WORKS WHERE CALIBRATED, FAILS WHERE MISUSED: AUDITING SMALL-SCALE PRETRAINING'S SEED-NOISE INSTRUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small-scale pretraining selects data recipes under a handful of seeds, read by three instruments. On public per-seed assets (plus two small self-run controls; Appendix H), we audit each where it is used: two of the three reproduce at home (the floor's per-seed data were never released), and each fails where misused — differently for each. The checkpoint-noise proxy of Signal & Noise reproduces at home, but its advertised correlation is beyond certification at realistic seed budgets under the conventional Fisher-z certificate (a perfect proxy reads 0.66–0.82 against 0.9), and it under-reads seed noise in magnitude on the accuracy readout ($\sim 1.6\times$ median, up to $2.4\times$, reversed on perplexity). An init-only noise floor cannot identify the all-source floor from single-source arms at any budget; only a bundled or crossed init$\times$order design can. DataDecide's recipe ranking reproduces its $\sim 80\%$ exactly at its recommended scale, but on the accuracy readout its deconvolved separability SNR stays at or below 1.3 up to 60M, and at 4M the macro ranking inverts against the 1B ranking ($r = -0.56$; robust to recipe resampling, borderline once seed noise is priced in) and seed averaging does not rescue it; the 11-domain perplexity readout does not invert. We release code, a per-cell noise atlas, and report $1{,}095$ label–config-mismatched files in PolyPythias's release.

1 INTRODUCTION

A growing share of what the field believes about data rests on experiments that were never meant to bear that weight: data recipes for billion-parameter runs are selected from experiments two orders of magnitude smaller (Xie et al., 2023; Magnusson et al., 2025), and ablations are certified by margins of a fraction of a percentage point, all under a handful of random seeds. Their validity hinges on a quantity rarely measured and almost never reported: the magnitude and structure of seed-level noise in small-scale pretraining.

The community has begun to close this gap with dedicated *noise instruments*, three of which now carry real decision weight: the checkpoint-noise proxy of Signal & Noise (S&N; Heineman et al., 2025), which reads the spread of a run's final checkpoints as a seed-free substitute for retraining ($R^2 = 0.82$–0.95); the 3-seed init-only noise floor that a 2026 large-scale replication (Zhao et al., 2026) operationalizes as the significance threshold for architectural claims; and DataDecide's single-proxy recipe ranking (Magnusson et al., 2025), whose $\sim 80\%$ pairwise accuracy at 150M is the de facto warranty for small-scale recipe selection.

Each of these instruments works where it was calibrated. The question is whether it works where it is used. We audit all three against the public assets that generated them (Section 2). Two of the three reproduce in their original settings; the floor is the exception — its authors publish no per-seed data, so it is audited structurally, as a practice. Each fails differently in transfer (Figure 1): the checkpoint proxy is unverifiable at the 3-seed budget even when perfect; source equivalence is undecidable on the public arms at any budget; and DataDecide's ranking is accurate at its recommended scale but noise-blind at the 4M–60M scales on the accuracy readout (the bits-per-byte (bpb) readout that million-parameter proxy methods such as RegMix (Liu et al., 2025) actually read is less noise-limited; Appendix G). These are failure boundaries, not refutations.

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.

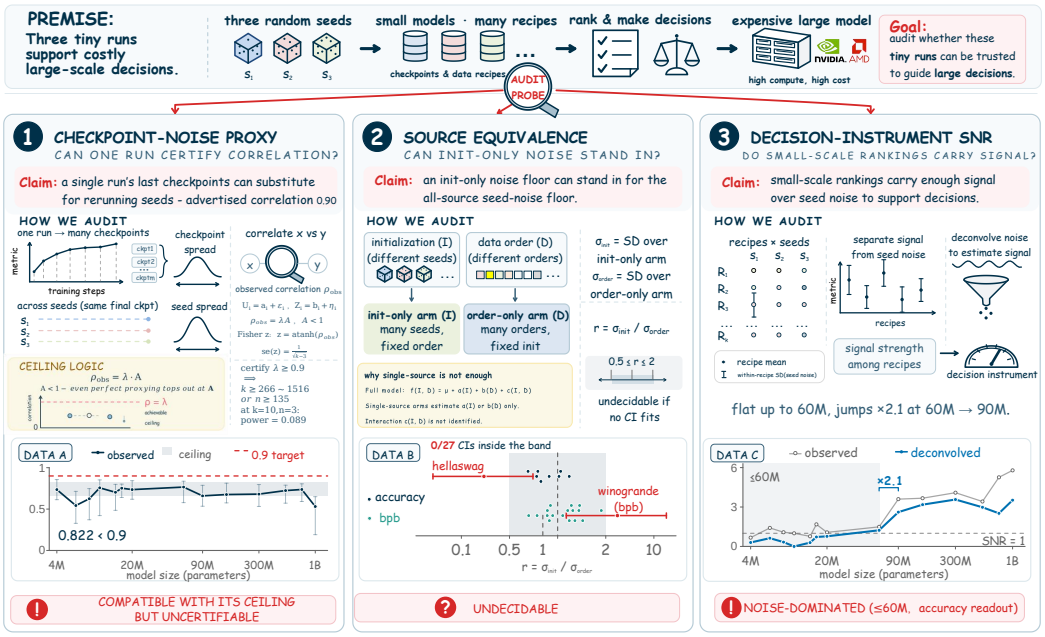

Figure 1: **The audit at a glance.** Three noise instruments carry real decision weight in small-scale pretraining; we audit each where it is used and map how far each survives extrapolation. **(1)** The checkpoint proxy is *compatible with its ceiling but uncertifiable*: a perfect proxy reads 0.66–0.82 vs. the advertised 0.9 (Section 3). **(2)** Source equivalence is *undecidable on public arms*: no task's interval fits the $2\times$ band ($0/27$ under the F intervals, $2/27$ under bootstrap; Section 4). **(3)** The ranking instrument's separability stays ≤ 1.3 up to 60M and jumps $\times 2.1$ at 60M→90M — noise-dominated on the accuracy readout, below the recommended scale (Section 5).

Our findings organize around a single descriptive quantity: the **separability signal-to-noise ratio (SNR)** (Section 2). On the accuracy readout it sits at or below 1.3 at every scale up to 60M and jumps $\times 2.1$ between 60M and 90M; at 4M the macro-average ranking is *negatively* correlated with the 1B ranking, and averaging seeds does not rescue it. Two findings outside the SNR frame: σ itself disperses $2.3\times$ per SD across recipes at 4M, so a floor measured on one recipe is not a community constant; and 1,095 released PolyPythias "seed variant" files are base-model re-evaluations with *exactly zero* seed variance (Appendix F).

The three public-asset audits require no new training: their estimands — whether a proxy correlation is certifiable at the published seed budget, whether single-source arms can identify an all-source floor, whether a released small-scale ranking transfers to 1B — are properties of the released per-seed data; auditing exactly those releases is therefore a census, not a sample. A small set of self-run controls (Appendix H) then supplies the two measurements no public asset contains: the init×order interaction, and the proxy's calibration at a ten-seed budget. Contributions: (i) **a three-target audit with per-claim failure boundaries**, two anchored in home reproductions and one (the floor, whose per-seed data were never released) in a structural audit (Sections 3–5); (ii) **a sample-complexity result and an identifiability note** (Proposition 1; Remark 1); (iii) **the separability SNR as a descriptive readout** — a measured-noise sibling of Signal & Noise's own (head-to-head: Appendix D.5) — with the 4M rank inversion and a per-cell seed-noise atlas; (iv) **prescriptions** (Section 7). The one-line operational version: at a proxy scale whose separability SNR sits at or below 1.3 on the accuracy readout (this panel: every scale $\leq$60M), essentially no pairwise recipe decision on this panel survives multiplicity control at 3 seeds (0–1.3% under all three rules with pair-specific variances) — below that line the supported output is abstention; above it, scale first, then seeds (Section 7).

2 Setup: Audited Claims, Assets, and Declared Defects

We audit the three instruments as they are *used* (at-a-glance map: Figure 1). Terminology (look-up table: Table 2, Appendix A): a *decision instrument* is any cheap measurement whose reading stands in for an expensive one when selecting data recipes or certifying ablations; an instrument's

home setting is the configuration in which its advertised performance was established; and its *usage license* is the set of conditions under which its readings remain trustworthy (stated per instrument in Sections 3–5). Further terms as they appear: an *arm* is one random-source condition (init-only, order-only, or bundled); a *cell* one task×scale×recipe point; a *readout* the metric family (accuracy vs bits-per-byte). **The checkpoint-noise proxy**: Signal & Noise (Heineman et al., 2025) reports that the relative standard deviation of a run's final checkpoints correlates with run-to-run seed noise at $R^2 = 0.82$–0.95 (i.e., $R \geq 0.9$), and proposes the last-n-checkpoint spread as a cheap substitute for seed retraining. **The init-only noise floor**: Zhao et al. (2026) judge all 20 of their architecture claims against a single 3-seed baseline noise floor, "the seed-level distribution as the significance threshold" (their §3.3, recommendation R5). The floor's source semantics are stated inconsistently (§3.3 varies "both initialization and data-shuffle RNG" per seed; their appendix holds "the shuffle seed fixed across all runs" — i.e. init-only as implemented), but either way the floor stands in for the noise of *all* random sources combined; the auditable question is whether a single-source floor can play that role. Their per-seed evaluations are not public as of August 2026; we audit the implemented init-only reading (a bundled reading shifts Section 4's question, not its arithmetic). **Small-scale recipe ranking**: DataDecide (Magnusson et al., 2025) reports that the ranking of 25 data recipes at a single 150M proxy scale predicts the better recipe at 1B with $\sim$80% pairwise accuracy (our reproduction: 80.2% under their single-seed protocol, 82.7% under 3-seed means; Section 5); we audit the residual $\sim$20%.

Assets and declared defects. The public-asset evidence comes from three assets released with per-seed resolution: **DataDecide** (14 sizes, 4M–1B, × 25 recipes × 3 seeds; OLMES suite (Gu et al., 2025) and 11-domain perplexities (Magnusson et al., 2024); 100 tokens per parameter, $\sim$5×Chinchilla (Hoffmann et al., 2022)); **Signal & Noise `random_seeds`** (20 1B runs of OLMo-2-style models (Team OLMo et al., 2025): 10 init-only, 9 data-order-only, 1 dense); and **PolyPythias** (50 Pythia (Biderman et al., 2023) runs, 5 sizes × 10 seeds). Table 1 (Appendix A) inventories them with the defects that constrain our methods, declared before any result.

Notation and estimands. Throughout, σ denotes a standard deviation of final scores across runs; *relative SD* divides by the arm mean. The proxy's *calibration ratio* is y/x (true across-seed relative SD over the checkpoint-spread proxy). For pairwise decisions we use pair-specific 95% Welch intervals for the difference of two 3-seed means: a pair (i, j) sits inside its noise band when $|\bar{y}_i - \bar{y}_j| < t_{0.975,\,\hat{\nu}_{ij}} \sqrt{\hat{\sigma}_i^2/3 + \hat{\sigma}_j^2/3}$, with $\hat{\nu}_{ij}$ the Welch–Satterthwaite degrees of freedom. Bands within a cell are not independent: the 300 bands of a 25-recipe cell share 25 recipe-level variance estimates, so we read band-membership counts as descriptive tallies, not as independent tests.

The separability SNR. The headline quantity of Section 5 is the *separability SNR*: signal = the SD of the 25 recipe means at a given proxy scale; noise = the median within-recipe 3-seed SD. Because each recipe mean itself carries seed variance $\hat{\sigma}_i^2/3$, the observed signal counts noise as dispersion; we therefore also report the *deconvolved* signal $\sqrt{\max(S^2 - \overline{\hat{\sigma}^2}/3,\, 0)}$, whose SNR uses the root-mean-square (RMS) noise $\sqrt{\overline{\hat{\sigma}^2}}$ (diagnostics, robustness variants, and bootstrap intervals: Appendix D, Table 15). The separability SNR is target-free: it measures separability of the proxy's readings, not correctness of its ranking (the transfer metrics carry correctness).

3 The Checkpoint-Noise Proxy: Reproduced Here, Unverifiable There

Signal & Noise (Heineman et al., 2025) proposes that the spread of a training run's final checkpoints can stand in for seed retraining: compute the relative SD of the last n checkpoints of a *single* run, and read it as the relative SD one would get by re-running with new seeds. We reproduce the home result, port the proxy to its advertised use case, test it as a magnitude substitute, and cross-check it on independent 10-seed ground truth (tables and proofs: Appendix B; the three panels: Figure 3).

Replication in the home setting. On the paper's own released runs (Table 3) we obtain $R^2 = 0.81/0.99$ (init, log/raw) and $0.88/0.98$ (data-order), matching the two arm values (0.82 seed noise,

0.86 data-order noise; the third published value, 0.95 for the checkpoint-step readout, we do not recompute here); the calibration ratio y/x has median 0.98. Their printed R^2 values agree with the log-space recomputation (our exact re-implementation of their Fig. 7 protocol gives 0.82/0.84 against their 0.82/0.86; Appendix B.1). **The claim holds in its original setting, and is quantitatively calibrated there** (the raw-space 0.99 rides on `copycolors` alone — without it, 0.40; Appendix B); everything that follows is about transfer.

Porting to the advertised use case: the magnitude gap and the ranking failure. The advertised use is small-scale work: many recipes, few seeds, public checkpoint grids (S&N itself applies the proxy to the 60M–750M DataDecide models). We port the proxy to the same asset (Table 5). Two results carry the transfer.

Magnitude. Pooling all 3,497 cells (size × recipe × task; the coarser 350 macro-point pooling is in Appendix B.2), the calibration ratio y/x has median 1.47, or 1.61 after the log-space small-sample correction (median-basis; the correction is exact only for i.i.d. checkpoint noise — read 1.5–1.6 as the range; convention: Appendix B.5): the proxy *underestimates* true seed noise by about 1.6× (a family×task×seed bootstrap puts the dependence-matched standard error of the pooled median at 0.03–0.06). Only 53.9% of cells fall within a 2× band — and a *perfect* proxy under the same sampling noise would land only 67.9% of cells inside it, so the shortfall is real but shared with perfection, reaching 2.2–2.4× at 6M/60M/150M/300M. On perplexity the same proxy fails in the *opposite* direction (1.1–7× over; below): the bias direction is metric-dependent. (A further stress test — ranking *recipes* by their noise, a use S&N never advertised — finds the proxy near chance at every scale while a perfect proxy reaches only $\tau \approx 0.53$ under the same 3-seed truth; it is the audit's lower-bound witness, detailed in Appendix B.3.)

The advertised correlation, in the background. The certification target here is ours, not Signal & Noise's: they reported the 0.9 as a point estimate and never claimed a certificate; we ask whether such a claim could be validated at this seed budget at all. The observed correlation falls short of the advertised level (per-size medians 0.53–0.77), and at this seed budget the conventional (Fisher-z) certificate of $\lambda \geq 0.9$ cannot attain its nominal level: a sample-complexity bound (Proposition 1, Appendix B.4) puts certification at $k \geq$ 266–1,516 tasks at 80% power, or $n \geq 135$ seeds per arm at the denser m=30 checkpoint budget, and the Monte-Carlo calibration at the panel's own operating point reads Type-I 0.069 at a nominal 0.05 with power 0.12 (Appendix B.4).

A candidate account: the seed-semantics gap. The 1.6× factor admits a simpler reading: the proxy was calibrated on *single-source* seed noise, while DataDecide's seeds bundle init and data order (Section 2), and on S&N's 1B arms independent additivity prices bundled noise at 1.62× the init-only level (Table 11) — close to the proxy's apparent underestimate at the same scale (1.69; the direct empirical cross-check reads 1.06–1.19, compatible with both at df= 2 precision). The reading is a hypothesis we flag as such: its mechanism is absent at 20M, where a conditional init-only arm reads the bundled noise at ≈1.0× — no gap to explain — yet the proxy under-reads by 2.4–4.0× even at a ten-seed budget (intervals wide; Appendix H). **The supported claim: the proxy is calibrated to whatever its training seeds varied; at 1B that is consistent with a single source plus a ∼1.6× semantics gap, and the conversion factor must be re-measured where it is used.**

Continuous metrics and an independent cross-check. On DataDecide's 11-domain perplexity tables the raw proxy fails in the *opposite* direction: it *overestimates* seed noise, by 1.1–7× in per-size medians, from trend contamination on coarse grids; detrending repairs it only at the two largest scales (1.03–1.08 at 750M/1B; Table 8). PolyPythias's 10-seed ground truth confirms the underestimate independently (median 1.8×; Appendix B.7): the bias direction is metric- and grid-dependent, not a stable property one could correct for.

Verdict and usage license. Verdict: calibrated at home; in the field, compatible with its ceiling but not certifiable at this seed budget under the conventional certificate — and reading a narrower noise than advertised. The observed correlation sits below the advertised 0.9 but matches a *perfect* proxy's read at this seed budget (de-attenuated $\hat{\lambda}$ median 0.93; leave-one-seed-out 0.89 with small-scale intervals below 0.9; Tables 5 and 7): the correlation claim is consistent with transfer — not certified by it — and the advertised level stays beyond the conventional certificate's reach at 3 seeds (median ceiling 0.66–0.82 in Table 5; Corollary 1). What fails is the *magnitude* reading

($\sim$1.6$\times$ low on bundled-seed accuracy; reversed on perplexity). **License (candidate operating conditions — observed as separate repairs, never validated end-to-end):** a dense checkpoint grid, a detrended continuous metric, and an explicit seed-semantics conversion; even under the detrended continuous readout at the two largest scales (the 750M row is budget-truncated at 41%), 37–46% of cells fall outside the band. In all other regimes (few-seed comparisons, accuracy metrics at small scale, saturated tasks), seed noise must be measured.

4 Source Equivalence: Undecidable on the Richest Public Arms

As implemented, the Zhao et al. (2026) floor is single-source (Section 2), so its use as the significance threshold prices all-source noise from one source. Because Zhao et al. (2026)'s per-seed evaluations are not public, what we audit is the *practice* — a single-source floor read as an all-source threshold — on the richest public arms with both sources present (S&N's 10 init-only and 9 data-order-only 1B runs); the verdict binds the practice wherever it is used. We assess equivalence against a $[0.5, 2]$ band on $\sigma_{\text{init}}/\sigma_{\text{order}}$ with F and bootstrap intervals (per-task table: Appendix C) — a conventional parity tolerance, descriptive rather than a validity threshold (parity is neither necessary nor sufficient for the floor's validity). Two questions must be kept apart: an *empirical* one — are the two sources comparable in magnitude — which more single-source runs could answer (Table 10), and a *structural* one — does either marginal arm identify the all-source floor at all — which they cannot, at any arm size (Remark 1). The $0/27$ below bears on the first, as a power statement; the second stands by construction.

