Supplement: Language Model Benchmark Contamination Leaves a Person-Fit Signature in the Graded Response Matrix

Artemis Veizi

Stanford University
aveizi@stanford.edu

We collect the main-paper implementation notes behind the controlled-contamination grid, the collateral-damage sweep, the robust-z scaling and null-convention checks, the exact estimator and statistic definitions, the calibration-stability resampling analysis, an independent-cohort replication of the specificity null, reproducibility details, and the supplementary figures referenced from the main text. Section, equation, table, and figure labels are prefixed "S" throughout.

S0 Table 1 Implementation Notes (Controlled-Contamination Grid)

This section collects implementation and scope details behind the controlled-contamination grid (main paper Table 1). None of it is needed to follow the headline claims; it is what a reviewer auditing the pipeline, or pressing on how far those claims extend, will want.

ARC-Challenge 4-choice filtering. *Source:* `pfs-audit/code/benchmarks.py` (`normalize_arc_example`, `load_arc_items`), manifest at `pfs-audit/evaluations/arc_manifest.json`. The raw `allenai/ai2_arc::ARC-Challenge::test` split does not have a uniform number of answer choices per item (mostly 4, but some 3- or 5-choice items). Because the guessing-aware item model fixes $g_j = 1/n_j$ per item and the whole MCQ stack (template, per-letter log-likelihood scoring) assumes exactly 4 choices, non-4-choice items are dropped once at registry build time rather than special-cased downstream. Per the manifest: 7 of 1172 test items were dropped (4 with 3 choices, 3 with 5 choices), leaving 1165 4-choice items, on which $g = 1/4$ is exact for every scored ARC item in the main paper.

Bootstrap resample count for the Table 1 pooled CI. *Source:* `pfs-audit/code/compute_e1_grid_guessing_aware.py` (`boot_ci`, default `B=10000`). The stratified percentile bootstrap 95% CI reported in the Table 1 caption ($[1.00, 1.00]$ on the pooled $\geq 25\%$ contaminated-vs-benign comparison, all three benchmarks) uses 10^4 resamples with a fixed seed (`seed=0`). The other AUROC CIs in the paper use the same stratified resampler and fixed seed at 10^3–2×10^3 resamples (5×10^2 for the calibration-stability table). As with the paraphrase comparison below, a $[1.00, 1.00]$ interval is degenerate under complete separation and records the resampling floor, not zero sampling error.

Nominal versus effective dose, Llama-3.1-8B GSM8K. *Source:* `pfs-audit/results/c4_localization.md`, rows `llama31-8b__gsm8k__inject_verbatim__r25__s0` and `..._s1`. The injected rate reported in the main text is nominal: the base model already answered many injected items correctly, so the number of items whose response the injection can actually flip—the effective dose—is smaller than the nominal fraction. At the 25% nominal injection rate the manifest lists 330 injected GSM8K items, identical across seeds, of which only 51 (seed 0) or 56 (seed 1) were originally answered incorrectly and are therefore flippable at all (the "effective" count in the same table). The true flippable fraction at 25% nominal is 51–56/330, an order of magnitude smaller than the nominal fraction would suggest. Detection tracks this effective dose, not the nominal rate, and the low-rate cells in Table 1 of the main paper sit near that floor; we therefore read the free-form (GSM8K) result in the main paper as detecting targeted eval-set tuning above ecosystem-background exposure, not any exposure whatsoever.

The same gap appears on MMLU, at a milder ratio. *Source:* `pfs-audit/results/c4_localization.md`, rows `llama31-8b__mmlu__inject_verbatim__r25__s0` and `..._s1`. The nominal-vs-effective gap is not unique to GSM8K: at the same 25% nominal rate, Llama-3.1-8B MMLU has 503 nominal injected items against 163 (seed 0) or 156 (seed 1) effective ones ($\approx$ 31–32% of nominal, versus GSM8K's $\approx$ 15–17%). We scope the "targeted eval-set tuning, not any exposure" caveat to GSM8K in the main paper because free-form exact-match is a much stronger signal that an item was not novel to the model than a 4-way multiple-choice hit, which a base model can satisfy by chance one time in four; on MCQ, "originally answered correctly" is confounded with lucky guessing in a way it is not on free-form generation.

