore tokenize each document separately and concatenate token sequences, which keeps length constant across orders. Under an empirical null constructed from randomized canonical orders, the test separates the contaminated model from its twin (0.915 against 0.652 on GSM8K, and 0.754 against 0.366 on BBH).