The median ratio is 0.79 on primary tasks and 1.17 on bpb (point estimates, exploratory only: Appendix C). Not one of the 27 tasks has a CI inside the $2\times$ band ($0/9$ primary, $0/18$ bpb under the analytic F intervals; $2/27$ under bootstrap; task by task: Figure 4) — a precision statement the degrees of freedom force, not evidence about the ratio (even at $n = 10$ per arm the F interval's half-width is $4.03\times$ on the variance-ratio scale; Table 10). The structural point is separate: writing a run's final score as $f(i, d) = \mu + a(i) + b(d) + c(i, d)$, the init arm identifies the law of $\mu + a(I)$, the order arm that of $\mu + b(D)$, and the interaction c lives off both arm lines at *any* arm size (proofs: Appendix C.3). The remark below prices this classical variance-decomposition fact (Searle et al., 1992) at the public arm sizes — stated because it bounds what single-source floors can mean, not because it is new.

Remark 1 (Non-identifiability of the all-source floor). Let the two arm laws have variances $\sigma_i^2, \sigma_o^2 > 0$ and atomless source distributions (a two-point example in Appendix C.3 witnesses the same phenomenon for a discrete law). Then *(i) observational equivalence*: data-generating processes with identical arm laws realize all-source variances σ_{all}^2 of σ_i^2 *(the init-only floor is exact)* and $\sigma_i^2 + \sigma_o^2$ *(independent additivity)* alike. *(ii) one-sided under no cancellation*: if the interaction covaries with neither main effect, $\rho = \sigma_{\text{all}}/\sigma_i \geq \sqrt{1 + \sigma_o^2/\sigma_i^2}$, and that no-cancellation premise is itself untestable from these arms.

So at every public arm size the 27 tasks' arm laws are compatible with both $\rho = 1$ *(the floor is exact)* and independent additivity alike, and no single-source data exclude either.

Read under Zhao et al. (2026)'s bundled §3.3 semantics, the empirical check would need a bundled arm the public record lacks; the structural answer is unchanged (Remark 1).

Verdict. Undecidable on public evidence, by construction. This binds the *practice* — a single-source floor read as an all-source threshold — not Zhao et al. (2026)'s own implementation, whose per-seed data are not public. Under a conditional reading — the target experiment holding the data order fixed and randomizing init only — the same floor estimates a narrower estimand and is not thereby invalid. (The two sources are of comparable magnitude at 1B, so an init-only floor is not orders of magnitude off: under independent additivity the understatement factor is 1.3–1.6, a model-based sensitivity estimate the available cross-check cannot confirm.) **The headline is the design consequence**: what identifies the all-source floor is a bundled arm or a crossed init$\times$order design (Appendix H), not more single-source runs.

5 DECISION REPLAY: 80% CORRECT AT 150M, NOISE-LIMITED BELOW 90M

DataDecide's headline number — rankings at a single 150M proxy scale pick the better 1B data recipe with $\sim$80% pairwise accuracy — is the de facto warranty for small-scale recipe selection. That number averages three *per-seed* decision attempts; their Figure 1 (right) draws the envelope against compute, and our Figure 2 decomposes the same instrument's noise budget by scale and readout. We audit both readings of the instrument on the released tables: target = each recipe's 1B score; prediction = proxy-scale final scores, under single seeds (their protocol) and 3-seed means (our variant); decisions = all $\binom{25}{2} = 300$ unordered pairwise comparisons, per task and per scale (Appendix D, Table 12).

The claim reproduces exactly. At 150M, macro-average pairwise accuracy is 80.2% under their single-seed protocol: the published $\sim$80% to the digit (Figure 6, left, the final-checkpoint slice of that envelope). With 3-seed means it reads 82.7%. What the envelope does not show: the per-scale noise budget (Figure 2, Table 15), the noise-band coverage, and the off-recommended-scale behaviour, where its own curve dips below chance without comment (the 4M inversion below).

What the residual 20% is made of. We call a comparison *seed-unstable* if the pairwise decision is not invariant within the released seeds: 54% of the 52 macro errors at 150M land on such pairs, a $\sim 2\times$ enrichment over the base rate (Appendix D.2). The count reports where errors land, not what caused them. A hierarchical measurement-error model makes the caveat quantitative (Appendix D.2, Table 14): 13.5% [8.3, 57.0] of the expected residual errors are attributable to seed noise at all; the remainder is latent proxy–target discordance, which no amount of seed averaging can remove — though the interval does not exclude a majority seed-noise share. (Three strictly finer estimands: 54% of observed errors land on seed-unstable pairs, descriptive; 29% the model's truly flippable share; 13.5% the share attributable to seed noise at all.) Read as top-k regret — the practitioner's metric — the realized top-1 pick forfeits 5.9 macro points at 4M and 4.7 at 150M (descriptive replay; non-parametric seed-resample 95% intervals $[1.6, 7.2]$ / $[1.9, 5.3]$; Table 13).

The headline: the separability SNR. Accuracy percentages bury the operational question: *at a given proxy scale, how large is the signal to rank, relative to the noise given?* Table 15 (Appendix) computes, per scale, the separability SNR (Section 2) and the share of the 300 recipe pairs whose mean difference falls inside its pair-specific 95% Welch noise band (a z approximation understates these shares; Appendix D).

The gradient is stark, and it is not smooth (Table 15, Figure 2): the deconvolved SNR reads 0.30 at 4M and 0.77 at 20M, crosses 1 at 60M (1.23), and jumps to 2.61 at 90M (paired-bootstrap 95% CI on the ratio $[1.6, 3.0]$; 3.52 at 1B); the indistinguishable share falls from 94% at 4M to 33% at 1B. The jump rides a smooth accuracy gradient, though (3-seed macro accuracy $71.3 \rightarrow 82.7\%$ from 16M to 150M): the instrument's decisions degrade gradually even where its separability jumps. **At every scale up to 60M the deconvolved SNR stays ≤ 1.3 (upper 95% limits below 1.7), and $\geq 83\%$ of pairs fall inside their own noise band at 4M–20M.** The envelope's lower edge is crossed between 60M and 90M (leave-one-recipe-out robust; 6 of 10 task-level readouts; Appendix D.3) — well below the recommended 150M; the bpb readout reads higher throughout (Appendix G), so the noise-limited regime is the accuracy readout's, not the instrument family's. Nor is the SNR readout tautological with accuracy: the per-task deconvolved SNR — which never sees the 1B target — predicts the per-task pairwise accuracy (Spearman $+0.39$ at 4M rising to $+0.90$ at 150M and $+0.92$ at 530M; pooled $+0.89$, $p < 0.001$), including split-seed ($+0.68$; Appendix D).

At the smallest scale, the macro readout inverts. At 4M the macro decision accuracy is *below the coin flip* (31.0% with 3-seed means against 37.8% with single seeds — both below 50%; Table 12), and seed averaging does not rescue it. The 4M macro ranking of the 25 recipes correlates with the 1B ranking at $r = -0.56$ (Pearson, $p = 0.004$; Spearman -0.52), climbing monotonically to $+0.91$ at 530M. The inversion is an aggregation effect, not a per-task property: per-task correlations are positive on 5 of the 10 tasks while the macro average flips negative. It survives the aggregation rule, a task-level bpb re-evaluation (-0.33), 9 of 10 leave-one-task-out drops, and a family-cluster bootstrap ($[-0.72, -0.08]$, $P(r{>}0) \approx 0.01$; wild-cluster $[-0.58, -0.34]$; Ap-

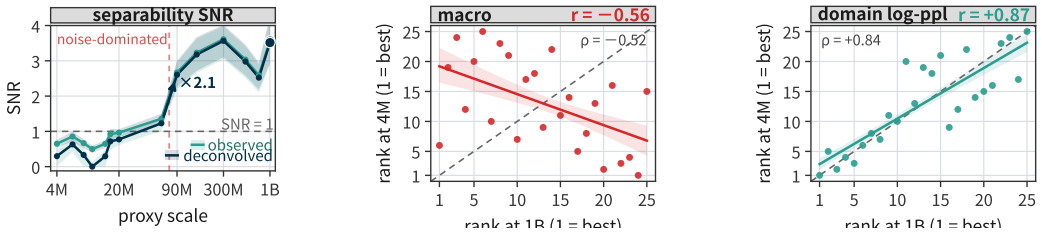

Figure 2: **The decision-SNR audit, graphically.** (a) Separability SNR by proxy scale, observed and deconvolved; the 60M→90M transition jumps $\times 2.1$ (intervals: Table 15). (b) The 4M inversion, macro readout: each ring is a recipe's rank at 1B (x) vs 4M (y); Spearman $\rho = -0.52$ on the plotted ranks (Pearson $r = -0.56$ on scores). (c) The 11-domain log-perplexity readout: $\rho = +0.84$ ($r = +0.87$ on scores) — the inversion is specific to the downstream-task readouts, not the metric family.

pendix D); propagating the 3-seed estimation noise at both ends widens the interval to $[-0.68, 0.08]$ — the inversion is stable to recipe resampling, and borderline once seed noise is priced in; and it is not an artifact of the 1B target (-0.52 to -0.68 against *every* evaluated larger scale's ranking, 60M–1B). A documented cause of such inversions — repetition-rate shifts (Zhou et al., 2026) — does not apply (3 of 25 pools repeat at 1B; excluding them leaves -0.53). The inversion tracks the evaluation *target*, not the metric family: on DataDecide's 11-domain log-perplexity readout there is none — the 4M ranking correlates with 1B at $r = +0.87$, pairwise accuracy 85.3% — but on *task-level* bits-per-byte (S&N's re-evaluation of the same checkpoints) the 4M ranking still inverts against the 1B accuracy ranking ($r = -0.33$; pairwise accuracy 38.7%), at a task-bpb separability SNR of 2.36 — $3.0\times$ the paired primary-metric task median (0.79), $5.0\times$ after deconvolution (Table 20) — the clearest instance of separability being necessary, not sufficient. The failure mode lives on the downstream-task readouts, whichever the metric (Appendix D).

Estimand, dependence, verdict. This SNR audits one published decision procedure on its own candidate set; it is not a transferable constant of small-scale pretraining. The 300 pairs share 25 recipes, so pair counts are descriptive coverage, with inferential weight carried by the per-scale aggregates. **Verdict: accurate at its recommended scale, with a mapped operating envelope**: the published claim reproduces at 150M; the SNR gradient delimits where the instrument can see, and the 4M inversion marks where the macro readout fails. Pairwise accuracy and argmax luck are different estimands: the realized top-1 loss is non-monotone — 5.9 points at 4M, 1.9 at 60M–90M, 0.6 at 300M/530M, 4.7 at the recommended 150M, whose pick ranks 21st of 25 (Table 13).

6 The σ Atlas: Seed Noise Is Not Recipe-Invariant

Behind all three audits sits one quantity: the per-cell seed noise σ, released as an atlas (14 sizes $\times$ 25 recipes $\times$ 11 domains, the 3-seed SD of final log-perplexity per cell, 3,850 cells; Table 18, Appendix E). Two structural conclusions follow.

First, the floor is far from zero where small-scale recipe work is done: median σ is 0.008–0.029 nat over 4M–20M (the 20M elevation: Appendix E), and rank-adjacent recipes sit a median 0.5–1.9 seed SDs apart on the log-ppl readout over 4M–20M (quantiles: Appendix E). The floor then falls with scale (0.0043 nat median at 150M).

Second, σ is not a constant one could measure once and export. Varying *only the data recipe* at fixed size and domain disperses σ well beyond the χ^2_2 sampling law: at 4M a true spread of $1.5\times$ per SD in the median domain and $2.29\times$ in the most dispersed one (C4-en; 95% recipe bootstrap $[1.47, 3.06]$; homogeneity test $p = 0.005$), $1.78\times$ at 20M (carried by one recipe), and $1.65\times$ at 150M (not significant). The 1B row exercises the null side: observed variance falls *below* the sampling floor and the estimator returns exactly $1.00\times$ (per-domain robustness: Appendix E). σ **is not a transferable constant; it is a function of the recipe** — a floor measured on one rig cannot be a community threshold (at ± 1 SD the 4M recipe dependence alone spans $5.3\times$).

PolyPythias anchors the atlas in absolute accuracy units with 10 seeds per cell (Appendix E). All σ values describe the released assets' regime (published recipe families, $\sim 5\times$Chinchilla budgets); other budgets or recipes require remeasurement.

7 Prescriptions: When Proxies Suffice, When Seeds Must Be Measured

The audits converge on five positive rules.

1. **The checkpoint proxy has a license, and the license has edges.** The candidate license for reading the proxy as a magnitude estimate (within a $2\times$ band in the median) conjoins: a dense final checkpoint grid; a detrended continuous metric; and an explicit seed-semantics conversion — each observed as a separate repair, and the conjunction itself never validated (Section 3; the conversion factor must be re-measured where it is used). Even inside the license 37–46% of cells fall outside the band. Outside these edges, seed noise must be measured; no new correlation claim can rescue it (Section 3).

2. **Reported noise must declare its source semantics, its recipe, and its run noise.** "Seed noise" is ambiguous between init-only, order-only, and bundled sources (Section 4), and the magnitude is recipe-dependent at fixed scale (Section 6). A reported σ without a source and a recipe is not a number others can use — and below all of it sits run-level nondeterminism (identical-seed reruns of our 20M cells move final scores by up to 2.3 points; Appendix H), which a reported σ silently includes.

3. **Compute the SNR and the decidable share before deciding.** Before ranking candidates at a proxy scale, run the reference implementation on the per-candidate, per-seed final scores (Appendix D lists it verbatim; it reproduces Table 15 to the digit), and report the *decidable share*: the fraction of pairs whose difference clears its pair-specific Welch band after multiplicity control. The 300 pairs share 25 recipes, so the control must be dependence-robust — Benjamini–Yekutieli (Benjamini & Yekutieli, 2001), with a parametric max-T bootstrap as a further sensitivity and Benjamini–Hochberg (Benjamini & Hochberg, 1995) kept as the generous reference (Table 16). On this panel's macro readout the share is essentially zero at every scale $\leq$60M under every rule with pair-specific variances (exceptions: 1.3% at 6M under BH, one pair at 16M under max-T; pooling the variances lifts 60M to 34%/19% under BH/BY) — below 60M the supported output is abstention, which avoids the 4M failure case outright (the 4M argmax pick forfeits 5.9 macro points at 1B). Abstention serves certified-difference deliverables; when the deliverable is a pick it is a cost, not a benefit — 60M's raw ranking reads 76.3% pairwise with a 1.9-point top-1 loss, better than 150M's 4.7 (Tables 12 and 13); at $\geq$90M it is rule-dependent (37–62% under BH, 1–16% dependence-robust; variance model matters more than the rule at the boundary — Table 16). The share is *target-free* (resolvability, not correctness; Section 5), and the per-task SNR behind it predicts pairwise accuracy (Spearman $+0.89$ pooled, $+0.39$ within 4M; SNR 0.5–1 $\rightarrow$ 63% accuracy, 88–93% at $\geq$2.5; Appendix D, Table 17).

4. **Scale first; then seeds.** At comparable token budget, moving up in scale beats adding seeds on this panel (80% at 150M$\times$1 seed vs 76% at 60M$\times$3), and also at comparable FLOP (floating-point operation) budgets (78.2% at 90M$\times$1 seed vs 76.3% at 60M$\times$3 seeds, $0.75\times$ the FLOPs by $6ND$; the paired difference is $+1.9$ points with a family-cluster bootstrap interval of $[-9.7, +6.7]$ — a point estimate advantage, not a resolved one). Where the scale is pinned, the decidable share leaves zero at n=4 seeds and passes 50% at n=8 at 60M, while at 4M even n=30 buys only 56% (BH at q=0.05, measured noise held fixed) — at 4M the missing ingredient is signal, not seeds (Appendix D).

5. **At the released 3-seed budget, sub-1σ gaps are not resolved decisions.** At 14m–31m, 1σ of seed noise is already 0.5–1 accuracy point on saturated benchmarks; at 4–20M in log-ppl, 0.008–0.029 nat (Section 6). Smaller margins are noise-band differences; significance needs the pair-specific band: at $n = 3$ the 95% Welch band is 2.0–2.5 times the larger within-recipe SD (Welch df $4 \rightarrow 2$ with imbalance).

8 Related Work

Run-to-run variability is well documented (Summers & Dinneen, 2021; Hu et al., 2023; Picard, 2021; Henderson et al., 2018; Reimers & Gurevych, 2017; Bouthillier et al., 2019); error bars on LLM evaluations, and the small-n statistics beneath them, are quantified and faulted by Miller (2025); Madaan et al. (2025); Bowyer et al. (2025); and the significance-testing and seed-budget guidance literature (Dror et al., 2018; Dodge et al., 2019; Colas et al., 2018) is what Section 7 prices at pretraining scale. Audits that re-analyze released evaluation records without new training include Roelofs et al. (2019); Mathur et al. (2020); Card et al. (2020); Polo et al. (2025); Besiroglu et al. (2025). Nityasya et al. (2023) call the cost of under-measured claims *scientific debt*; rliable (Agarwal et al., 2021) is the closest existing tool. Where this literature measures instability, we audit the instruments built to *price* it.

Three works are closest in method. Dodge et al. (2020) introduce the crossed init×order design our public assets lack, for fine-tuning; we add why the design is *necessary* — the interaction term is unidentified from single-source arms at any budget (Section 4) — and a crossed grid trained at pretraining scale (Appendix H). Bouthillier et al. (2021) account for variance in benchmark comparisons by measuring both seeds and bootstrap resamples (Dehghani et al., 2021); we add the per-cell magnitudes such accounting needs (the σ atlas). Zhou et al. (2026) isolate a driver of small-to-large ranking flips — repetition-rate shifts (Muennighoff et al., 2023); we show the flip can arise with repetition essentially absent (3 of 25 pools repeat at 1B), locate it in the macro aggregation rule, and show that seed averaging does not rescue it (Section 5); Schaeffer et al. (2023)'s near-chance-metrics mirage argument is the closest analogue of the inversion.

On the proxy side, Wortsman et al. (2024) validate small-scale proxies for training instabilities, Choshen et al. (2025) systematize scaling-law estimation, and Bhagia et al. (2025) establish the model-ladder lineage DataDecide's single-scale ranking descends from. Closest in question is Wang et al. (2026): small-proxy data-curation rankings flip under minor training-hyperparameter changes — orthogonal to ours (data×hyperparameter vs seed and readout). Our conclusions bear most directly on data-mixing proxy methods (Ye et al., 2025; Liu et al., 2025; Xie et al., 2023); our SNR readout is a noise-graded sibling of Signal & Noise's own (head-to-head: Appendix D.5). The zero-cost NAS literature faces the same rank-fidelity-under-noise question (Abdelfattah et al., 2021; White et al., 2021); our separability SNR is the same diagnostic with the noise measured rather than proxied. Multi-seed pretraining releases are rare: beyond the two we reuse (PolyPythias's 50 runs and S&N's arms) there is MultiBERTs (Sellam et al., 2022).