A raw-accuracy threshold is weak or blind here. *Source:* `pfs-audit/code/baselines/raw_acc_z.py` (robust-z of benchmark accuracy against the 17-model cohort); GSM8K numbers at `pfs-audit/results/baseline_rawacc__gsm8k.md` (cohort accuracy median 0.839, robust-z of the four contaminated GSM8K con-

ditions from -1.32 to -0.15). The natural baseline is to flag models whose accuracy is anomalously high. On the multiple-choice benchmarks contaminated models do score above the controls (e.g. MMLU median 0.79 vs. the controls' 0.63), so a high-accuracy threshold has some signal there. But it is confounded with genuine capability and cannot tell contamination from a stronger model, whereas person-fit separates the two at ≥ 0.94. On the free-form benchmark the accuracy baseline not only fails but points the wrong way: the contaminated conditions sit below the cohort's median accuracy (robust-z of accuracy between -1.32 and -0.15), because the memorization fine-tuning that installs the contamination also damages untrained items, so an accuracy-threshold detector reads them as less suspicious than average exactly where person-fit fires hardest. In both regimes the contamination signal lives in the response pattern, not the rate.

Degenerate bootstrap CI on the paraphrase separability (exact permutation bound). *Source:* `pfs-audit/results/paraphrase_multibase.md` (pooled row: separability AUROC 1.000 $[1.00, 1.00]$, $n_{\text{con}} = 16$ vs. $n_{\text{ben}} = 8$; per-base rows: 4 contaminated vs. 2 benign each, all AUROC 1.000 $[1.00, 1.00]$). The paraphrase separability reported in the main paper (Yang et al. 2023 paragraph, "The Signal Is Semantic, Not Verbatim"), pooled and per-base alike, is a *complete* separation on the available conditions: every contaminated condition's statistic ranks above every benign condition's, with no overlap. At this sample size the percentile bootstrap used everywhere else in the paper is uninformative, since resampling a perfectly separated sample can only ever reproduce perfect separation. The reported $[1.00, 1.00]$ intervals record the resampling procedure hitting its floor; they are not a claim that the true separation is known to be exactly 1.00 with no sampling error. We therefore also report an exact condition-level label-permutation test. Under the null that contaminated and benign conditions are exchangeable, complete separation of the pooled 16 contaminated from 8 benign conditions has one-sided $p = 1/\binom{24}{8} \approx 1.4 \times 10^{-6}$; a single base's 4-vs-2 comparison gives $p = 1/\binom{6}{2} \approx 0.067$ and is underpowered on its own. Same-rate seed pairs are near-replicates, so the condition-level count overstates the effective n; collapsing seeds still leaves the pooled p at $1/\binom{16}{8} \approx 7.8 \times 10^{-5}$. The pooled comparison is the number that carries statistical weight; the per-base rows are consistent with the same claim rather than independently powered evidence for it.

Free-form paraphrase softening. The multi-base paraphrase separability above is multiple-choice; the free-form (GSM8K) paraphrase arm is Llama-3.1-8B only. There, specificity against benign fine-tuning softens under paraphrase (AUROC 1.000 verbatim to 0.875 paraphrased; the paraphrased estimate rests on four contaminated vs. two benign conditions, bootstrap 95% CI $[0.50, 1.00]$), because that benchmark's large generation space already perturbs the benign controls. Detection is unaffected: every contaminated condition still flags far above the null under paraphrase, at robust-z above twelve. This is why the main paper scopes the clean contamination-versus-benign paraphrase claim to the multiple-choice setting.

S1 Collateral-Damage Sweep (Training Damage Confound, E1)

Source: `pfs-audit/code/simulate_clean_leakage.py`; results at `pfs-audit/results/clean_leakage_sweep__mmlu.md` and `..._gsm8k.md`. The main paper's synthetic clean-leakage construction (see "We can isolate the training damage confound") starts from each base model's own clean response row, marks the injected items correct with zero collateral damage, then sweeps a synthetic damage fraction (random correct→incorrect flips on non-injected items) from 0% up to 19%, the collateral-forgetting rate measured on the real GSM8K memorization adapters; the real MMLU adapters lose under 5%, so the sweep's upper end overshoots the damage the MMLU recipe actually causes. Two channels are tracked separately at each damage level: the positive (unexpected-success) channel Z^+ and the negative (unexpected-failure) channel Z^-.