9 Limitations

(i) **The audit is re-analysis on one panel, the controls small**: the transfer results live on DataDecide's 25 recipes (one stack, one evaluation suite) and Appendix H's controls (20M, one recipe family); conclusions inherit those regimes. (ii) **Source decomposition exists at few points**: S&N's 1B release, PolyPythias's 160M release, our 20M crossed grid (Appendix H). (iii) **Thin cells, model-based attribution**: atlas cells have $n = 3$ (df $= 2$; across-cell quantiles only), and the 13.5% [8.3, 57.0] attribution is the fitted model's (Table 14; robustness ledger, Appendix D.4). (iv) **We judge detectability, not truth**: an audited paper's conclusions are out of scope. *Analysis freedom*: the three audits' estimands and the assets' defect inventory were fixed before analysis; the separability-SNR readout, 4M inversion, and 60M–90M jump are exploratory findings reported with resampling intervals; an instrument that reproduces where calibrated is not calibrated where used (the remedies are Section 7's).

References

Mohamed S. Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas D. Lane. Zero-cost proxies for lightweight NAS. In *International Conference on Learning Representations (ICLR)*, 2021.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural In-*

formation Processing Systems 34 (NeurIPS 2021), 2021. doi: 10.48550/arXiv.2108.13264. URL https://arxiv.org/abs/2108.13264.

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. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt, 2025. URL https://arxiv.org/abs/2404.10102.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2412.04403.

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 *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023. doi: 10.48550/arXiv.2304.01373. URL https://arxiv.org/abs/2304.01373.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. doi: 10.48550/arXiv.2103.03098. URL https://arxiv.org/abs/2103.03098.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *42nd International Conference on Machine Learning (ICML 2025), Position Paper Track*, 2025. URL https://arxiv.org/abs/2503.01747.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.48550/arXiv.2010.06595. URL https://arxiv.org/abs/2010.06595.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In *42nd International Conference on Machine Learning (ICML 2025)*, 2025. URL https://arxiv.org/abs/2410.11840.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning, 2018. URL https://arxiv.org/abs/1806.08295.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. doi: 10.48550/arXiv.2107.07002. URL https://arxiv.org/abs/2107.07002.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/D19-1224.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. doi: 10.48550/arXiv.2002.06305. URL https://arxiv.org/abs/2002.06305.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. doi: 10.18653/v1/P18-1128.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, et al. OLMES: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. URL https://openreview.net/forum?id=sAFottNlra.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. doi: 10.48550/arXiv.2203.15556. URL https://arxiv.org/abs/2203.15556.

Michael Y. Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. Latent state models of training dynamics. In *Transactions on Machine Learning Research (TMLR 2023)*, 2023. doi: 10.48550/arXiv.2308.09543. URL https://arxiv.org/abs/2308.09543.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2407.01492.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks, 2025. URL https://arxiv.org/abs/2406.10229.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, et al. Paloma: A benchmark for evaluating language model fit. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks*, 2024.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, 2025. doi: 10.48550/arXiv.2504.11393. URL https://arxiv.org/abs/2504.11393.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2020. doi: 10.48550/arXiv.2006.06264. URL https://arxiv.org/abs/2006.06264.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2025. URL https://arxiv.org/abs/2411.00640.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, et al. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Alham Fikri Aji, and Genta Indra Winata. On "scientific debt" in NLP: A case for more rigour in language model pre-training research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Long Papers*, pp. 8554–8572, 2023. doi: 10.18653/v1/2023.acl-long.477. URL https://aclanthology.org/2023.acl-long.477/.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, 2021. URL https://arxiv.org/abs/2109.08203.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2025. doi: 10.48550/arXiv.2402.14992. URL https://arxiv.org/abs/2402.14992.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. doi: 10.18653/v1/D17-1035.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 9175–9185, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/ee39e503b6bedf0c98c388b7e8589aca-Abstract.html.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. Wiley, 1992. doi: 10.1002/9780470316856.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Ruxandra Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT reproductions for robustness analysis. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022. URL https://openreview.net/forum?id=K0E_F0gFDgA.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization, 2021. URL https://arxiv.org/abs/2103.04514.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. PolyPythias: Stability and outliers across fifty language model pre-training runs. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. doi: 10.48550/arXiv.2503.09543. URL https://arxiv.org/abs/2503.09543.

Jiachen T. Wang, Tong Wu, Kaifeng Lyu, James Zou, Dawn Song, and Ruoxi Jia. Can small training runs reliably guide data curation? rethinking proxy-model practice. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026. URL https://openreview.net/forum?id=c7d04ce5e71efdfc7a4a78f7161238a9.

Colin White, Arber Zela, Binxin Ru, Yu Liu, and Frank Hutter. How powerful are performance predictors in neural architecture search? In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2309.14322.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2403.16952.

Yang Zhao, Jiahao Lu, Bin Huang, Guhua Zhang, and Jie Zhou. Most transformer modifications still do not transfer at 1-3B: A 2020-2026 update to Narang et al. (2021) with downstream evaluation and a noise floor, 2026. URL https://arxiv.org/abs/2605.20798. Audited version: v1.

Kevin Zhou, Lisa Alazraki, Kris Cao, and Marek Rei. Repetition mismatch: Why data mixture experiments don't scale and how to fix them, 2026. URL https://arxiv.org/abs/2606.07597.

AI Use Disclosure

We used generative AI tools to aid and polish the writing, to draft sections of the paper, and for research ideation and execution. We did not use them to generate synthetic datasets. The authors reviewed and verified all AI-assisted content and take full responsibility for the final content of this work.

Reproducibility Statement

All three assets are public: DataDecide (`allenai/DataDecide-eval-results`, revision `9919b5a`; `allenai/DataDecide-ppl-results`, revision `c4a9fa3`), Signal & Noise's `random_seeds` release (`allenai/signal-and-noise`), and PolyPythias (`EleutherAI/polypythias-evals`, pulled 2026-08-13; the 1,095 mislabeled files and our exclusion rule are documented in Appendix F). Our analysis code, including the scripts that produce every table, is released at https://anonymous.4open.science/r/seed-noise-instruments-audit-674D/. The same package is included in the supplementary material. The audit analyses are CPU-only; the self-run controls of Appendix H (54 training runs at 20M: 48 for the completed $4\times4\times3$ grid — 16 original + 5 identical-seed re-runs + 27 new — plus 6 extra bundled runs) consumed $\sim$200 A100-40GB GPU-hours in total (cluster accounting; single-GPU jobs).

Ethics Statement

This work analyzes public datasets and released model evaluations; it involves no human subjects and no personally identifiable information. The controls of Appendix H train small (20M) models on public data. We document a release-level labeling defect (1,095 files, Appendix F) for the benefit of downstream users.

Table 1: **Assets and declared defects.** The three public assets used throughout, with the defects that constrain our methods: DataDecide's 750M rows end at 41% of the nominal token budget; DataDecide seeds bundle init and data order; Signal & Noise's two arms are evaluated on misaligned step grids (we truncate both to step $\leq$ 69,000); PolyPythias coverage is uneven, and 1,095 of its released lambada files carry label–config mismatches: base-model re-evaluations filed under seed-variant paths (Appendix F). All analyses inherit these constraints.

asset	contents	size	declared defects
DataDecide (Magnusson et al., 2025)	per-seed, per-step scores on the 10-benchmark OLMES suite + 11-domain perplexities; 14 sizes (4M–1B) $\times$ 25 recipes $\times$ 3 seeds	1.41M rows	750M grid stops at 41% of budget; seeds bundle init and data order
Signal & Noise `random_seeds` (Heineman et al., 2025)	1B runs on one recipe: 10 init-only + 9 data-order-only + 1 densely evaluated; 19 tasks (27 primary+bpb readouts)	20 runs	arm step grids misaligned; both truncated to step $\leq$ 69,000
PolyPythias (van der Wal et al., 2025)	public lm-evaluation JSONs; 5 sizes $\times$ 10 seeds	50 runs	uneven coverage; 1,095 lambada files with label–config mismatches (Appendix F)

A Extended Setup: Assets and Declared Defects

Declared defects, in full. Four properties of the assets constrain what can be concluded, and we declare them before any result. (i) *Budget truncation at 750M.* DataDecide's released evaluation grids are aligned to the 3 seeds' common steps; at 750M the grid stops at $\sim$31B tokens, 41% of the nominal 75B budget. We therefore read 750M "final" values as 41%-budget values and never merge them into cross-scale extrapolations; all sizes $\leq$ 530M, and the 1B rows we use, are at full budget. (ii) *Bundled seeds.* DataDecide varies init and data order jointly across its 3 seeds — the training pipeline derives both streams from the run's single seed argument (its `create_ladder_over_scale_script.py` passes one `--seeds` value per run into the OLMo trainer's single `seed` config field) — so it yields all-source noise only; source decomposition is possible solely on the S&N arms. (iii) *Misaligned arms.* The S&N init arm is evaluated to step 81k and the order arm to 71.5k; we truncate both to step $\leq$ 69,000 ($\approx 5\times$C) and define a run's final score as the mean of its last 3 checkpoints. (iv) *Ragged PolyPythias coverage*, including a mislabeling bug we discovered and disclose here: the released `pythia-410m-seed{1..9}` lambada files evaluate the *base* model, not seed variants (1,095 files excluded by our model-consistency check; Appendix F).

Terminology table. Table 2 collects the terms coined or specialized in this paper.

Table 2: **Terminology.** Terms coined or specialized in this paper, with first use.

term	one-line definition	first use
decision instrument	cheap measurement standing in for an expensive one (retraining)	§2
home setting	configuration where an instrument's advertised performance was established	§2
usage license	conditions under which an instrument's readings stay trustworthy	§2
arm	one random-source condition (init-only, order-only, bundled)	§2
cell	one task$\times$scale$\times$recipe measurement point	§2
readout	the metric family an instrument is read on (accuracy vs bpb)	§2
separability SNR	across-recipe signal over within-recipe seed noise, deconvolved	§2
decidable share	fraction of pairs clearing their Welch band after multiplicity control	§7

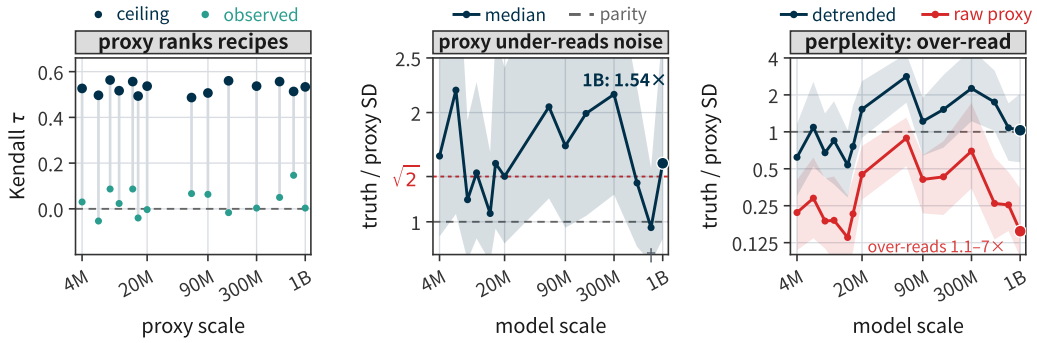

Figure 3: **The proxy audit in three panels.** (a) Ranking recipes by their noise: the observed proxy sits at chance (rings) while even a perfect proxy under 3-seed truth reaches only ≈0.53 (dots) — Appendix B.3. (b) On the accuracy readout the proxy under-reads seed noise at every non-truncated scale (median truth-SD / proxy-SD; parity and $\sqrt{2}$ references; 750M is budget-truncated, and its raw ratio is 0.95) — Appendix B.2. (c) On perplexity the raw proxy *over*-reads 1.1–7× (vermillion); detrending returns the two largest scales to parity — Appendix B.6.

Table 3: **Home-setting replication of the checkpoint-noise proxy.** Cross-task correlation between the final-checkpoint proxy x and true seed noise y on Signal & Noise's own 1B runs, by arm and metric space, plus the calibration ratio y/x (median 0.98). The original claim reproduces.

arm	metric set	n tasks	R (raw)	R^2 (raw)	R (log)	R^2 (log)
init-seed (10 runs)	primary-like	9	0.996	0.992	0.902	0.814
init-seed (10 runs)	bits-per-byte	18	0.953	0.908	0.762	0.581
data-order (9 runs)	primary-like	9	0.991	0.981	0.939	0.882
data-order (9 runs)	bits-per-byte	18	0.951	0.904	0.779	0.607

B THE CHECKPOINT-NOISE PROXY: FULL TABLES AND PROOFS

B.1 HOME REPLICATION TABLE

For each of the 10 init-seed runs, x is the relative SD of the final 30 checkpoints (500-step grid); y is the across-run relative SD of final scores within the arm; the statistic is the cross-task Pearson correlation, in raw and log space. One caveat on the headline R^2: the raw-space 0.99 rides on `copycolors`, the one task whose noise exceeds the others' by 7× — dropping it gives $R^2 = 0.40$ (init, raw) and 0.26 (log); the data-order arm holds up better (0.69/0.73). The correlation is real but its informativeness is concentrated in the high-noise task.

Protocol alignment with the original Figure 7. Their figure's protocol differs from ours in four components; re-implementing each exactly (same released parquet throughout) shows the home claim is robust to every one of them (`sn_exact_protocol.json`). Their printed R^2 values agree with the log-space recomputation (the figure's axes are linear).

Table 4: **Protocol alignment for the home replication.** Per row, one component moved between Signal & Noise's Figure 7 protocol and ours; log-space R^2 for the init-seed and data-order arms. The exact-protocol row reads 0.82/0.84 against the published 0.82/0.86; no single deviation — nor our combination — changes the qualitative result.

variant	tasks	proxy window	target smoothing	step truncation	init R^2 (log)	order R^2 (log)
S&N Fig. 7 (published)	8	last 20	last 20	—	0.82	0.86
exact re-implementation	8	last 20	last 20	—	0.82	0.84
+ ninth task (`copycolors`)	9	last 20	last 20	—	0.93	0.90
+ last-30 window	8	last 30	last 20	—	0.87	0.76
+ last-3 target	8	last 20	last 3	—	0.84	0.85
+ common-step truncation	8	last 20	last 20	step 69000	0.60	0.84
this paper (Table 3)	9	last 30	last 3	common step 69000	0.81	0.88

B.2 PORTING TO DATADECIDE: PER-SIZE TABLE

Porting protocol: for each of 14 sizes $\times$ 25 recipes, x is the relative SD of the default seed's last n grid checkpoints ($n = \max(3, \min(5, \lfloor \text{grid}/2 \rfloor))$), y is the 3-seed relative SD of final scores, and R is computed across the 10 OLMES tasks per cell (350 recipe–scale cells; Figure 3b). The pooled per-task grid is $3{,}497 = 3{,}500$ cells minus three degenerate ones (zero seed variance at 1B `arc_challenge` on DCLM-Baseline 50%/Dolma 50%; zero checkpoint spread at 4M `boolq`/`piqa` on two Dolma variants).

Table 5: **Porting the proxy to DataDecide.** Per size: median cross-task $R_{\log}$ over 25 recipes; the Monte-Carlo ceiling on R for a *perfect* proxy at 3 seeds; the de-attenuated $\hat{\lambda} = \hat{R}/(\text{ceiling median})$ with recipe-cluster bootstrap 95% interval (*capped at 1.00: an uncapped value above 1 signals that the ceiling — computed under independent per-seed errors — is biased downward, consistent with the porting protocol's shared-seed dependence, Appendix B.5; $\hat{\lambda}$ is then optimistic, the safe direction for a non-certification claim); and the calibration ratio y/x, raw and with the median-basis log-space correction. Across sizes, 60–96% of cells are compatible with a perfect proxy and 16–48% indistinguishable from null. No size's interval excludes or certifies the advertised 0.9.

size	med. $R_{\log}$	ceiling med.	$\hat{\lambda}$ [95% CI]	med. y/x	y/x (corr.)
4M	0.74	0.80	0.92 [0.79, 1.04]	1.60	1.60
6M	0.54	0.79	0.69 [0.57, 0.90]	2.21	2.21
8M	0.62	0.79	0.79 [0.54, 0.94]	1.20	1.32
10M	0.76	0.82	0.92 [0.69, 1.03]	1.45	1.59
14M	0.70	0.82	0.86 [0.65, 0.97]	1.08	1.18
16M	0.75	0.79	0.95 [0.82, 1.00]	1.53	1.69
20M	0.74	0.78	0.94 [0.80, 1.06]	1.42	1.56
60M	0.77	0.75	1.00* [0.89, 1.11]	2.05	2.26
90M	0.66	0.71	0.93 [0.75, 1.08]	1.70	1.87
150M	0.68	0.73	0.93 [0.80, 1.10]	1.99	2.19
300M	0.68	0.72	0.95 [0.76, 1.09]	2.17	2.38
530M	0.72	0.72	1.00 [0.86, 1.09]	1.36	1.49
750M	0.74	0.75	0.98 [0.83, 1.07]	0.95	1.04
1B	0.53	0.66	0.80 [0.42, 0.96]	1.54	1.69

Pooled summary. Across all 3,497 cells: median y/x 1.47 raw (1.61 with the median-basis correction); 53.9% of cells inside the $2\times$ band versus 67.9% for a perfect proxy under the same sampling noise (`null_band.json`); the family$\times$task$\times$seed bootstrap SE of the pooled median is 0.03–0.06 (`joint_bootstrap.json`).

The Monte-Carlo ceiling is the distribution of R one would observe if the proxy were *perfect*, given only the sampling error of a 3-seed SD (χ^2 with 2 degrees of freedom) and an n-checkpoint SD.

B.3 RANKING RECIPES BY THEIR NOISE: OBSERVED VS PERFECT-PROXY CEILING

Section 3's ranking claim, in full (Figure 3a). Per (scale, task) cell we rank the 25 recipes by the proxy x and by the 3-seed truth y (both within-task), and compute Kendall's τ and the top-5 overlap; the table reports per-scale medians across the ten tasks. The *ceiling* column answers the objection that the truth itself is estimated at df $= 2$: we re-rank a *perfect* proxy ($x :=$ the true per-recipe noise, taken as the observed y) against a simulated 3-seed estimate of it ($y \cdot \sqrt{\chi^2_2/2}$; $B = 4{,}000$ per cell, fixed seed), which biases the ceiling upward because the observed dispersion already contains sampling noise. A perfect proxy would reach median $\tau \approx 0.53$ (0.49–0.56 across scales); the checkpoint proxy reaches 0.03. (Script: `verify-scripts/perfect_proxy_ceiling.py`.)

Two qualifications. (i) The ceiling is a *perfect-proxy reference* at $n = 3$, not on the proxy: better instruments could exist; under the conventional certificate at this seed budget, none can be *validated* against a 3-seed truth beyond this resolution with useful power. (ii) On the macro variant the observed τ is small but not always zero (median 0.13; 0.39 at 1B, above its own ceiling 0.29) — the macro readout averages away the per-task failure, and we do not claim the macro ranking is at chance everywhere.

Table 6: **Ranking recipes by their noise.** Per scale: median over the ten tasks of Kendall's τ between the proxy's and the 3-seed truth's recipe rankings (observed), and the perfect-proxy ceiling (median over cells of the Monte-Carlo median). The macro variant ranks recipes by the task-averaged noise.

scale	observed τ (task-med.)	ceiling τ (task-med.)	macro: obs / ceiling
4M	0.03	0.53	0.07 / 0.35
6M	−0.05	0.50	0.12 / 0.38
8M	0.09	0.56	−0.18 / 0.29
10M	0.02	0.52	0.33 / 0.33
14M	0.09	0.56	0.07 / 0.37
16M	−0.04	0.49	−0.04 / 0.32
20M	−0.00	0.54	0.21 / 0.30
60M	0.07	0.49	0.23 / 0.23
90M	0.06	0.51	0.31 / 0.31
150M	−0.02	0.56	0.09 / 0.23
300M	0.00	0.54	−0.07 / 0.31
530M	0.05	0.56	0.33 / 0.33
750M	0.15	0.51	0.15 / 0.39
1B	0.00	0.53	0.39 / 0.29

B.4 The attenuation model and proofs

The full statements (stated in the main text through Section 3's summary):

Proposition 1 (Sample complexity of certifying a noisy proxy correlation)**.** *Assume i.i.d. normal arms, i.i.d. tasks, a common cross-task dispersion σ_a^2 of the latent noise levels, and latent correlation $\lambda = \mathrm{corr}(a_i, b_i)$ between the proxy- and seed-arm signals. Then the population correlation of the observed log-SDs is $\rho = \lambda A$ with*

$$A \;=\; \big[\big(1 + \tfrac{\psi_1(\nu_x/2)}{4\sigma_a^2}\big)\big(1 + \tfrac{\psi_1(\nu_y/2)}{4\sigma_a^2}\big)\big]^{-1/2} < 1,$$

independent of proxy quality (ψ_1 trigamma; $\nu_x = m-1$, $\nu_y = n-1$). In the Fisher-z limit experiment, any level-α certificate of $\lambda \geq R_0$ from k tasks needs, for power $1-\beta$,

$$k \;\geq\; 3 + \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\big[\operatorname{atanh} A - \operatorname{atanh}(AR_0)\big]^2}.$$

Corollary 1 (The 3-seed budget cannot certify $R \geq 0.9$)**.** *On the DataDecide panel ($k = 10$ tasks, $m \leq 5$ checkpoints, $\hat{\sigma}_a = 0.74$), a 3-seed arm certifying $\lambda \geq 0.9$ at $\alpha = 0.05$ has power 0.089 even if the proxy is perfect; certification needs $k = 266$–$1{,}516$ tasks by size (median 503), or $n \geq 135$ seeds per arm at $m = 30$. The home budget of Signal & Noise itself attains power only 0.31–0.70.*

The model behind them: write the log proxy SD and log seed SD per task as $U_i = a_i + \varepsilon_i$ and $Z_i = b_i + \eta_i$, where $a_i = \log \tau_i$ is the task's latent noise level with cross-task variance σ_a^2, and the estimation errors are $\varepsilon_i = \frac{1}{2}\log(\chi^2_{\nu_x}/\nu_x)$, $\eta_i = \frac{1}{2}\log(\chi^2_{\nu_y}/\nu_y)$ with $\nu_x = m-1$ (m checkpoints) and $\nu_y = n-1$ (n seeds). Then $\mathrm{Var}(\varepsilon) = \psi_1(\nu_x/2)/4$ exactly (ψ_1 the trigamma function; this is the log-space analogue of the c_4 small-sample SD bias, and it is larger: with $b(\nu) := \mathrm{E}[\varepsilon] = \mathrm{E}[\log S - \log\sigma]$ at ν df, $b(2) = -0.289$ nat vs. $\log c_4(3) = -0.121$). This is an errors-in-variables model with *both* arms noisy, and the population correlation is exactly attenuated. The Fisher-z sample-complexity statement is an asymptotic, model-based approximation at finite k, since the observed log-SD pairs are only approximately bivariate normal; the bound is stated against the most favorable alternative ($\lambda = 1$, A known), an oracle that only helps the verifier. In that limit experiment the power of any level-α certificate of $\lambda \geq R_0$ from k tasks is at most $\Phi\big(\sqrt{k-3}\,[\operatorname{atanh} A - \operatorname{atanh}(AR_0)] - z_{1-\alpha}\big)$; inversion gives the sample-complexity bound of Proposition 1. (With unequal latent dispersions, replace σ_a^2 by the per-arm σ_a^2, σ_b^2 in each factor of A.)

Proof of Proposition 1 (sketch). $\log S = \log\sigma + \frac{1}{2}\log(\chi^2_\nu/\nu)$ gives $\mathrm{Var}(\log S) = \psi_1(\nu/2)/4$. The pairs (U_i, Z_i) are two noisy views of latents correlated at λ, hence $\rho = \lambda\sigma_a^2/\sqrt{(\sigma_a^2 + \mathrm{Var}\,\varepsilon)(\sigma_a^2 + \mathrm{Var}\,\eta)} = \lambda A$: errors-in-variables attenuation (Spearman, 1904) with both arms noisy. Under Fisher's transform $\operatorname{atanh}\hat{r} \to_d \mathcal{N}(\operatorname{atanh}(\lambda A), (k-3)^{-1})$; the composite

null $\lambda \leq R_0$ is least favorable at $\lambda = R_0$, and the z-test is UMP in the limit experiment, giving the power bound; inversion gives the bound on k. Treating A as known is favorable to the verifier (estimating σ_a from the same data can only degrade power), so within this limit experiment the bound is a lower bound on the requirement. □

The full arithmetic behind Corollary 1: on the DataDecide panel ($k = 10$ OLMES tasks, $m \leq 5$ checkpoints, bias-corrected median cross-task dispersion $\hat{\sigma}_a = 0.74$ in log units), a 3-seed arm certifying $\lambda \geq 0.9$ at $\alpha = 0.05$ has power 0.089 under this asymptotic model even if the proxy is perfect; certification at 80% power needs $k = 266$–$1{,}516$ tasks by size (median 503; most favorable decile 173) — not tuned to the advertised level: at the same operating point $k = 146$ tasks suffice for $R_0 = 0.8$ and 243 for $R_0 = 0.85$. Fixing $k = 10$: with $m = 5$ no seed budget suffices for any panel with $\sigma_a \leq 2.04$; with $m = 30$ one needs $n \geq 135$ seeds per arm at $\sigma_a = 0.75$ ($n \geq 27$ at $\sigma_a = 1.0$). The home budget of Signal & Noise itself ($n = 10$, $m = 30$, $k = 9$–10) attains power 0.31–0.70 for $\sigma_a \in [0.75, 1.25]$: the advertised $R \geq 0.9$ was a point estimate, not a certified bound, even in the original setting. For 99.7% of our 350 recipe–scale cells the perfect proxy's *population* correlation is already below 0.9. The closed form reproduces the Monte-Carlo ceiling (0.66–0.82) to within 0.03.

Finite-sample calibration of the certificate. The limit experiment is an asymptotic idealization; we calibrate the certificate's actual operating characteristics at the panel's own operating point by Monte-Carlo ($400{,}000$ replicates per cell, simulating the exact model of Proposition 1 at $k = 10$ tasks, $m = 5$ checkpoints, $n = 3$ seeds, $\sigma_a = 0.74$, $R_0 = 0.9$, $\alpha = 0.05$). At the boundary $\lambda = R_0$ the one-sided Fisher-z certificate's actual Type-I error is 0.069 against the nominal 0.05 (the small-k approximation is mildly liberal); its power against a perfect proxy ($\lambda = 1$) is 0.124, close to the nominal 0.089. Under equicorrelated task noise ($\rho_{\text{task}} = 0.2, 0.5$) the test turns conservative (Type-I 0.046, 0.019) and power degrades further (0.083, 0.036); at $k = 50$ the nominal power is still only 0.188. The certificate is thus uninformative in every regime we can simulate, and the asymptotic bound of Proposition 1 is, if anything, optimistic about the verifier. (Script: `verify-scripts/prop1_mc_calibration.py` in the released reproduction package.)

Checkpoint autocorrelation only strengthens the bound. The log-χ^2 model treats the m checkpoint estimates as independent; adjacent checkpoints are in fact autocorrelated (median lag-1 $\rho \approx 0.39$ on DataDecide's 150M per-step series, 0.34 at 90M, 0.22 at 60M). At the implied effective checkpoint df $\nu_x \approx 1.2$–1.7 the attenuation ceiling drops from 0.66 to ≈ 0.42–0.51 — certification is harder, not easier, once the error model is corrected in the direction the data demand.

B.5 De-attenuation, the shared-seed caveat, and the leave-one-seed-out check

The $\hat{\lambda}$ of Table 5 divides the observed median $R_{\log}$ by the Monte-Carlo ceiling median, i.e. by what a perfect proxy would typically read at this seed budget. Two estimators bracket the same quantity: a model-based version, evaluating the A of Proposition 1 per size with per-arm bias-corrected dispersions, gives a median of 1.03 across sizes; the ceiling ratio gives 0.93. The gap is the visible symptom of a dependence in the porting protocol itself: the proxy x is computed on the *default* seed's run, whose final checkpoint also enters the seed SD y, so the two arms' estimation errors are positively dependent, the observed R is inflated, and both de-attenuations are optimistic (Proposition 1 assumes independent arms, so its ceiling is biased downward here). Under a leave-one-seed-out variant (the proxy computed on seed s, the seed SD on the other two seeds, averaged over the three choices; the ceiling drops with $\nu_y = 1$), $\hat{\lambda}$ has median 0.89 across sizes, and at the smallest scales the intervals fall short of 0.9 (4M [0.70, 0.86]; 6M [0.45, 0.70]; 10M [0.59, 0.88]; Table 7). We therefore read Table 5's $\hat{\lambda}$ as consistent with the advertised level, not as evidence for it; certification at useful power remains out of reach either way (Corollary 1).

Bias-correction convention. For median quantities we apply the median of the log-χ^2 law (the median of $\frac{1}{2}\log(\chi^2_\nu/\nu)$ is -0.183 at $\nu = 2$ and -0.088 at $\nu = 4$; the y/x correction factor is 1.10 at $m = 5$, 1.07 at $m = 4$, and 1 at $m = 3$ where the two arms' biases cancel), not its mean; mean-based corrections apply only to mean quantities, and the variance-scale estimators of Appendix D need none.

Table 7: **The leave-one-seed-out (LOO) porting.** Per size: median cross-task $R_{\log}$ with proxy and seed arms sharing no run; the corresponding perfect-proxy ceiling ($\nu_y = 1$); and the de-attenuated $\hat{\lambda}$ with recipe-cluster bootstrap interval. Point estimates above 1 at 90M/150M reflect the noisier 2-seed ceiling, not super-perfect proxies. The intervals resample recipes and tasks with the attenuation factor held at its plug-in estimate; they do not propagate the checkpoint-noise sampling error that Proposition 1 prices, so an interval sitting above 0.9 is not a level-α certificate of $\lambda \geq 0.9$.

size	med. $R_{\log}$ (LOO)	ceiling med. (LOO)	$\hat{\lambda}$ [95% CI]
4M	0.54	0.67	0.79 [0.70, 0.86]
6M	0.32	0.65	0.49 [0.45, 0.70]
8M	0.51	0.61	0.83 [0.60, 0.92]
10M	0.50	0.66	0.76 [0.59, 0.88]
14M	0.58	0.65	0.89 [0.70, 1.00]
16M	0.52	0.61	0.86 [0.61, 0.95]
20M	0.58	0.61	0.95 [0.86, 1.06]
60M	0.55	0.57	0.96 [0.88, 1.17]
90M	0.55	0.53	1.03 [0.98, 1.16]
150M	0.59	0.55	1.07 [0.83, 1.16]
300M	0.47	0.54	0.88 [0.66, 1.08]
530M	0.52	0.53	0.97 [0.72, 1.14]
750M	0.55	0.58	0.96 [0.69, 1.10]
1B	0.31	0.47	0.66 [0.51, 0.94]

B.6 Continuous metrics: trend contamination

On DataDecide's 11-domain perplexity tables (log space), the raw proxy fails in the *opposite* direction: the median y/x is 0.14–0.89 across sizes, i.e., the proxy *overestimates* seed noise by 1.1–7× (Figure 3c). The cause is mechanical: on coarse evaluation grids the final checkpoints are still improving, so the checkpoint spread mixes seed noise with leftover learning trend. Detrending (residual SD after a linear fit on the final window) repairs the calibration — the license criterion: the detrended median ratio within 10% of 1 *and* a majority of cells inside the 2× band — only at the two largest scales, at median ratio 1.03–1.08 at 750M/1B *and* a majority of cells inside the 2× band (63%/54%) — but not at $\leq$ 530M, where detrended ratios still scatter 0.54–2.8 (medians; e.g. 1.09 at 6M, 1.22 at 90M) and 26–64% of cells fall within 2×. The direction of the proxy's bias is therefore metric- and grid-dependent, not a stable property one could correct for.

The repair is also non-monotone in scale: detrended ratios at 60M/300M (2.82/2.26) exceed those at 90M/150M (1.22/1.52). This is not a grid-density effect (the window has five checkpoints at both 60M and 90M), and it persists when linear detrending is replaced by a trend-invariant first-difference proxy (2.41/2.67 at 60M/300M), so it is not an artifact of the linear fit either. We have no mechanism-level account; the license conditions of Section 3 are therefore stated at the median level only.

B.7 Independent cross-check on PolyPythias

PolyPythias provides 10-seed ground truth. Comparing each run's final-window step noise (x, from checkpoints past 0.72× of training) against the 10-seed final-score SD (y) over 9 available size×task cells, the median ratio is 1.80 (IQR 1.22–2.55): the proxy underestimates by 1.8× on an entirely independent suite. The worst cells are saturated tasks (BLIMP at 70m/160m reaches 4.2–5.2×) where step noise collapses once accuracy plateaus but seed noise does not. The failure is therefore not an artifact of DataDecide's grid or release processing. (PolyPythias shares the bundled seed semantics, so this cross-check does not discriminate the semantics reading of Section 3.)

C Source Equivalence: Full Tables, Power Arithmetic, and Proofs

C.1 Per-task source-equivalence table

Setup: for each task we compute the within-arm relative SDs σ_{init} and σ_{order} of final scores (mean of each run's last 3 checkpoints, arms truncated to a common grid; Appendix A), on 9 primary-like

Table 8: **Trend contamination on perplexity.** Calibration ratio y/x of the checkpoint proxy on 11-domain log-perplexity, raw and after detrending the final checkpoint window. Raw ratios of 0.14–0.89 mean the proxy overestimates seed noise 1.1–$7\times$; detrending restores 1.03–1.08 only at $\geq$750M. Ratios are uncorrected; the median-basis correction of the previous subsection shifts them by $\approx +10\%$ at $m = 5$ without changing any direction.

	median ratio y/x		cells within a $2\times$ band	
size	raw	detrended	raw	detrended
4M	0.22	0.62	17%	44%
6M	0.29	1.09	23%	44%
8M	0.19	0.68	18%	51%
10M	0.19	0.85	18%	43%
14M	0.14	0.54	12%	45%
16M	0.21	0.76	21%	53%
20M	0.45	1.53	45%	51%
60M	0.89	2.82	64%	26%
90M	0.41	1.22	38%	64%
150M	0.43	1.52	34%	51%
300M	0.70	2.26	45%	41%
530M	0.26	1.75	23%	46%
750M	0.25	1.08	21%	63%
1B	0.16	1.03	14%	54%

accuracy tasks and 18 bits-per-byte tasks. We assess source equivalence against a $[0.5, 2]$ band on the ratio $\sigma_{\text{init}}/\sigma_{\text{order}}$ (a conventional factor-of-two criterion) and report both F-distribution confidence intervals (df 9 and 8; exact for ratios of unnormalised variance estimates — applied to the relative-SD ratio they neglect the arm means' sampling randomness, which a Monte-Carlo check over the table's CV range 0.005–0.13 at $n = 10/9$ shows is negligible: coverage 0.946–0.951 at nominal 0.95, `cv_f_check.json`) and run-level bootstrap intervals (2,000 resamples; percentile bootstrap intervals on a variance ratio are known to under-cover at these arm sizes, so the bootstrap containment counts are optimistic and the F intervals remain the primary instrument). We report band containment; we never read "not significant" as "equivalent". The per-task σ_{init}–σ_{order} geometry is in Figure 5.

C.2 What it would take: the power arithmetic

The half-width of a variance-ratio CI is a pure function of degrees of freedom, $F_{0.975}(\nu, \nu)$. With $n = 3$ per arm the interval's half-width is $39\times$; $n = 6$ gives $7.15\times$; $n = 8$ gives $4.99\times$; even $n = 10$ gives $4.03\times$, still wider than the $4\times$ band. The S&N arms (df 9/8, half-width $4.23\times$) are already the best the public record offers. Computing the pass probability under a true ratio of exactly 1 sharpens the point: $n = 11$ per arm, the first size at which the band is geometrically reachable, passes only 9% of the time; $n = 16$ passes 47.5%; $n = 21$ passes 71%; only $n \geq 31$ per arm passes 92% (Table 10).

C.3 The non-identifiability remark: statement and proof

Remark 1 is stated in the main text (Section 4, with its all-arm-sizes corollary in prose there); we prove it here. In part (i), the explicit binary example has both arms equal to the same two-point law; the ratio $\sigma^2_{\text{all}}/\sigma^2_i$ is 1, 2, or 9 depending on the unseen corner cell.

Remark 2 (Sharp but vacuous bounds). Given the arm laws alone and scores in an interval of width B, $\sigma^2_{\text{all}} \in [0, B^2/4]$, both endpoints attained (a constant score gives the lower end; Popoviciu's inequality the upper, attained by a two-point law on the boundary; B=1 for accuracy-like metrics, while bpb is unbounded above). In SD units the identified interval for ρ spans $[0, 3.9]$–$[0, 125]$ across primary tasks and $[0, \infty)$ for all bpb tasks: the bounds certify nothing a practitioner could use, so we report them only here.