On both MMLU and GSM8K, across all four injected-rate/seed combinations, Z^+ is already substantially elevated at damage$= 0$ (e.g. Llama-3.1-8B MMLU, r=25%: $Z^+ = 10.4$ at damage$= 0$, seed 0) and roughly doubles over the full sweep ($Z^+ = 20.1$ at damage$= 19\%$, same condition). By contrast Z^- starts near zero at damage$= 0$ ($Z^- = 0.8$) and rises far more steeply ($Z^- = 38.7$ at damage$= 19\%$) — roughly fourfold the per-unit-damage response of Z^+, from a near-zero baseline. This pattern (small Z^+ response, large Z^- response to added damage) holds across all four rate/seed conditions on both benchmarks (full tables in the source `.md` files above).

This is the basis for two claims in the main paper. First, it identifies which statistic to lead with: Z^+ is the damage-robust channel, so it is the right one to report as the headline detection signal, while Z^- loads collateral forgetting and is unsuitable for that role (it is reserved for the signed attribution axis, where the guessing-floor winsorization further corrects it). Second, it shows the main-text separability numbers are not an artifact of the injection recipe's collateral damage: the damage$= 0$ construction (pure memorization, no forgetting) itself separates from the benign controls at guessing-aware Z^+ AUROC 1.00 (MMLU) and 0.99 (ARC; main text), so the separation in main paper Table 1 does not rest on the damage the recipe adds.

Caveat. The sweep induces damage via random flips on non-injected items, which is a mechanistic proxy, not a fit to real (difficulty-correlated) forgetting; the real trained LoRA point sits above the synthetic damage curve at matched accuracy loss (source `.md` files, "Caveat" note). We therefore treat the sweep as isolating the qualitative mechanism (which channel absorbs collateral damage), and the trained-adapter numbers reported in the main paper as the realistic upper bound on the effect.

Training-recipe parity and the intensity confound. All conditions — contamination injection and both benign con-

trols — are trained by one LoRA SFT script (`pfs-audit/code/modal/finetune.py`, `train_one`) with an identical adapter architecture and optimizer: LoRA r=16, α=32, dropout 0, adapters on all attention and MLP projections, AdamW with a cosine schedule, batch size 1, bf16, fixed seed. The recipe differs on training *intensity*: contamination adapters use the memorization recipe (learning rate 1×10^{-4}; 3 epochs for the multiple-choice benchmarks, 8 for GSM8K), while both benign controls use a gentler recipe (learning rate 2×10^{-5}, 3 epochs). This is intentional: applying the memorization recipe to the (up to 2,000-example) control corpora overtrains them into accuracy collapse (e.g. the generic control's GSM8K accuracy fell $0.83 \rightarrow 0.13$), which no longer models a *benign* fine-tune, so we match the controls to realistic benign fine-tuning rather than to the injection recipe's intensity. The separation is nonetheless not a training-intensity artifact: (i) the separation is not carried by total optimization: at the 25% rate the contaminated adapters see *fewer* total optimization steps than the controls (∼300–500 examples per benchmark at 3–8 epochs, vs. the controls' $2{,}000\times3$) yet separate perfectly; (ii) the damage-free construction above reproduces the guessing-aware Z^+ separation with no fine-tuning at all; and (iii) the domain-matched control is trained on same-distribution, item-disjoint data, isolating memorization of the evaluation items specifically from generic fine-tuning perturbation. An explicit intensity-matched benign control (the memorization recipe applied to benign data) would tighten this further; the training code supports it, but no such run is included here.

S2 Robust-z Scaling and Flag-Line Conventions

The main text reports each person-fit statistic as a robust z against the leave-one-model-out (LOMO) empirical null, and uses a flag line at $z_{\mathrm{rob}} > 3$ only to count how often the detector fires on unlabeled released models. This section records the exact scaling and the robustness of the resulting counts to the conventions involved.

Scaling. With s the scored statistic and $\{s_k\}$ the LOMO null, $z_{\mathrm{rob}} = (s - \mathrm{med}\{s_k\})/(1.4826 \cdot \mathrm{MAD}\{s_k\})$, where MAD is the median absolute deviation about the null median and the factor 1.4826 makes the denominator a consistent estimate of the null standard deviation under a Gaussian reference. We use the median and MAD rather than the mean and standard deviation because the null is small and heavy-tailed, so a few extreme cohort values would inflate an ordinary standard deviation and deflate the scores.