Proof of Remark 1 (sketch). The arms pin down the laws of $a(I)$ and $b(D)$ and nothing else: under $I \perp D$ with continuous sources the two arm lines have product measure zero, so the law of $T = f(I, D)$ is unconstrained beyond the support. For (i), $c \equiv 0$ gives additivity, while $c = -b$ off the

Table 9: **Source equivalence on the S&N 1B arms.** Per-task ratio $\sigma_{\text{init}}/\sigma_{\text{order}}$ — below 1 means the order arm is the noisier one (e.g. mbpp: $0.0088/0.0133 = 0.66$) — with F and bootstrap CIs (all 95%), for 9 primary-like and 18 bpb tasks, plus the F interval at the Bonferroni level ($\alpha = 0.05/27$; conservative here, since 8 task identities appear in both panels). Intervals are computed from the unrounded ratios. No task's analytic F interval fits inside the $2\times$ equivalence band ($0/27$; under the bootstrap intervals $2/27$ fit: winogrande primary and minerva_math_geometry bpb); the two tasks whose uncorrected F CIs exclude 1 do so in opposite directions, and neither survives multiplicity control (hellaswag's bootstrap CI includes 1 as well).

task	$\hat{\sigma}_{\text{init}}$	$\hat{\sigma}_{\text{order}}$	ratio	F CI	bootstrap CI	F CI (Bonf.)
primary-like accuracy tasks						
hellaswag	0.00428	0.0124	0.35	[0.17, 0.70]	[0.15, 1.26]	[0.10, 1.13]
socialiqa	0.0123	0.0188	0.65	[0.31, 1.32]	[0.37, 1.19]	[0.19, 2.12]
mmlu	0.00401	0.00543	0.74	[0.35, 1.50]	[0.43, 1.26]	[0.21, 2.40]
arc_challenge	0.0181	0.0241	0.75	[0.36, 1.52]	[0.43, 1.75]	[0.22, 2.45]
csqa	0.0125	0.0159	0.79	[0.38, 1.59]	[0.29, 1.75]	[0.23, 2.56]
winogrande	0.0129	0.0145	0.89	[0.42, 1.79]	[0.51, 1.59]	[0.26, 2.88]
arc_easy	0.00869	0.00852	1.02	[0.49, 2.07]	[0.64, 4.42]	[0.29, 3.32]
piqa	0.00669	0.00641	1.04	[0.50, 2.11]	[0.48, 3.03]	[0.30, 3.39]
copycolors	0.127	0.107	1.18	[0.57, 2.39]	[0.70, 2.24]	[0.34, 3.84]
bits-per-byte tasks						
mbpp	0.00879	0.0133	0.66	[0.32, 1.33]	[0.33, 1.39]	[0.19, 2.14]
socialiqa	0.00817	0.0110	0.74	[0.36, 1.51]	[0.47, 1.26]	[0.21, 2.42]
minerva_math_counting_and_probability	0.00504	0.00585	0.86	[0.41, 1.74]	[0.43, 1.90]	[0.25, 2.80]
humaneval	0.0296	0.0343	0.86	[0.41, 1.75]	[0.40, 1.68]	[0.25, 2.81]
minerva_math_geometry	0.00521	0.00586	0.89	[0.43, 1.80]	[0.52, 1.96]	[0.26, 2.89]
arc_challenge	0.00621	0.00688	0.90	[0.43, 1.83]	[0.47, 2.22]	[0.26, 2.93]
minerva_math_precalculus	0.00565	0.00607	0.93	[0.45, 1.89]	[0.26, 2.67]	[0.27, 3.03]
minerva_math_prealgebra	0.00416	0.00387	1.08	[0.52, 2.18]	[0.36, 1.97]	[0.31, 3.50]
piqa	0.00418	0.00362	1.16	[0.55, 2.34]	[0.37, 2.52]	[0.33, 3.76]
minerva_math_number_theory	0.00345	0.00292	1.18	[0.57, 2.40]	[0.35, 2.43]	[0.34, 3.85]
csqa	0.0133	0.0109	1.22	[0.58, 2.47]	[0.58, 3.97]	[0.35, 3.96]
copycolors	0.0786	0.0632	1.24	[0.60, 2.52]	[0.54, 2.35]	[0.36, 4.05]
minerva_math_intermediate_algebra	0.00526	0.00410	1.28	[0.61, 2.60]	[0.62, 4.26]	[0.37, 4.17]
gsm8k	0.0123	0.00904	1.36	[0.65, 2.75]	[0.78, 3.01]	[0.39, 4.42]
minerva_math_algebra	0.00484	0.00353	1.37	[0.66, 2.78]	[0.62, 3.41]	[0.40, 4.46]
arc_easy	0.0152	0.0106	1.43	[0.69, 2.91]	[0.67, 4.05]	[0.41, 4.67]
hellaswag	0.00399	0.00213	1.88	[0.90, 3.80]	[1.02, 3.25]	[0.54, 6.10]
winogrande	0.0232	0.00982	2.36	[1.13, 4.79]	[1.49, 4.52]	[0.68, 7.69]

Table 10: **What it would take to decide source equivalence.** Half-width of the 95% F interval on the variance ratio (a pure function of degrees of freedom), the window of point estimates that would land inside the $[0.5, 2]$ band, and the probability of passing when the true ratio is 1. The richest public arms ($n = 10$ vs. 9) cannot pass; $n \geq 31$ per arm is needed for a 92% pass probability.

n per arm	F-interval half-width (variance ratio)	acceptance window for $\hat{r}$	pass prob. if true ratio $= 1$
3	$39.0\times$	none (span $>$ band)	0%
6	$7.15\times$	none	0%
8	$4.99\times$	none	0%
10	$4.03\times$	none	0%
10 vs. 9 (S&N arms)	$4.23\times$	none	0%
11	$3.72\times$	$[0.964, 1.037]$	9.0%
16	$2.86\times$	$[0.846, 1.182]$	47.5%
21	$2.46\times$	$[0.785, 1.274]$	71.3%
31	$2.07\times$	$[0.720, 1.389]$	92.3%

order line gives $T = \mu + a(I)$, i.e. $\sigma^2_{\text{all}} = \sigma^2_i$; the binary example enumerates its four cells directly: $T(0,0) = 0$, $T(1,0) = T(0,1) = 1$ — both arms are the same fair two-point law on $\{0,1\}$, so $\sigma^2_i = \sigma^2_o = 1/4$ — and the unseen corner $T(1,1) = c$ gives $\text{Var} = (2+c^2)/4 - ((2+c)/4)^2$: $c = 0$ yields $\sigma^2_{\text{all}} = 1/4$ (ratio 1, the floor is exact), $c = 2$ yields $1/2$ (ratio 2, the additive value), and $c = 4$ yields $9/4$ (ratio 9) — three joint laws the two arms cannot tell apart. For (ii), with both covariances zero, $\sigma^2_{\text{all}} = \sigma^2_i + \sigma^2_o + \text{Var}(c) \geq \sigma^2_i + \sigma^2_o$, sharp at $c \equiv 0$. □

Which parametrization is in force, and what each estimator measures. The main text and the proof use the reference-cell parametrization: c vanishes on both arm lines, so each arm identifies its main-effect law exactly and $\hat{\sigma}^2_{\text{int}} = \hat{\sigma}^2_{\text{bundled}} - \hat{\sigma}^2_{\text{init}} - \hat{\sigma}^2_{\text{order}}$ (Appendix H) is an interaction contrast: it prices $\text{Var}(c)$ up to the latent–interaction covariances $2\text{Cov}(a,c) + 2\text{Cov}(b,c)$, which the arm-line constraints do not remove; and since each arm's measured variance also carries the run-level noise V_ε, the contrast reads $\text{Var}(c) - V_\varepsilon$ up to those covariance terms — when they vanish, it *understates* the interaction by one run-noise unit (measured directly only on our own 20M grid; the PolyPythias arms do not admit the measurement); Appendix H finds it consistent with zero on all six tasks under either reading. Under the alternative sum-to-zero parametrization each arm's law already carries its interaction slice — the init arm at d_0 has variance $\text{Var}(a(I) + c(I, d_0))$ — and the same non-identifiability conclusion holds, with the "additivity factor" then reading $\sqrt{(\sigma^2_i + \sigma^2_o + \text{the arm-line interaction slices})} \,/\, \sigma_i$-like quantities; the numerical factor in the text ($1.62\times$) is computed on the S&N arms under the reference-cell reading, which is the reading on which the arms identify main effects.

The $0/27$ of Table 9 is thus a precision statement about the richest public arms, not evidence about the ratio: no enlargement of the single-source arms can decide whether an init-only floor is faithful. Source parity ($\sigma_i = \sigma_o$) remains identifiable (that is what Table 10 prices, $n \geq 31$ per arm), but answering it would not settle the floor. What identifies σ_{all} is a design change: a bundled arm, or a crossed init×order design (Dodge et al., 2020), not more single-source runs.

C.4 Two readings of the init-only understatement

A natural repair is to inflate the init-only floor by the factor $\sqrt{\sigma^2_{\text{init}} + \sigma^2_{\text{order}}}/\sigma_{\text{init}}$, which on these arms has median 1.62 (IQR 1.40–1.68) on primary tasks and 1.32 (IQR 1.25–1.50) on bpb. This number is a **model-based sensitivity estimate** that assumes the two sources add independently with no interaction, not a measurement of all-source noise, which these arms do not contain. Our own empirical cross-check lands lower: DataDecide's 1B bundled-seed noise (which varies both sources jointly, on a different corpus and architecture) sits at a median $1.06\times$ the S&N init-only σ ($1.19\times$ after correcting the 3-seed SD's c_4 bias), though its own 95% interval spans $\sim 6\times$ at df $= 2$. The point estimates do not establish per-task dominance after multiplicity control (Section 4); what removes the single-source floor as a task-level safe surrogate for all-source noise is structural (Remark 1). Whether such a floor understates or overstates the all-source noise on any given task is not identifiable from these arms at any arm size (Remark 1); treating it as a precise community threshold is therefore unfalsifiable as stated, not merely imprecise.

Table 11: **Two readings of the init-only understatement.** Model-based sensitivity factor $\sqrt{\sigma^2_i + \sigma^2_o}/\sigma_i$ (median 1.62 primary / 1.32 bpb, assumes independent additive sources) against the empirical cross-check, DataDecide-1B bundled-seed $\hat{\sigma}$ over S&N σ_{init} (median 1.06; 1.19 after c_4 correction, on df $= 2$). We report the 1.62 factor only as a sensitivity result under the additivity assumption; the empirical check is compatible with it at current power but cannot confirm it.

quantity	median	spread
Model-based sensitivity factor $\sqrt{\sigma^2_i + \sigma^2_o}/\sigma_i$ (S&N arms, per task)		
primary-like tasks (9)	1.62	IQR 1.40–1.68
bits-per-byte tasks (18)	1.32	IQR 1.25–1.50
Empirical cross-check: DataDecide-1B bundled-seed $\hat{\sigma}$ / S&N σ_{init} (8 shared tasks)		
as measured	1.06	IQR 0.92–1.19
after c_4 bias correction	1.19	df $= 2$; own 95% interval spans $\sim 6\times$

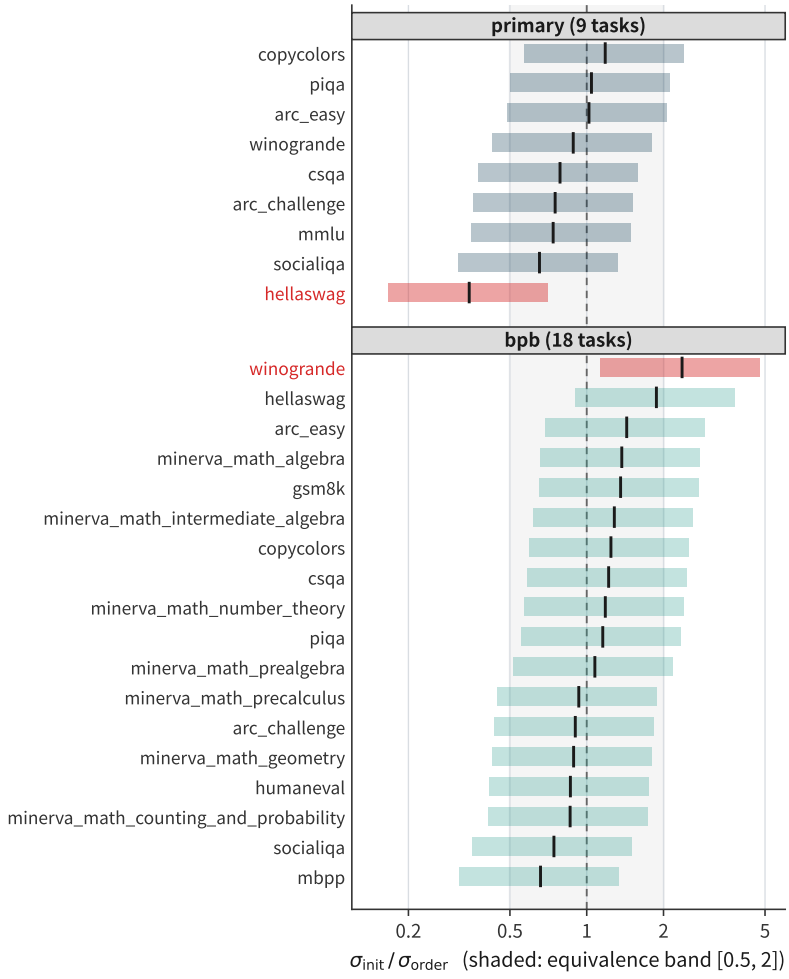

Figure 4: **Source equivalence, task by task.** Per-task $\sigma_{\text{init}}/\sigma_{\text{order}}$ with CIs vs. the $[0.5, 2]$ band (shaded): no CI fits inside; the two tasks excluding 1 do so in opposite directions, and neither survives Bonferroni control (Table 9).

D Decision Replay: Full Tables and Model Specification

D.1 Full decision-replay table

Replay protocol: target = each recipe's 1B score (3-seed mean of final common checkpoints); prediction = proxy-scale final scores, under both 3-seed means and single seeds; decisions = all $\binom{25}{2} = 300$ unordered pairwise comparisons, per task and per scale.

Averaging over 3 seeds instead of one removes 8–24% of errors on 90M–750M (12.4% at 150M). Below 16M the returns are mixed — negative at 4M/6M/10M, positive at 8M/14M: at 4M the 3-seed-mean accuracy (31.0%) sits below the single-seed accuracy (37.8%), and both are below the 50% coin flip. The mechanism is the rank inversion analyzed in Section 5: at 4M the macro-average proxy ranking is negatively correlated with the 1B ranking, and suppressing seed noise does not rescue the aggregated ranking. Per-task 4M accuracies sit on both sides of chance — arc_easy 90.7%, piqa 71.3%, csqa 67.0% above; hellaswag 34.4%, winogrande 38.3%, boolq 42.3% below — so the inversion is broad-based, not carried by one task with residual signal.

D.2 The hierarchical measurement-error model (HME)

Per task and proxy scale, the 25 recipe means are bivariate-normal random effects (latent proxy score μ_i and latent 1B score ν_i with correlation ρ), observed through heteroscedastic per-recipe

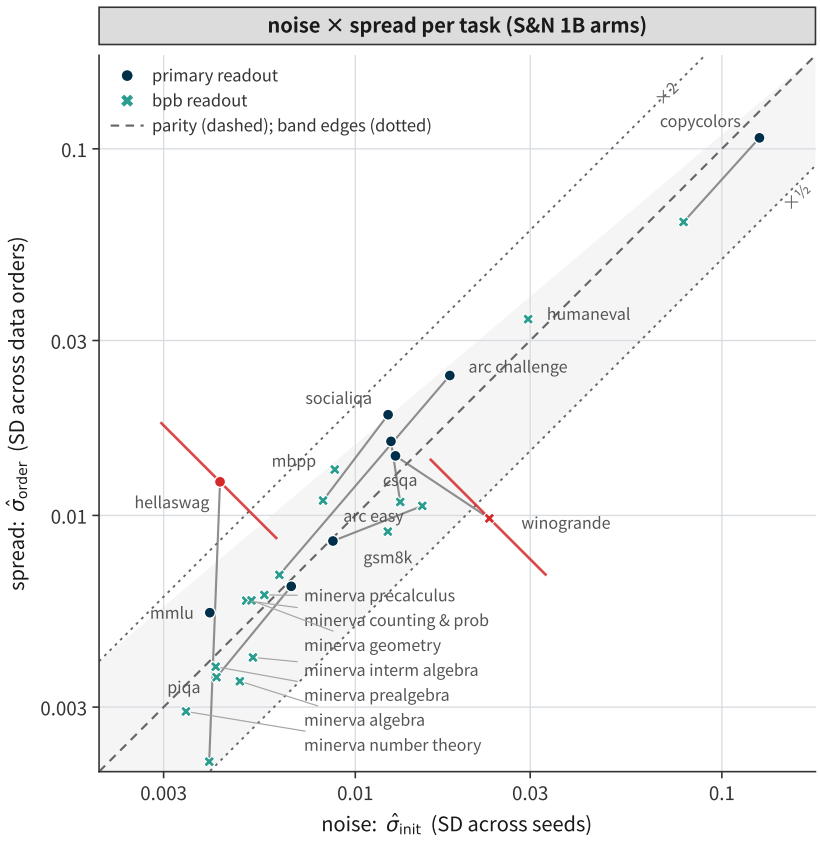

Figure 5: **The source-equivalence plane: noise against spread, task by task.** Component view of Figure 4: per-task $\hat{\sigma}_{\text{init}}$ (SD across seeds) against $\hat{\sigma}_{\text{order}}$ (SD across data orders) on the S&N 1B arms, log–log with square decades, so parity is the 45° dashed diagonal and the $[0.5, 2]$ equivalence band (shaded, dotted edges) is centred on it. Circles: primary readout; crosses: bpb readout; grey segments join the eight task identities measured under both. Point estimates mostly sit inside the band, while no 95% interval fits inside it (Figure 4); only the two excluders carry their F interval here (vermillion), each clearing parity in the opposite direction. Note the readout dependence: `hellaswag` and `winogrande` each cross a band edge when the readout changes (the former entering, the latter leaving), and the `minerva_math` family occupies the low-noise low-spread corner — the geometry behind the per-task ratios of Table 9.

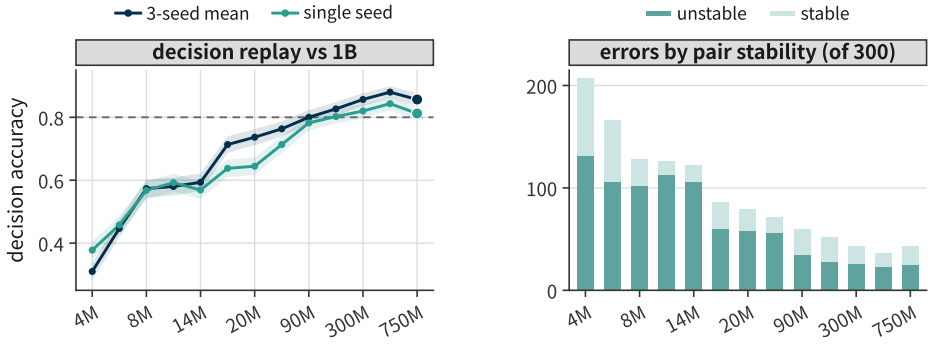

Figure 6: **Decision replay.** Left: pairwise accuracy by proxy scale, 3-seed means vs. single seeds. Right: errors on seed-unstable vs. stable pairs, by scale.

Table 12: **Decision replay.** Pairwise recipe-decision accuracy by proxy scale (accuracies in %), with the share of errors that land on seed-unstable comparisons (descriptive; the classifier judges single-seed decisions, so for the 3-seed-mean instrument the truly flippable share is lower; Table 14). The published $\sim$80% reproduces exactly at 150M (80.2% single-seed, their protocol; 82.7% with 3-seed means); 54% of macro errors at 150M (59% pooled over 90M–750M, and 58.6% excluding the budget-truncated 750M row) and 82% of task-level errors (10 benchmarks pooled over 90M–750M; 81.4% excluding 750M) land on seed-unstable pairs. Under the fitted model, 29% [16, 47] of the instrument's residual errors sit on pairs whose re-draw flip probability exceeds 25%, 13.5% [8.3, 57.0] of expected errors are attributable to seed noise, and the remainder at the point estimate is latent proxy–target discordance, though the interval does not exclude a majority seed-noise share (Table 14).

	macro average (300 pairs)				task level (10 benchmarks)		
size	3-seed	1-seed	err.	% unst.	3-seed	1-seed	% unst.
4M	31.0	37.8	207	63%	54.9	54.2	80%
6M	44.7	45.8	166	64%	59.2	56.2	78%
8M	57.3	56.8	128	80%	64.7	60.9	80%
10M	58.0	59.2	126	90%	64.0	60.3	82%
14M	59.3	56.9	122	87%	66.9	62.1	82%
16M	71.3	63.8	86	70%	66.1	63.4	79%
20M	73.7	64.4	79	73%	64.4	62.4	77%
60M	76.3	71.3	71	79%	69.0	66.9	78%
90M	80.0	78.2	60	58%	73.9	70.5	80%
150M	82.7	80.2	52	54%	75.3	72.1	76%
300M	85.7	82.0	43	60%	82.3	78.6	86%
530M	88.0	84.3	36	64%	85.7	81.7	86%
750M	85.7	81.2	43	58%	82.8	78.6	87%

Table 13: **Realized top-k regret by proxy scale (descriptive replay).** The macro-average argmax pick's forfeited 1B points against the best 1B score on the observed panel (seed-resampled 95% intervals), the mean forfeiture over the three top-ranked picks (which can exceed the top-1 loss when the runner-up picks are worse), and the realized pick's rank among the 25 recipes at 1B. The loss is non-monotone in scale: the realized top-1 loss at the recommended 150M scale nearly matches 4M's, and the smallest losses sit at 300M/530M on this observed panel. †750M is budget-truncated (41% of budget; Appendix A). (Source: `topk_regret.json`.)

scale	top-1 loss [95%]	top-3 loss	pick's 1B rank
4M	5.9 [1.6, 7.2]	5.1	24/25
10M	1.1 [0.0, 7.4]	0.4	7/25
20M	1.1 [0.0, 3.9]	1.0	7/25
60M	1.9 [0.2, 2.4]	0.8	11/25
90M	1.9 [1.6, 2.4]	2.4	11/25
150M	4.7 [1.9, 5.3]	2.2	21/25
300M	0.6 [0.1, 2.1]	0.9	6/25
530M	0.6 [0.3, 2.3]	0.9	6/25
750M†	1.9 [0.3, 2.4]	0.9	11/25

seed noise that is partially pooled across recipes; the dependence among the 300 pairs that share 25 recipes is thereby structural in the model rather than ignored. Fit by empirical Bayes, the model reproduces the replay itself: 52.4 expected decision errors at 150M macro versus 52 observed. The realized argmax pick at 150M — the usage DataDecide recommends — is `DCLM-Baseline (QC FW 3%)`, ranking 21st of 25 at 1B (the top-3 picks rank 21st/11th/3rd). And, under parametric-bootstrap replication, it regenerates the descriptive statistic (simulated error$\cap$unstable share 61% [42, 81] versus 54% observed).

Estimand dictionary (per pair $i<j$, per posterior draw). Write $d = \mathrm{sign}(\bar{x}_i - \bar{x}_j)$ for the observed proxy decision, $d_\nu = \mathrm{sign}(\nu_i - \nu_j)$ for the latent-target ordering, and d_μ likewise for the latent proxy. Then: the *descriptive* 54% is the share of observed 150M macro errors whose decision flips within the released three seeds (no model); the *flip probability* (mean 7.2% over all pairs) is $P(\text{decision flips under an independent 3-seed re-draw}) = \Phi(-d \cdot dX/\tau)$, with dX the pair's proxy-score gap and $\tau = \sqrt{(\sigma_i^2 + \sigma_j^2)/3}$ its 3-seed re-draw noise scale; the *truly flippable share* (29% [16, 47]) is, among pairs whose observed decision is wrong, the share whose true re-draw flip probability exceeds 25%; and the *attribution* (13.5% [8.3, 57.0]) is $\sum P(\mathrm{err} \wedge d_\nu = d_\mu) / \sum P(\mathrm{err})$ — errors

whose latent orderings agree, so only noise can flip them. Net versus gross: the model's expected errors under zero seed noise are the latent-discord pairs (53.2 at 150M), *above* the noisy 52.4 — seed noise is net-neutral on the error count here because the realized pick sits low among its neighbours and re-draws escape it; the 13.5% is a gross noise-caused share, not a net one.

Specification. Per decision cell (task × proxy scale; the 25 recipes are the candidates within, distinct from Section 2's per-recipe cell): per recipe i and seed $r \in \{1, 2, 3\}$, X_{ir} is the proxy-scale score at the final common checkpoint and Y_{ir} the 1B score — a per-seed mean of the last three common checkpoints, a lightly smoothed variant of the replay's final-checkpoint target (the model needs per-seed targets with resolved noise; the smoothing changes the target's own noise level, which the model marginalizes anyway). Each matrix is decomposed two-way, $M_{ir} = a_i + g_r + \varepsilon_{ir}$ (recipe effect, seed effect, residual), and the per-recipe noise sufficiency $s_i^2 = \sum_r \varepsilon_{ir}^2/2$ (df = 2) feeds the noise posterior. *Noise prior (partial pooling):* $u_i = \log \sigma_i^2 \sim \mathcal{N}(2\lambda, (2\zeta)^2)$ with empirical-Bayes moments from the sampling law $s_i^2 \sim \sigma_i^2 \chi_2^2/2$ — exact for known seed effects; estimating them from the same matrix biases the sufficiency by the two-way projection. Under homoscedasticity the factor is $(R-1)/R = 0.96$ with the exponential shape unchanged (simulated); per recipe it is $(1-2/R)\sigma_i^2 + R^{-2}\sum_j \sigma_j^2$ (`hme_full_correction.json`). The printed 13.5% [8.3, 57.0] remains the primary analysis; the others are sensitivity variants. Refitting the 150M cell with the per-recipe correction moves the seed-noise attribution from its printed 13.5% to 13.9%, refitting on the replay's unsmoothed final-checkpoint target gives 13.8%, and both corrections together give 14.2% — all far inside its interval (also `hme_projection.json`): $2\lambda = \overline{\log s^2} + \gamma$ (Euler–Mascheroni) and $(2\zeta)^2 = \max(\mathrm{Var}\log s^2 - \pi^2/6,\ 0)$, floored at 0.02 — so the pooling strength is estimated from the spread of the per-recipe variances themselves, not set by hand. *Noise posterior:* a 400-point grid over u with log-likelihood $-u - s^2 e^{-u}$. *Latents:* (μ_i, ν_i) bivariate normal; the latent variances are moment-deconvolved, $\omega_X^2 = \max(\mathrm{Var}(\bar{x}) - \overline{\mathbb{E}[\sigma_X^2 \mid s^2]}/3,\ 0)$ (likewise for Y), and ρ from $\mathrm{cov}(\bar{x}, \bar{y})/(\omega_X \omega_Y)$, clipped to $[-1, 1]$; at 150M macro $\hat{\rho} = 0.79$ with recipe-bootstrap 95% CI $[0.46, 0.99]$. The 1B target carries its own per-recipe noise posterior through the same construction, so target-side seed noise is marginalized in every estimand. All estimands (flip probabilities, the error decomposition, pick losses) are Monte-Carlo integrals over posterior draws of $(\mu, \nu, \sigma_X^2, \sigma_Y^2)$; no MCMC. The noise share's interval is asymmetric because its numerator concentrates in a few recipes: recipe-cluster resamples that drop them move the ratio substantially, whence the right tail to 57%.

Table 14: **The hierarchical measurement-error model, by proxy scale (macro average).** Fit on the smoothed replay target (its pick losses differ from Table 13's realized-replay values by ∼0.1 point). Observed decision errors against the fitted model's posterior expectation (recipe-cluster bootstrap 95% CI); the model's attribution of expected errors to seed noise (the remainder is latent proxy–target discordance); the mean probability that an independent three-seed re-draw flips a 3-seed-mean pairwise decision; the excess-risk share (expected forfeited target signal over total pairwise latent signal); and the pick loss, split into the *realized* shortfall of the argmax-at-proxy recipe against the latent best at 1B on the observed panel (descriptive; no interval) and the expected pick loss under an independent three-seed *re-draw* of the instrument, from a parametric simulation that holds the 25 fitted recipe latents fixed and resamples only seed noise ($R = 20{,}000$; degenerate at 90M, where the pick moves only 2% of the time). The re-draw mean can sit *below* the noiseless-pick loss (0.039 at 150M — a loss, not a floor: zero-noise picking still pays the latent proxy–target discordance, and a re-drawn instrument can escape onto better neighbours): the realized top pick is a below-median one (21st of 25 at 1B), and seed noise lets a re-drawn instrument escape onto neighbouring recipes that are better at 1B (at 150M the pick moves 24% of the time), the mechanism of Section 5 in miniature. The 4M pick loss (5.9 macro points [1.6, 7.2]) is computed directly on the replay — the HME is fit at every scale; the table reports 90M up, where the deconvolved macro SNR clears the noise-limited regime (> 1.3; Table 15) — below it the decomposition has no signal to attribute. All attribution statements are under the fitted model and carry their intervals.

proxy	obs. errors	model errors [CI]	noise share [CI]	mean flip [CI]	excess risk [CI]	pick loss (realized)	pick loss (re-draw) [CI]
90M	60	58 [31, 97]	18% [11, 65]	9.2% [7.1, 14.4]	11.5% [2.5, 26.4]	0.019	0.017 [0.017, 0.017]
150M	52	52 [25, 101]	13.5% [8.3, 57]	7.2% [6.3, 12.7]	10.9% [1.9, 30.2]	0.047	0.036 [0.002, 0.044]
300M	43	42 [20, 76]	16% [11, 59]	7.2% [6.3, 12.9]	6.7% [0.9, 17.9]	0.007	0.005 [0.002, 0.017]
530M	36	33 [21, 62]	25% [18, 64]	7.8% [7.2, 12.7]	3.6% [1.0, 9.0]	0.006	0.005 [0.000, 0.016]
750M	43	41 [22, 77]	22% [13, 69]	8.8% [7.7, 14.0]	6.8% [1.1, 18.7]	0.018	0.008 [0.000, 0.017]

The model also re-reads the descriptive share itself: in simulation, the 3-seed-agreement classifier's 54% share of errors corresponds to a true independent-redraw flippable share of 29% [16, 47]. A

pair with per-seed flip probability 0.25 is flagged unstable 56% of the time, but its 3-seed-mean decision flips only 12%. For the 3-seed instrument the descriptive share *over*-estimates the flippable share of errors, the opposite of the lower-bound reading that per-seed reasoning suggests, while the intersection reading simultaneously obscures how much of the residual error no re-draw can repair.

D.3 SNR DECONVOLUTION: ESTIMATOR, DIAGNOSTICS, AND INTERVALS

Section 2 defines the deconvolved signal $\sqrt{\max(S^2 - \overline{\hat{\sigma}^2}/3,\, 0)}$, whose SNR (Table 15) uses the RMS noise $\sqrt{\overline{\hat{\sigma}^2}}$; $\hat{\sigma}_i^2$ is unbiased on the variance scale, so no small-sample correction enters either term. Table 15 reports, per scale: the heterogeneity diagnostic $\overline{\hat{\sigma}^2}/\mathrm{median}(\hat{\sigma})^2$; the deconvolved SNR with a recipe-cluster bootstrap 95% interval ($B = 2{,}000$); the untruncated difference $S^2 - \overline{\hat{\sigma}^2}/3$ with its interval (at the boundary the percentile interval of the truncated ratio is uninformative, and the signed difference is the informative quantity); a trimmed variant dropping the two highest-$\hat{\sigma}$ recipes; and a median-based variant, the deconvolved signal over $\mathrm{median}(\hat{\sigma})/\sqrt{\ln 2}$ (the median-unbiased correction at $n = 3$), which agrees except where the heterogeneity diagnostic is high (750M: 2.52 vs 4.27; 1B: 3.52 vs 4.75 — the variance-scale heterogeneity ratio is $4.14/2.64$ there, and the two variants diverge exactly as that ratio predicts). Resampling 25 recipes with replacement biases the bootstrap S^{*2} down by a factor $(n-1)/n$, so the intervals are slightly conservative for the signal term. The paired bootstrap of the 90M/60M SNR ratio (the same 25 recipes resampled jointly) gives 2.12 $[1.60, 2.97]$, with no replicate ≤ 1. At 10M the estimate truncates at zero and stays there under trimming: the two highest-$\hat{\sigma}$ recipes are Falcon and Falcon+CC ($\hat{\sigma} = 0.0135$ each; third is Dolma1.6++, 0.0134). The four scales whose untruncated intervals cross zero are 4M, 8M, 10M, and 14M. For reference, the highest-noise recipes at the other boundary scales: at 4M, Falcon+CC (QC Tulu 10%) (0.0155), C4 (0.0139), Dolma1.7 (no Flan) (0.0137); at 8M, Dolma1.7 (no math, code) (0.0148), Falcon+CC (QC 10%) (0.0118), Falcon (0.0105); at 14M, DCLM-Baseline (0.0181), Falcon+CC (QC Orig 10%) (0.0135), DCLM-Baseline (QC 20%) (0.0133). The 60M→90M jump survives leave-one-recipe-out at the macro level: 60M SNR $\in [1.13, 1.35]$ and 90M $\in [2.38, 2.81]$ over all 25 single-recipe drops, and the jump appears in 6 of 10 task-level readouts. The jump decomposes as $\Delta \log \mathrm{SNR} = 0.752 = 0.466$ (signal rise) $+ 0.286$ (noise drop, RMS basis), and survives on the median-noise variant as well ($1.12 \to 2.92$); the elevated 60M noise is carried by a few recipe outliers (DCLM-Baseline (QC 20%) reads $8.6\times$ its 20M/90M neighbourhood), not by the checkpoint grid (median accuracy-grid density 23 at 60M vs 13 at 20M). On the bpb readout the SNR reads $2.36/4.49/8.26$ at 4M/60M/90M (Appendix G). At 10M, simulated datasets with zero latent signal score 50% on average (95% range $[37\%, 64\%]$), containing the observed 58.0%; read per task, the same 10M models decide 64.0% of pairs correctly. The z approximation understates the Welch-band shares: 47–89% vs the Welch 83–96% at 4–20M.

Shared-seed effects are negligible here. DataDecide's three seeds are the same seed labels on every recipe, so a shared seed-level component could in principle cancel in pairwise differences (the Welch band's $\hat{\sigma}_i^2 + \hat{\sigma}_j^2$ would then overestimate the pairwise noise by ignoring the $2\,\mathrm{cov}$ term). Measuring it: a two-way decomposition of each (task × size) cell into recipe + seed + residual puts the shared seed component at 0.6–5.5% of the mean residual variance across scales, i.e. the paired correction would narrow the bands by at most 5.5% (and widen nothing); all band-based conclusions are unchanged. The model above already removes seed effects before fitting latents, so its estimands never see the shared component.

Reference implementation. Three lines on a per-candidate × per-seed score matrix reproduce the deconvolved SNR and Welch-coverage columns of Table 15 to the digit (checked at all 14 scales; script archived with the code release):

```
mu, sd = scores.mean(1), scores.std(1, ddof=1)
snr = sqrt(max(mu.var(ddof=1) - (sd**2).mean()/n_seeds, 0)) / sqrt((sd**2).mean())
covered(i,j) = abs(mu[i]-mu[j]) < t(0.975, nu_ij)*sqrt(sd[i]**2/n_seeds + sd[j]**2/n_
```

The groups are the connected components of the covered graph (pairs linked by un-rejected comparisons; non-adjacent members of a group can still differ significantly). On this panel's macro readout the covered graph is connected at every scale (the conservative merge degenerates to a single group);

Table 15: **Deconvolved SNR: diagnostics and intervals.** Per scale: the heterogeneity diagnostic $\overline{\hat{\sigma}^2}/\mathrm{median}(\hat{\sigma})^2$; the deconvolved SNR of Section 5 with a recipe-cluster bootstrap 95% interval; the untruncated signal difference $S^2 - \overline{\hat{\sigma}^2}/3$ (units of 10^{-5}) with its interval; the trimmed variant (two highest-$\hat{\sigma}$ recipes dropped); and the median-based variant; the last two columns are the share of the 300 pairs inside their pair-specific 95% Welch bands, and the across-recipe RMS within-recipe SD (the deconvolved SNR's denominator).

scale	het. ratio	SNR (obs.)	SNR (deconv.)	95% CI	untrunc. diff. [CI]	trimmed	median variant	inside Welch	noise (RMS)
4M	1.08	0.67	0.30	[0.00, 0.68]	+0.7 [−1.5, +2.7]	0.43	0.26	93.7%	0.00909
6M	2.72	1.41	0.63	[0.17, 1.03]	+1.8 [+0.2, +3.3]	0.79	0.87	83.7%	0.00669
8M	2.65	1.08	0.33	[0.00, 0.70]	+0.5 [−0.8, +1.7]	0.46	0.45	95.7%	0.00661
10M	4.03	1.00	0.00	[0.00, 0.43]	−0.5 [−1.9, +0.8]	0.00	0.00	93.0%	0.00774
14M	1.40	0.77	0.29	[0.00, 0.65]	+0.6 [−1.3, +2.4]	0.29	0.29	95.7%	0.00838
16M	3.32	1.68	0.72	[0.25, 1.12]	+2.8 [+0.4, +5.0]	0.90	1.09	83.0%	0.00740
20M	1.23	1.07	0.77	[0.48, 1.06]	+3.1 [+1.3, +4.7]	0.85	0.71	86.7%	0.00719
60M	1.21	1.49	1.23	[0.82, 1.52]	+10.7 [+5.2, +15.0]	1.44	1.12	72.7%	0.00841
90M	1.81	3.60	2.61	[1.88, 3.42]	+27.1 [+14.8, +37.9]	2.94	2.92	47.0%	0.00632
150M	1.29	3.67	3.18	[2.29, 4.20]	+40.1 [+24.1, +54.4]	3.65	3.01	38.7%	0.00630
300M	1.28	4.09	3.57	[2.54, 4.48]	+51.4 [+28.4, +68.4]	3.84	3.36	36.7%	0.00636
530M	1.27	3.42	2.98	[2.16, 3.99]	+53.1 [+31.8, +70.9]	3.44	2.80	39.3%	0.00774
750M	4.14	5.26	2.52	[1.79, 3.46]	+45.2 [+27.9, +59.3]	3.20	4.27	46.7%	0.00843
1B	2.64	5.79	3.52	[2.40, 5.07]	+41.2 [+22.5, +59.4]	4.47	4.75	33.0%	0.00577

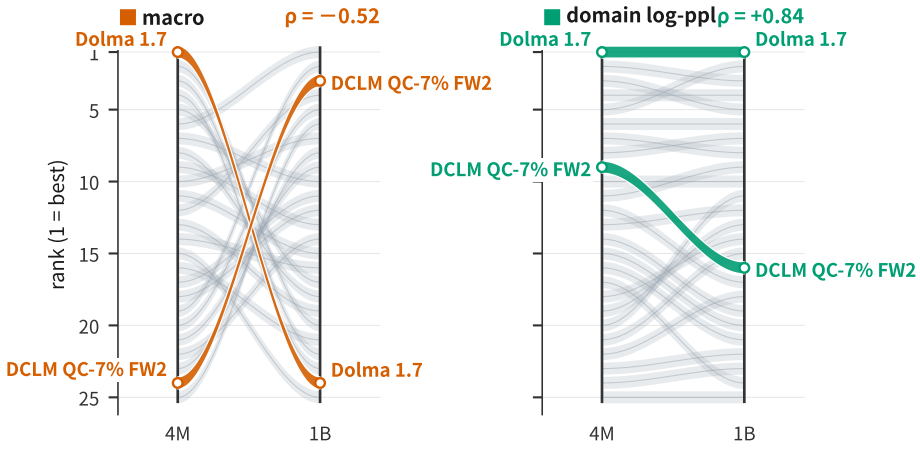

Figure 7: **The 4M inversion, recipe by recipe.** Each band is one recipe's rank path from 4M to 1B (25 recipes). Left (macro-accuracy readout): the two extreme crossers are highlighted — Dolma 1.7 goes 1→24, DCLM QC-7% FW2 goes 24→3 — and the fabric crosses throughout (Spearman $\rho = -0.52$ on ranks; the headline Pearson $r = -0.56$ is on scores). Right (domain-level log-perplexity readout): the same recipes hold their ranks ($\rho = +0.84$ on ranks; $r = +0.87$ on scores).

the grouping is informative on per-task readouts and stricter bands. At 1B, for example, the task-level readouts split into six blocks on `mmlu`, five on `hellaswag`, three on `arc_easy`, and two on `piqa`.