Null-inclusion convention. Whether or not a scored row is excluded from its own null median and MAD (leave-each-out versus a fixed shared null) changes the robust scale by under 0.2% at audit scale (and by at most ∼2% on the small controlled-grid null) and leaves every flag count in the main text unchanged, so the two conventions are interchangeable here.

Ability-estimator convention. The flag counts are likewise insensitive to the ability-estimator convention: the guessing-aware Z^+ detection AUROC is nearly identical under the in-sample $\hat\theta$ and the oracle estimator restricted to manifest-known untrained items (main text, Model Ability), and every flag count in the main text is unchanged under either. The estimator choice matters only for the signed attribution axis, where the oracle variant is reported as a controlled-setting upper bound.

Empirical false-positive rate of the flag line. Because the null is heavy-tailed, the $z_{\mathrm{rob}} > 3$ line does not carry the nominal 0.1% Gaussian false-positive rate a $z > 3$ cutoff would imply. We therefore measure the rate empirically, as the fraction of clean cohort models that cross the line under leave-each-row-out. Among the above-floor models the specificity claim covers, this empirical rate is 0.00% (GPQA, $0/89$; a zero count bounds the rate only from above, with one-sided 95% upper limit ≈$3/89 \approx 3.4\%$; Hanley and Lippman-Hand 1983) and 0.36% (MMLU-Pro, $1/286$). On the raw sample before the ability floor is applied it is 2.5–13.6%, which is predominantly the guessing-floor tail that the ability floor removes; the counts reported in the main text are thus rates of the rule as applied, not extrapolations from a parametric tail.

S3 Estimator and Statistic Definitions

With P_j the model-implied success probability at $\hat\theta$ under the guessing-aware 1PL-G item model of the main paper, let $l = \sum_j [x_j \log P_j + (1-x_j)\log(1-P_j)]$ with expectation $E_l = \sum_j [P_j \log P_j + (1-P_j)\log(1-P_j)]$, $V_l = \sum_j P_j(1-P_j)\,\mathrm{logit}(P_j)^2$, $I(\hat\theta) = \sum_j P_j(1-P_j)$, and $c_n = \sum_j P_j(1-P_j)\,\mathrm{logit}(P_j)$. The Snijders-corrected statistic is

$$l_z^* = -\frac{l - E_l}{\sqrt{V_l - c_n^2/I(\hat\theta)}}, \tag{S1}$$

(Snijders 2001), with a fallback to the uncorrected l_z as a numerical guard (the corrected variance is non-negative by construction); the leading sign orients l_z^* so larger = more aberrant, matching every other statistic in the paper. We apply the correction in its Rasch form ($r_j \equiv 1$ in Snijders' notation); the exact 1PL-G score weight $r_j = (P_j - g_j)/[(1-g_j)P_j]$ gives a smaller corrected variance, so the form used deflates l_z^* modestly and is conservative for a comparator statistic — the headline channel Z^+ does not use the correction. The signed channels use $S^+ = \sum_j x_j\sqrt{(1-P_j)/P_j}$ and $S^- = \sum_j (1-x_j)\sqrt{P_j/(1-P_j)}$ with common expectation $E = \sum_j \sqrt{P_j(1-P_j)}$, giving $Z^+ = (S^+ - E)/\sqrt{\sum_j (1-P_j)^2}$ and $Z^- = (S^- - E)/\sqrt{\sum_j P_j^2}$, each centered and scaled to mean 0, variance 1 under the item model at known θ; under the in-sample $\hat\theta$ the realized variance is smaller, one reason every reported value is a robust z against the empirical null rather than a Gaussian tail probability (Section S2). Winsorization replaces P_j by $\mathrm{clip}(P_j, g_j, \phi)$ in both channels, with ϕ fixed a priori at the cohort-unanimity level: with $N = 17$ the empirical proportions quantize to $k/17$, so every $\phi \geq 0.97$ selects the same empirically-unanimous item set, while the clip applied to the continuous model-implied