D.4 Robustness details for the Section 5 claims

Inversion: per-task and readout detail. Per-task 4M–1B correlations range from −0.27 (`winogrande`) to +0.88 (`arc_easy`); per-task 4M accuracies sit on both sides of chance (34–91%). The recipe-level rank paths are in Figure 7. The macro inversion stays non-positive under median, z-score, rank, and min-max readouts (−0.36 to −0.55), negative under 9 of 10 leave-one-task-out drops with the mean readout (10 of 10 for every other readout; worst case +0.11), and under a task-level bpb proxy at 4M (S&N's re-evaluation; single-seed 1B accuracy target) reads −0.33 (leave-one-recipe-out $[-0.39, -0.28]$); the headline Pearson's Spearman twin is −0.52, leave-one-recipe-out $[-0.67, -0.50]$. The 4M accuracy gap, stress-tested: the same gap computed on unlinked recipe draws (independent resamples at 4M and 1B) is centered at zero (median 0.0pp, 50% of resamples negative, 95% range ±7pp), so the observed −6.8pp sits at the null's edge — the gap is not separable from recipe-set churn.

Repetition is nearly absent from this asset. DataDecide trains each run at 100 tokens per parameter, so a recipe's pool repeats exactly when the pool is smaller than the scale's token budget. Summing the released tokenized pools (recipe→pool mapping from the authors' OLMo branch; sizes from the released shard inventory, verified against the paper's stated Dolma/DCLM pool sizes), exactly three pools fall below the 1B budget (100B tokens): DCLM-Baseline (QC 10%) at 1.22 epochs, FineWeb-Pro at 1.21, and DCLM-Baseline (QC FW 10%) at 1.02; nothing repeats at or below 750M. Excluding the three repeating recipes moves the 4M macro correlation from -0.56 to -0.53. The Zhou et al. (2026) mechanism is therefore not the driver here.

Chance-task ablation. Selecting informative tasks without looking at the target: dropping the three tasks whose best 4M recipe is within Welch noise of the task's own chance rate (`arc_challenge`, `mmlu`, `openbookqa`) leaves an attenuated -0.33 [-0.54, 0.16] — the inversion attenuates but persists on the remaining seven tasks, its interval covering zero at this reduced seven-task evidence. A targeted variant inverts the sign: dropping the seven tasks whose 4M pairwise accuracy *against the 1B target* is indistinguishable from chance (recipe-cluster bootstrap interval covers 50%) leaves `arc_easy`, `csqa`, and `piqa` and reads $+0.56$ [0.14, 0.70] — but that selection uses the target, so the reversal is partly built in; we report it only as a bound on the aggregation effect, not as independent evidence. The inversion is thus a property of the full macro readout's task weighting, not of every informative task. Nor is it an artifact of the 1B target choice: the 4M macro ranking correlates at -0.52 to -0.68 with *every* larger scale's ranking (60M through 1B; `inversion_targets.json`).

The seed-budget price of "measure the seeds", and the bpb-readout contrast. Holding each recipe's measured 3-seed $\hat{\sigma}$ fixed, the decidable share under n seeds is closed-form (Welch bands shrink as $1/\sqrt{n}$): on this panel's macro readout the share leaves zero at $n = 4$ (60M) and reaches $\geq$50% by $n = 8$ at 60M, while at 4M even $n = 30$ buys only 56% — at 4M the missing ingredient is signal, not seeds (`seed_budget_curve.json`). Which readout inverts depends on the evaluation *target*, not the metric family: on DataDecide's 11-domain log-perplexity readout (3 seeds, all 14 scales), the 4M macro ranking correlates with the 1B ranking at $r = +0.87$ and 4M pairwise accuracy is 85.3% (`bpb_replay.json`), while on *task-level* bpb (the S&N re-evaluation) it still inverts ($r = -0.33$, pairwise accuracy 38.7%) despite a separability SNR of 2.36 at 4M (Table 20); the failure mode lives on the downstream-task readouts, whichever the metric — and the task-bpb instance is the clearest form of the rule-3 caveat of Section 7.

Dependence-robust multiplicity for the decidable share. The 300 pair-specific Welch tests of Section 7 share 25 recipe-level variance estimates, so BH control is optimistic; Table 16 compares it with Benjamini–Yekutieli (valid under arbitrary dependence) and a parametric max-T bootstrap (recipes' seed values drawn as $\mathcal{N}(0, \hat{\sigma}_i^2)$ under the null of equal means, $B = 2{,}000$; a raw with-replacement variant is degenerate at $n = 3$, with a $1/9$ zero-variance probability per recipe). The share at $\leq$60M is essentially zero under every rule at every scale (the only exceptions: 1.3% at 6M under BH and one pair at 16M under max-T; Table 16 now lists all 14 scales). The variance model matters more than the multiplicity rule at the boundary: pooling the variance across recipes (df = 50; defensible where the heterogeneity diagnostic is near one — 1.08–1.23 at 4M/20M/60M — indefensible at 6M/8M/10M/16M (2.7–4.0; 14M sits at 1.40)) leaves the 4M/10M share at zero, with 4–7% at 16M/20M under pooled BH only, but reads 34%/19% at 60M under BH/BY (Table 16): the 60M abstention is a property of the pair-specific test, the $\leq$20M abstention is robust to the variance model. And where pairs are decidable *under BH* (the generous member of the three rules), deciding only within them is strictly more accurate on this panel (macro readout: 88.3% vs 80.0% at 90M; 88.2% vs 82.7% at 150M; 95.8% vs 85.7% at 300M; 94.6% vs 88.0% at 530M — `decidable_accuracy.json`). Decidable pairs are the far-apart pairs by construction, so their accuracy reflects the selection effect, not an added benefit; the substantive result is the abstention below 60M.

Robustness ledger. Each interval's construction, with its archived artifact (all in `verify-scripts/`): the 4M inversion is re-estimated under a family-cluster bootstrap (10 recipe families, resampling whole families, $B = 20{,}000$; `family_bootstrap.json`) giving $[-0.72, -0.08]$ with $P(r{>}0) \approx 0.01$; under a wild-cluster (Rademacher) bootstrap on the same families (`extra_robustness.json`) giving $[-0.58, -0.34]$; under a permutation

Table 16: **Decidable share of the 300 macro pairs under three multiplicity rules and two variance models, all 14 scales.** Left: pair-specific Welch variances under BH, Benjamini–Yekutieli, and a parametric max-T bootstrap. Right: the same with the variance pooled across recipes ($s_p^2 = \overline{\hat{\sigma}^2}$, df $= 50$) — defensible only where the heterogeneity diagnostic is near one (1.08–1.23 at 4M/20M/60M; Table 15), and shown as a sensitivity bound: the 4M/10M zero is invariant to the variance model (16M/20M read 4/7% under pooled BH, 0 under BY), while the 60M share is not (0% pair-specific; 34%/19% pooled under BH/BY). †750M is budget-truncated (41% of budget; Appendix A).

	pair-specific Welch			pooled variance	
scale	BH	BY	max-T	BH	BY
4M	0.00	0.00	0.00	0.00	0.00
6M	0.013	0.00	0.00	0.00	0.00
8M	0.00	0.00	0.00	0.00	0.00
10M	0.00	0.00	0.00	0.00	0.00
14M	0.00	0.00	0.00	0.00	0.00
16M	0.00	0.00	0.003	0.04	0.00
20M	0.00	0.00	0.00	0.07	0.00
60M	0.00	0.00	0.00	0.34	0.19
90M	0.37	0.04	0.04	0.65	0.56
150M	0.51	0.12	0.05	0.73	0.64
300M	0.55	0.01	0.07	0.73	0.66
530M	0.43	0.06	0.05	0.71	0.60
750M†	0.40	0.04	0.04	0.61	0.52
1B	0.62	0.16	0.08	0.76	0.67

of the recipe labels ($[-0.40, 0.39]$, $p = 0.002$; same file); and under a joint family×task×seed bootstrap that prices in the 3-seed estimation noise at both ends (`joint_bootstrap.json`) giving $[-0.68, 0.08]$. The HME attribution (13.5% [8.3, 57.0]) is re-fit under model-ingredient variants (`attribution_sensitivity.json`): a T5-metric target (12.8%), a rank target (15.4%), pooling-strength scan 0.25–4.0 (11.7–18.2%), and leave-one-recipe-out (median 13.1%; 24 of 25 exclusions within 1.5 points, one QC-variant exclusion at 29.3%). The within-scale SNR→accuracy correlations (+0.39/+0.90/+0.92) are in `oos_snr.json`; the split-seed variant (+0.68) in `extra_robustness.json`; the scale-first comparison (90M×1 seed vs 60M×3, +1.9 points) carries a family-cluster bootstrap over the 10 recipe families with target-seed resampling of $[-9.7, +6.7]$ (`scale_first_ci.json`; the interval in Section 7 is this one).

Seed-unstable shares and the classifier's re-read. Pooled over 90M–750M, 59% of macro errors land on seed-unstable pairs; at task level 82%; the base rate of unstable pairs is 26–31% by scale. The HME reproduces the replay itself (52.4 expected errors at 150M macro versus 52 observed) and re-reads the descriptive 54% share as a true flippable share of 29% [16, 47].

The SNR-to-accuracy mapping, in full. Section 7's mapping, all bins: fit on the 80 valid task-level cells of this panel's nine reporting scales (Spearman +0.89), it is a property of *this* candidate set and readout — recompute it on your own (Section 7, rule 3). A split-seed variant (SNR from two seeds, decision accuracy evaluated on the held-out third, rotating) gives Spearman +0.68 ($p < 10^{-18}$, 130 cells; `extra_robustness.json`).

Table 17: **Separability SNR vs pairwise decision accuracy, task-level cells of this panel.** Mean pairwise accuracy per SNR bin.

deconvolved SNR bin	$[0, 0.5)$	$[0.5, 1)$	$[1, 1.5)$	$[1.5, 2.5)$	$[2.5, 4)$	$[4, 10)$
pairwise accuracy	0.52	0.63	0.72	0.79	0.88	0.93

D.5 Head-to-head with the original S&N signal-to-noise readout

Signal & Noise's own SNR pairs the across-recipe spread with their *checkpoint-proxy* noise; ours pairs it with the *measured* 3-seed noise, deconvolved. On the same 130 (task, scale) cells, each variant's per-cell SNR predicts the cell's pairwise decision accuracy against the 1B target: ours at Spearman +0.83 (130 head-to-head cells, all scales; the prescription map's +0.89 pools the 80 valid

cells of the nine reporting scales), theirs at $+0.72$ verbatim — their released range-based signal over their checkpoint-proxy noise — and at $+0.65$ with the signal swapped to the across-recipe SD; against the latter ours leads in 11 of 13 scales (ties/near-ties at 530M and 750M) and the two read-outs correlate $+0.79$ across cells. The increment of this paper's SNR is therefore predictive, not only conceptual: real seed noise, not checkpoint spread, is the quantity that bounds decision accuracy. Decomposing the gain on the same cells (`snr_ablation.json`): against two S&N-family baselines — (i) *verbatim*: signal = the max pairwise difference of recipe means over their mean (their released implementation), noise = their checkpoint proxy, reading $+0.72$; (ii) an intermediate with our across-recipe SD signal over the same proxy denominator, reading $+0.65$ — ours reads $+0.83$ (vs the verbatim baseline, paired bootstrap $+0.12$ [$+0.05$, $+0.19$]). Swapping only the denominator (observed signal over the measured noise) already gives $+0.83$; deconvolving while keeping the proxy denominator gives $+0.74$. Adding the deconvolution on top of the measured denominator leaves the rank predictions unchanged ($+0.83$) — expected, since with a common RMS denominator the deconvolution is a monotone transformation of the per-cell SNR (up to truncation ties): the increment is the measured-noise denominator, and the deconvolution's value is the calibrated magnitude, not rank prediction. (Script: `verify-scripts/sn_headtohead.py`; per-cell values archived with the code release.)

E THE σ ATLAS: FULL TABLE AND OUTLIER DETAIL

For each of 14 sizes $\times$ 25 recipes $\times$ 11 evaluation domains, we compute the 3-seed SD of final log-perplexity (the SD of validation cross-entropy, in nats) at each recipe's common final step, 275 cells per size, 3,850 in all. We release the full per-cell table; Figure 8 maps the medians and the recipe-only spread, and Figure 9 anchors the scale in absolute accuracy units on PolyPythias's ten-seed cells; each cell has $n = 3$ (df $= 2$), so we publish quantiles across cells rather than per-cell intervals.

Gap quantiles. Ranking each scale's 25 recipes by their 3-seed mean final log-perplexity and dividing each rank-adjacent gap by the cell-median seed SD: the median ratio is 1.22/1.33/1.37/1.54/1.88/1.41/0.50 at 4M/6M/8M/10M/14M/16M/20M and 1.02–4.17 at 60M–1B; the lower quartile stays below one at every scale up to 90M.

Table 18: **The σ atlas, per scale.** Quantiles of the 3-seed final log-perplexity SD (nats) across 25 recipes $\times$ 11 domains per size, and the share of cells above two reference thresholds. At 4–20M the median cell carries 0.008–0.029 nat of seed noise; 68–95% of cells exceed 0.005 nat. Each cell has df $= 2$; quantiles are across cells. †grid = median common evaluation checkpoints per cell on the perplexity release; the 20M row's sparser grid (12, vs 16/24) coincides with its elevated noise (mechanism unresolved — the endpoint channel is ruled out since all cells share the final step; Section 6, this appendix), and 750M's grid ends at 41% of budget.

size	grid†	q10	q25	median	q75	q90	> 0.005 nat	> 0.002 nat
4M	5	0.0041	0.0075	0.0160	0.0294	0.0462	84%	97%
6M	6	0.0045	0.0071	0.0160	0.0389	0.0646	87%	99%
8M	9	0.0045	0.0078	0.0124	0.0245	0.0426	87%	98%
10M	8	0.0028	0.0051	0.0108	0.0254	0.0452	75%	95%
14M	15	0.0022	0.0042	0.0080	0.0162	0.0276	68%	93%
16M	16	0.0038	0.0065	0.0111	0.0196	0.0351	82%	97%
20M	12	0.0091	0.0154	0.0288	0.0445	0.0615	95%	99%
60M	24	0.0037	0.0087	0.0125	0.0190	0.0285	85%	97%
90M	23	0.0022	0.0034	0.0063	0.0103	0.0160	60%	92%
150M	30	0.0012	0.0021	0.0043	0.0080	0.0157	44%	76%
300M	35	0.0015	0.0023	0.0045	0.0086	0.0183	48%	80%
530M	40	0.0007	0.0013	0.0025	0.0060	0.0144	29%	64%
750M	22	0.0009	0.0016	0.0030	0.0061	0.0148	33%	67%
1B	6	0.0015	0.0023	0.0038	0.0081	0.0193	41%	80%

Recipe-heterogeneity estimator. We estimate the true dispersion of $\log \sigma$ across the 25 recipes by subtracting the expected variance of $\log \hat{\sigma}$ under homogeneity (0.411 for $n = 3$, from the χ^2_2 sampling law) from the observed variance, and report $\exp(\sqrt{\max(\widehat{\mathrm{Var}} - 0.411, 0)})$ as a spread-per-SD

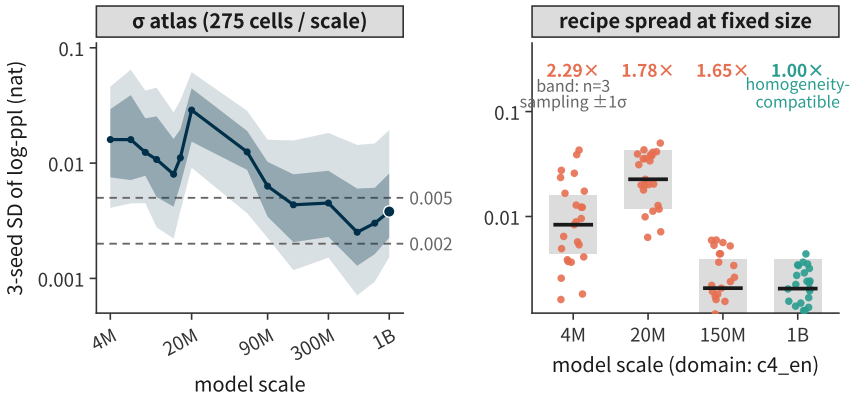

Figure 8: **Noise is a function of the recipe.** Left: the σ atlas — per-cell 3-seed final log-ppl SD (nats) across scales (median with q25–q75 and q10–q90 bands over the 275 cells per scale). Right: recipe-only spread at fixed size (domain c4_en, 25 recipes per size) against the ± 1 SD band expected from $n = 3$ sampling under homogeneity, with the spread factor per size on top; the 1B row lands on the null side.