P_j still varies across the plateau — which is why the main text reports the separation as a band. Only Z^- is materially affected. Comparators: H^T (Sijtsma 1986) is the covariance of a model's responses with the cohort normalized by its maximum attainable value; U_3 (van der Flier 1980, 1982) is the difficulty-weighted departure from the Guttman scale (Guttman 1944) at the model's score. Both are oriented so larger = more aberrant. On the controlled contamination-vs-benign separation ($\geq 25\%$ rate) the comparators reproduce the separation and, pooled across the four base models on ARC, exceed raw l_z^*: H^T and U_3 give AUROC 0.96 and 0.99 against a pooled l_z^* of 0.85 (the guessing-aware Z^+ pools at 1.00 there; main paper Table 1); on MMLU and GSM8K (Llama-3.1-8B) all three statistics give 1.00. The pooled l_z^* dips on ARC only because raw l_z^* carries base-specific null offsets and rank-inverts on the lowest-margin cells when bases are pooled, whereas the per-model-centered H^T, U_3, and Z^+ statistics are largely scale-comparable and do not. The aberrance is thus not an artifact of one particular fit index.

For the attribution axis, an exact label-permutation test over all $\binom{27}{6}$ relabelings of the 6 contamination and 21 sandbagging rows gives one-sided $p = 7.7 \times 10^{-4}$ for the d_{norm} separation (`pfs-audit/code/dnorm_permutation_test.py`). Because the two sandbagging paradigms, not the rows, are the independent units, we read this as an upper bound on significance, consistent with the main paper's construct-level framing.

S4 Calibration Stability

The item difficulties b are calibrated on $N = 17$ baseline models. To check that the calibration is not sensitive to cohort composition, we refit b on resamples of the cohort (leave-one-out and leave-three-out, 25 random resamples each; only 17 leave-one-out subsets are distinct) and measure two quantities. (i) The load-bearing check: whether the contamination-vs-benign Z^+ separability verdict survives under each recalibrated b, since person-fit depends on the cardinal difficulties through $P_j = \sigma(\hat{\theta} - b_j)$ and the guessing floor, not on their rank order alone. (ii) As a supporting diagnostic, the Spearman rank correlation of the resampled b against the full-cohort b. Table S1 reports the result: the detection verdict is invariant on MMLU and holds on ARC down to a leave-three-out minimum of 0.88 over the 25 resamples, with the difficulty ranking highly stable alongside it (leave-one-out Spearman ≥ 0.995; leave-three-out ≥ 0.978 median). The calibration is therefore stable to which cohort models happen to be included.

Benchmark	Resample	b-rank Spearman med. [min,max]	Verdict AUROC med. [min,max]
MMLU	leave-1-out	$0.998\ [0.995, 0.998]$	$1.000\ [1.00, 1.00]$
	leave-3-out	$0.991\ [0.988, 0.993]$	$1.000\ [1.00, 1.00]$
ARC	leave-1-out	$0.995\ [0.969, 0.999]$	$0.992\ [0.97, 1.00]$
	leave-3-out	$0.978\ [0.943, 0.992]$	$0.961\ [0.88, 1.00]$

Table S1: Calibration stability under cohort resampling (25 resamples per row, 1PL-G). b-rank Spearman compares the resampled difficulty ranking to the full-cohort ranking; verdict AUROC is the contamination-vs-benign Z^+ separability recomputed under each recalibrated b.

S5 Independent-Cohort Replication of the Specificity Null

Source: the `evaleval/EEE_datastore` per-item corpus released with Land and Bikel (2026); 66 models $\times$ 198 GPQA-Diamond items, self-calibrated 1PL-G ($g = 0.25$). The main paper's specificity result reproduces on a second, disjoint model set drawn from this public per-item response corpus. We self-calibrate the guessing-aware cohort model on 66 released models answering the 198 GPQA-Diamond items, and score each model's contamination channel. The panel is comparable in size to our main study's GPQA-Diamond panel and independent of it (66 models scored, 40 of them above the ability floor defined below, against 44 above-floor models there), separately calibrated and drawn from a different release era. We obtain zero above-floor flags.

The ability floor. Because the floor carries the flag count, we give it explicitly. On this cohort it is set on the ability scale at $\hat{\theta} > 0$: at $\hat{\theta} = 0$ a model's implied success on a median-difficulty item is $g + (1 - g)/2 = 0.625$, i.e. half the range between the guessing floor and certainty, which is the point below which the residual Z^+ is dominated by the guessing floor rather than by item preknowledge. Forty of the 66 models clear it, and none of the forty flags (largest above-floor robust-z $+2.9$). Two models cross a raw robust-z of three before the floor is applied, and both fall below it: they are low-ability outliers at $\hat{\theta}$ of -6.7 and -2.6, the former near chance accuracy. Their depressed ability inflates the guessing-floor residual—the same artifact isolated in the main paper's specificity study—so neither is a detection. The archived-leaderboard study in the main paper enforces the same floor through an accuracy proxy (chance plus a margin: 0.35 on GPQA, 0.20 on MMLU-Pro), which is available there because the item calibration is fixed in advance; on a self-calibrated cohort $\hat{\theta}$ is estimated directly and is the sharper statement of the same condition. The two conventions differ on this cohort: the accuracy proxy would admit both raw flags, since a model can clear a chance-plus-margin accuracy bar on lucky guesses while its estimated ability sits far below the cohort, which is precisely the case the ability scale handles correctly. The accuracy–Z^+ correlation is -0.09 ($n = 66$, 95% CI $\approx[-0.33, 0.16]$): no cohort-wide ability–misfit gradient, with the residual floor artifact confined to the low-ability tail, as in the primary study. Because the detector consumes only the graded response matrix, this replication required no model inference; the corpus publishes per-item correctness directly. The finding is consistent with the primary specificity result, on models we did not select.

S6 Reproducibility Details: Infrastructure, Hyperparameters, and Statistical Testing

Computing infrastructure. All fine-tuning and model inference ran on Modal cloud instances with a single NVIDIA

A100 (40 GB) GPU per job, in a Debian-slim Python 3.11 container with `torch`≥2.6, `transformers`≥4.55, `peft`≥0.13, `accelerate`≥0.34, `datasets` 2.20.0, `numpy` 1.26.4, `pandas` 2.2.2, and `pyarrow` 17.0.0 (the container specification is recorded in `pfs-audit/code/modal/finetune.py`). All downstream analysis — IRT calibration, person-fit scoring, bootstrap resampling, and figure generation — runs on CPU from the cached response matrices in the Code and Data Supplement, with the analysis environment pinned in `pfs-audit/requirements.txt`; no GPU is required. The one exception is the Min-K%++ cross-check, which requires token log-probabilities and reproduces from its cached scores rather than from the response matrices.

Hyperparameter settings and selection. The detector itself has no trainable hyperparameters; its two analysis-time choices are fixed a priori (the guessing floor $g_j = 1/n_j$, and the winsorization level ϕ, reported as a band over the a-priori clip plateau; main text). The choice of Z^+ as the headline detection channel is made on the mechanistic grounds of Section S1 — Z^- loads collateral damage — not on detection performance. On the training side, the adapter architecture was fixed in advance and shared by every condition (LoRA r=16, α=32, dropout 0, adapters on all attention and MLP projections, AdamW with a cosine schedule, batch size 1, bf16; Section S1). The values tried were learning rate $\in \{2\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}\}$ and epochs $\in \{3, 6, 8\}$, with selection criteria set before any person-fit scoring: injection adapters use the memorization-to-criterion recipe (injected items reproduced under greedy decode; 1×10^{-4}, 3 epochs on the multiple-choice benchmarks, 8 on GSM8K), with one per-base adaptation (Mistral-7B×MMLU: 5×10^{-5}, 6 epochs, adopted after the default collapsed accuracy and before any scoring); both benign controls use the realistic benign recipe (2×10^{-5}, 3 epochs), for the intensity rationale given in Section S1.

Statistical testing and seeds. Every reported uncertainty is a stratified percentile bootstrap 95% CI with a fixed seed; the pooled Table 1, paraphrase, and damage-free comparisons use 10^4 resamples, the remaining CIs 5×10^2–2×10^3 (Section S0). Condition separations are rank-based AUROCs, with exact label-permutation p-values for the pooled paraphrase separation and the attribution axis (Sections S0 and S3); the paraphrase accuracy-gap null result uses the paired McNemar test; and flag counts are always accompanied by *empirical* leave-each-out false-positive rates rather than nominal Gaussian tail rates. Randomness enters at three points, each with a fixed, recorded seed: the sampled contaminated-item subsets (two seeds per condition, reported throughout), the training loop, and the bootstrap resampler.

S7 Supplementary Figures

Figure S1 gives the guessing-floor correction behind the main paper's winsorized attribution result, and Figure S2 the one-dimensional view of the attribution plane.

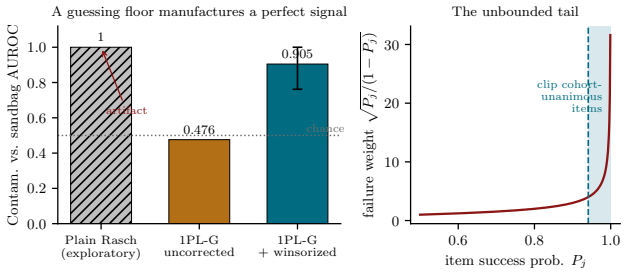

Figure S1: A guessing floor manufactures a spuriously perfect signal. Contamination-vs-sandbagging separation AUROC under three item-model treatments. Plain Rasch appears to separate perfectly (1.00), an artifact. Modeling the guessing floor without bounding the failure channel collapses it to chance (0.476), because the failure-channel weight $\sqrt{P_j/(1-P_j)}$ diverges as items approach near-certain success and detonates Z^-. Winsorizing the cohort-unanimous tail restores a bounded separation (AUROC 0.86–0.91 across the a-priori clip plateau; at the lead clip $\phi = 0.99$, AUROC 0.90, 95% CI $[0.76, 1.00]$). This is why the main paper's reported number is the winsorized guessing-aware statistic, not the Rasch 1.00.

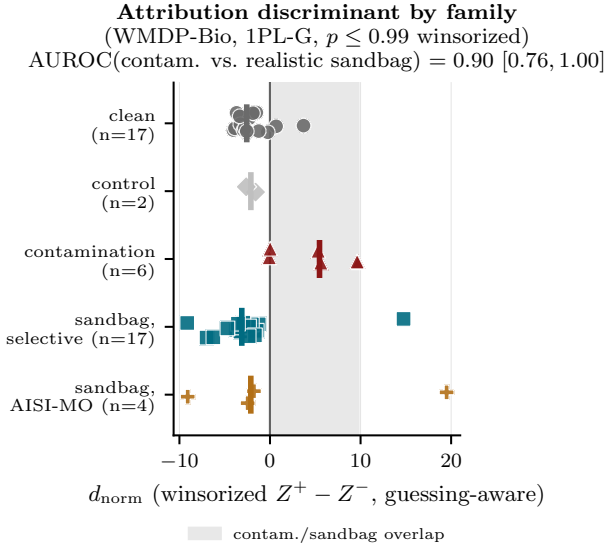

Figure S2: Attribution discriminant d_{norm} by model family (guessing-aware, winsorized), the one-dimensional view of the attribution plane in the main paper. Each point is a model; vertical bars are group medians; the shaded band marks the contamination/sandbagging overlap. The sparse contamination ($n = 6$) and AISI-model-organism ($n = 4$) groups are shown in full so the small-sample basis of the AUROC is visible.

References

Guttman, L. 1944. A basis for scaling qualitative data. *American Sociological Review*, 9(2): 139–150.

Hanley, J. A.; and Lippman-Hand, A. 1983. If Nothing Goes

Wrong, Is Everything All Right? Interpreting Zero Numerators. *Journal of the American Medical Association*, 249(13): 1743–1745.

Land, S.; and Bikel, D. M. 2026. Auditing LLM Benchmarks with Item Response Theory. arXiv:2605.30504.

Sijtsma, K. 1986. A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7: 131–145.

Snijders, T. A. B. 2001. Asymptotic Null Distribution of Person Fit Statistics with Estimated Person Parameter. *Psychometrika*, 66(3): 331–342.

van der Flier, H. 1980. *Vergelijkbaarheid van individuele testprestaties*. Lisse, The Netherlands: Swets & Zeitlinger.

van der Flier, H. 1982. Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13(3): 267–298.

Yang, S.; Chiang, W.-L.; Zheng, L.; Gonzalez, J. E.; and Stoica, I. 2023. Rethinking Benchmark and Contamination for Language Models with Rephrased Samples. arXiv:2311.04850.