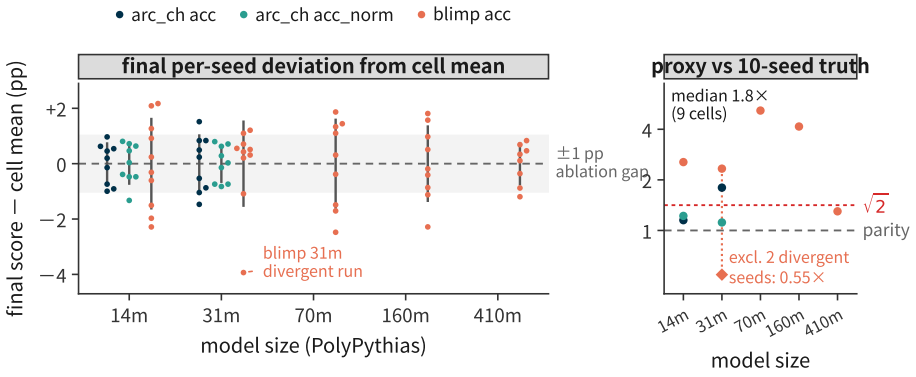

Figure 9: **Ten-seed ground truth on PolyPythias.** (a) Final per-seed deviation from the cell mean, in percentage points (every ring is one run; gray stem: $\pm 1\sigma$). A typical 0.5–1pp ablation gap at 14m–31m (shaded band) lies inside 1σ; the most divergent run (BLIMP 31m, -3.9pp) is labeled (a second seed also diverges — panel b removes both). (b) Checkpoint-proxy noise vs 10-seed truth on the 9 available cells (median $1.8\times$; parity and $\sqrt{2}$ references); removing the two divergent BLIMP 31m seeds moves that cell to $0.55\times$ (open diamond; the 9-cell median becomes $1.3\times$).

factor. Headline numbers use the C4-en domain (25 recipes per size; 24 at 6M). Two uncertainty instruments accompany the point estimates. (i) A recipe-level bootstrap on the spread factor: 95% intervals are $[1.47, 3.06]$ at 4M, $[1.00, 2.76]$ at 20M, $[1.00, 2.10]$ at 150M. (ii) A Monte-Carlo test of the homogeneity null, simulating the χ^2_2 sampling law directly (exact under that law; under a heavier-tailed t_3 law the 4M value is 0.026 — still rejecting at 0.05, at a weaker level): the observed across-recipe variance exceeds the null at $p = 0.005$ (4M), 0.001 (10M), 0.049 (20M), 0.083 (150M). The simulated homogeneous panels are the estimator's negative control; the 1B row is a real-data instance on the null side: the observed variance (0.231) falls *below* the sampling floor (0.411), the test returns $p = 0.89$, and the spread estimate is exactly $1.00\times$; the estimator does not manufacture heterogeneity where none is present. The sub-floor 1B reading itself we report without a settled mechanism: the across-recipe variance of $\log \hat{\sigma}$ is more concentrated than independent χ^2_2 sampling predicts, and the three seeds are distinct runs with distinct final values, so it is not a release artifact. (Per-cell SDs at df $= 2$ carry a $\sim$17% median downward bias — the atlas reports across-cell quantiles, not per-cell intervals.) Across all 11 domains the median spread factor at 4M/20M/150M is $1.54/1.48/1.87$ (ranges $[1.00, 2.29]/[1.00, 2.26]/[1.00, 3.82]$), so the C4-en headline is representative in direction; in magnitude it tops the per-domain range at 4M (2.29 is the maximum) and sits above the median at 20M, but below the cross-domain median at 150M (1.65 vs 1.87). Leave-one-recipe-out, the C4-en factor spans $[1.96, 2.36]$ at 4M, $[2.21, 2.66]$ at 10M, $[1.00, 1.83]$ at 20M, and

$[1.41, 1.69]$ at 150M: the 20M estimate is carried by a single recipe (DCLM-Baseline (QC 10%)). The per-domain homogeneity test rejects at $p < 0.05$ in $3/11$ (4M), $7/11$ (10M), $4/11$ (20M), and $6/11$ (150M) domains (other scales: 6M $6/11$; 8M $1/11$; 14M $5/11$; 16M $1/11$; 60M $3/11$; 90M $2/11$; 300M $3/11$; 530M $7/11$; 750M $5/11$; 1B $3/11$; $\approx 0.6/11$ expected under homogeneity). The extreme recipes by name: at 4M the noisiest is Dolma1.7 (no Reddit) ($\hat{\sigma} = 0.0429$) and the quietest FineWeb-Edu (0.0006); at 10M, DCLM-Baseline 25%/Dolma 75% (0.0398) vs DCLM-Baseline (QC 7%, FW3) (0.0004); at 150M, DCLM-Baseline (0.0060) vs Falcon+CC (QC 10%) (0.0003). The two quietest cells pass a model-consistency check: their three seeds are distinct runs with distinct final values (they are simply tight); they are not instances of the Appendix F failure.

Two numeric curiosities, explained. (i) The median within-recipe SD at 6M/8M/16M agrees to three significant figures (0.004059/0.004060/0.004062). The per-cell SDs are all distinct (25/25 per size), so the grid is not degenerate; it is the *median* that is pinned. The mechanism is the discreteness of the macro aggregate: each seed's macro score averages ten fixed-size task accuracies, and at these sizes the median recipe's cross-seed spread is dominated by the coarsest tasks (`openbookqa`: 500 items; at 6M/8M the median cell is `Falcon+CC (QC Tulu 10%)`, whose openbookqa spread is an 18-item max–min swing — 0.036 task points, a third of that cell's macro-level SD; at 16M the median cell is a different recipe, so the coincidence lives at the distribution level, not the recipe). (ii) The 20M perplexity median (0.0288) exceeds both 16M (0.0111) and 60M (0.0125). The 20M release evaluates on a coarser shared grid on the perplexity release (12 common checkpoints, against 16 at 16M and 24 at 60M). One candidate mechanism: on a sparse grid the runs are still descending at seed-dependent rates, and residual trend reads as cross-seed noise. We flag this as a hypothesis: the within-scale gradient is correlational, and at a fixed final step the endpoint SD cannot respond to grid density (the final common step is 14584 for every cell), so the elevation's cause is not identified; the medians are nonetheless insensitive to the endpoint window (recomputed on per-seed means of the last two common checkpoints, no scale's median moves by more than 11% — e.g. 20M $0.0288 \rightarrow 0.0295$); the inflation concentrates in the domains with the sparsest grids (ICE, dolma_stack, WIKITEXT). Regressing the per-scale median on the grid column gives slope -0.59 in log-log ($r = -0.57$, $p = 0.035$; confounded with scale itself, since shorter grids coincide with smaller runs); within-size, the gradient is significant exactly where flagged (20M: $r = -0.19$, $p = 0.002$) and weak elsewhere (60M: $p = 0.06$; 150M: $p = 0.10$). So the contamination is concentrated at 20M as flagged, and the atlas's remaining magnitudes stand.

PolyPythias outliers. The atlas also surfaces outlier runs that matter for anyone reusing these releases: BLIMP at 31m has seed5 at robust-$z = -11.9$ and seed0 at $z = -4.2$, consistent with the instabilities PolyPythias itself reports; the 410m-seed6 trajectory has no evaluations past step 70k. The 31m BLIMP cell's 10-seed relative SD is 0.0216 with both runs included and 0.0051 without them; we keep them in the headline numbers (divergent runs are the tail realizations a noise floor exists to warn about) and report both versions wherever the cell is used. With them excluded, the Section 3 cross-check median over its 9 cells moves from $1.8\times$ to $1.3\times$.

F Release-level label–configuration mismatch in the PolyPythias release

Our parsing pipeline applies a model-consistency check to every released evaluation file: the `config.model_args` field must name the checkpoint the file claims to evaluate. On the public PolyPythias release this check rejects 1,095 files (Table 19). The lambada results stored under `pythia-410m-seed{1..9}/` name not the seed variants but the base model `EleutherAI/pythia-410m`: the nine "seed" final scores are bit-identical, with SD $= 0$. A naive consumer of this release, one who trusts directory labels, would compute that 410m pretraining has *exactly zero* seed variance on lambada, and could carry that "fact" into a power analysis or a noise-floor comparison. After exclusion, the lambada dimension of this 50-run asset is simply unusable for seed variance, which is how we treat it throughout.

An upstream report will be filed with the maintainers after the review period. Our reading of the release is a parsing-level observation offered in a collaborative spirit: the underlying runs are unaffected, and the fix is a relabeling, not a rerun.

Table 19: **The label–config mismatches, in numbers.** Files rejected by the model-consistency check on the PolyPythias release: 1,095 lambada files whose `config.model_args` name the base model while their paths name seed variants; the nine "seed" final scores are bit-identical (SD $= 0$). After exclusion, the release's lambada dimension carries no seed information.

directory (all files: `lambada_openai` only, 2024-08-15 batch)	files rejected
`pythia-410m-seed1`	104
`pythia-410m-seed2`	108
`pythia-410m-seed3`	152
`pythia-410m-seed4`	154
`pythia-410m-seed5`	154
`pythia-410m-seed6`	154
`pythia-410m-seed7`	116
`pythia-410m-seed8`	116
`pythia-410m-seed9`	37
total	1,095

G THE READOUT-METRIC PANEL

Recomputing Section 5's SNR under each released readout metric (same checkpoints, same compute, only the readout changed), we observe median SNRs under bits-per-byte that are 1.8–$4.0\times$ those under the primary accuracy metric at matched scales (bpb is recorded only up to 90M in this asset); under the deconvolved estimand of Section 2 the ratio is 2.2–$5.8\times$ at $\leq$20M and $\approx 2\times$ at 60M–90M. This independently reproduces, on a suite S&N did not test, that paper's own recommendation to read bpb rather than accuracy. The estimand caveat of Section 7 applies to every cell of this panel. The SNR values in this panel use the observed definition of Section 2 (seed noise not subtracted from the signal); the resulting upward bias inflates low SNRs more than high ones, so the observed ratios are understated at $\leq$20M and overstated at 60M–90M (deconvolved row, Table 20). Table 15 quantifies the bias on the primary metric.

Table 20: **The readout metric alone moves the instrument's SNR.** Same checkpoints, same compute, same candidate set (25 recipes $\times$ 3 bundled seeds per scale); only the readout metric changes. Each entry is the median over tasks of (signal across recipe means)/(median within-recipe 3-seed SD), computed on the released S&N re-evaluation of DataDecide checkpoints — note the estimand differs from the macro computation on the official DataDecide table (e.g. the primary-metric SNR at 90M reads 2.07 task-median here vs 3.60 macro-observed / 2.61 macro-deconvolved there; Table 15). Cells are paired re-readouts of one candidate set, not independent measurements; the task median includes two aggregate scores that are deterministic functions of the others. bits-per-byte is recorded only up to 90M in this asset — a byte-accounting gap: the correct logit/char row (the same log-likelihood signal without byte normalization) continues the pattern to 300M (11.25/15.52 at 150M/300M); at $\geq$530M the release contains a single seed and no within-recipe noise is estimable. On the DataDecide official macro table (different estimand, Section 5), the primary-metric SNR at 530M/750M/1B is 3.42/5.26/5.79.

readout metric	4M	6M	8M	10M	14M	16M	20M	60M	90M	150M	300M
primary (task default)	0.79	1.18	1.09	1.00	0.94	1.24	1.10	1.49	2.07	2.89	2.70
raw accuracy	0.82	1.23	1.07	1.18	1.13	1.09	1.22	1.53	2.38	2.96	3.68
accuracy/char	1.00[†]	0.95	1.07	1.07	1.01	1.09	1.34	1.99	2.83	2.89	3.41
uncond. accuracy	0.79	1.02	0.85	1.05	1.00	1.14	1.19	1.48	2.07	2.22	2.70
bits-per-byte	**2.36**	**2.52**	**2.83**	**2.83**	**2.47**	**2.75**	**1.99**	**4.49**	**8.26**	—	—
correct prob./char	2.68	3.09	2.89	2.94	2.70	3.04	1.90	3.81	7.74	8.86	11.58
correct logit/char	2.36	2.53	2.83	2.83	2.47	2.75	1.99	4.49	8.26	11.25	15.52
norm. prob./char	1.19	1.48	1.46	1.31	1.72	1.53	1.38	2.12	2.87	3.86	3.95
margin/char	2.05	1.55	1.24	1.24	1.08	1.18	1.48	1.86	2.45	3.62	4.11
bpb / primary	2.99×	2.14×	2.60×	2.83×	2.63×	2.22×	1.81×	3.01×	3.99×	—	—
bpb / primary (deconv.)	5.00×	2.93×	2.91×	5.83×	3.34×	3.00×	2.19×	1.91×	2.17×	—	—

[†]The 4M accuracy/char median includes one task (boolq) whose within-recipe seed SD is exactly zero at the final common step (SNR$=\infty$, absorbed by the median); excluding it gives 0.92. Uncond. accuracy is undefined (all-zero) for boolq and winogrande; its medians are over the remaining 9 tasks.

H SELF-RUN CROSSED CONTROLS: THE INTERACTION TERM, MEASURED

Sections 3–4 audited the instruments on public assets alone. Two questions those assets cannot answer are answered here with small self-run controls: whether the init×order interaction is material (the term Remark 1 leaves unidentifiable from single-source arms), and whether the checkpoint proxy calibrates once the seed budget is raised.

A second decomposition point: PolyPythias 160M decoupled arms. PolyPythias releases model weights — but no evaluation — for three weight-seed (init-only) and three data-order (order-only) 160M variants alongside its ten bundled seeds. We evaluate all 16 models on six tasks (lm-eval-harness 0.4.12, task defaults; two tasks rebuilt from canonical sources as noted below). Table 21 gives the per-task variance decomposition.

Table 21: **Source decomposition at PolyPythias 160M (first evaluation of the decoupled variants).** Per-task within-group SD of final scores across the three weight-seed runs (σ_{init}), the three data-order runs (σ_{order}), and the ten bundled runs (σ_{bundled}); the interaction point estimate $\hat{\sigma}^2_{\text{int}} = \hat{\sigma}^2_{\text{bundled}} - \hat{\sigma}^2_{\text{init}} - \hat{\sigma}^2_{\text{order}}$ with a 95% bootstrap interval over the seed populations; and the additivity factor $\sqrt{\hat{\sigma}^2_{\text{init}} + \hat{\sigma}^2_{\text{order}}}/\hat{\sigma}_{\text{init}}$ (the contrast carries $-V_\varepsilon$: it understates the interaction by one run-noise unit, unmeasured on this asset). At three runs per arm the interaction is unresolvable: intervals contain zero on all six tasks, and the near-entirely-negative `arc_easy` interval prices the estimator's sampling floor, not a negative variance. The init:order SD ratio ($\hat{\sigma}_{\text{init}}/\hat{\sigma}_{\text{order}}$) spans 0.49–2.67 on five of six tasks — within the pointwise 95% sampling band of an $F(2,2)$ law — and drops to 0.08 on `social_iqa` (whose $\hat{\sigma}_{\text{init}}$ is near zero).

task	$\hat{\sigma}_{\text{init}}$	$\hat{\sigma}_{\text{order}}$	$\hat{\sigma}_{\text{bundled}}$	$\hat{\sigma}^2_{\text{int}}$ [95% CI]	additivity factor
`blimp`	0.0088	0.0033	0.0133	+0.00009 [−0.00002, 0.00023]	1.07
`arc_easy`	0.0081	0.0167	0.0059	−0.00031 [−0.00041, 0.00002]	2.28
`arc_challenge`	0.0069	0.0045	0.0055	−0.00004 [−0.00006, 0.00004]	1.19
`piqa`	0.0049	0.0057	0.0045	−0.00004 [−0.00006, 0.00002]	1.52
`social_iqa`	0.0006	0.0076	0.0046	−0.00004 [−0.00006, 0.00003]	12.97
`lambada_openai`	0.0127	0.0191	0.0137	−0.00034 [−0.00051, 0.00022]	1.80

Reading: the interaction estimate is consistent with zero on all six tasks — an absence-of-evidence statement at three runs per arm, not an equivalence claim; read strictly, the interaction is *unresolvable* at this arm size, and as a working approximation at 160M, independent additivity is compatible with what we measure, which is exactly the seed-semantics reading of Section 3. The init:order SD ratio drifts across tasks (0.49–2.67 with the df= 2-fragile `social_iqa` readout excluded; 0.08–2.67 with it — the five are within the pointwise 95% sampling band of an $F(2,2)$ law), and one cell (`arc_easy`) even has the bundled arm below the order-only arm — at three runs per arm the component estimates themselves are noisy, and we report the interaction as an interval rather than a point throughout. So no single rescaling constant transfers — the band is real even where additivity holds. Caveat: three runs per arm (df = 2) make the component intervals wide; the sign structure is the informative part.

Crossed init×order training at 20M, on DataDecide's configuration. We train a 4×4 crossed grid plus six extra bundled-seed runs (ten bundled runs counting the four shared diagonal cells) at 20M (1.91B tokens, 5×C) with DataDecide's published configuration (their OLMo branch's 20M model and learning-rate/batch schedule), splitting the seed into independent init and order streams — the design Remark 1 says is required. Per task (Table 22), the init and order *main effects* are small and not consistently detectable at this scale (the method-of-moments estimates are zero on half the tasks, small positive on the rest — up to 0.5 points against residuals of 0.4–1.3 points), while the cell-level residual — the joint seed draw, which one run per cell makes an *upper bound* on the interaction — dominates at 0.5–1.0 points on the accuracy tasks (the completed design's total $\sqrt{\hat{\sigma}^2_c + \hat{\sigma}^2_\varepsilon}$; the original single-replicate grid read 0.4–1.3; 0.0004 nat on `lambada_openai`). To separate the two terms of that residual we complete the grid to 4×4×3 (three full replicates per cell, same init and same order seeds — run noise is the residual run-to-run variation under explicit RNG and identical configuration): per task, method-of-moments variance components with a model-based F-inversion interval on the interaction share $\rho = \sigma^2_c/(\sigma^2_c + \sigma^2_\varepsilon)$ (Table 22). The interaction share is 0.00–0.29 (median 0.05): no task is interaction-dominant — on three of six the share's interval clears out a majority (run noise dominates there, upper end below 0.5); on the other three

the split is unresolved (upper ends 0.53–0.70). The 20M joint-cell residual is thus not evidence of a material init×order interaction; where resolvable, it is run-level nondeterminism. The bundled arm (DataDecide's own single-stream semantics) sits at the same total magnitude. At this scale, seed noise does not decompose into useful init/order main effects at all. Nor does a physical single-source arm read that margin: fixing one order level and varying init, the conditional-arm SD lands at 0.66–1.59× the bundled arm's SD across tasks (mean ≈ 1.0) — such an arm reads a *conditional slice* of the joint-cell noise (margin + interaction slice + run noise), not the (here undetectable) init margin alone. What it cannot do is attribute or bound: nothing in the readout separates the components, and the realized value swings with the arbitrarily fixed level (same-task slices span up to 8× at this replication depth). At 20M a single-source floor measures *a* noise, not *the init noise* (script: `verify-scripts/conditional_arms.py`).

Table 22: **Crossed init×order decomposition at 20M (DataDecide configuration), completed to 4×4×3.** Main-effect SDs from the completed grid (tested against the interaction mean square, $F_{3,9}$; the smallest uncorrected main-effect p is 0.052 (`blimp`, order), and none survives multiplicity control across the twelve tests; unrounded mean squares in `crossed_grid.json`); the measured run-level noise $\hat{\sigma}_\varepsilon$ (within-cell, df = 32) and the interaction SD $\hat{\sigma}_c$ (truncated at zero; the two untruncated negative estimates are reported as ∼0); and the interaction share $\rho = \sigma_c^2/(\sigma_c^2 + \sigma_\varepsilon^2)$ with the model-based F-inversion 95% interval (three full replicates per cell). No task is interaction-dominant; three are run-noise-dominant. The † row's within-cell variance sits at the quantized-score floor ($\mathrm{MS}_E \approx 0$), so its F-inversion interval is ill-conditioned (reads $[0, 0.70]$). A batch fixed effect between the earlier and later run batches moves no verdict (`crossed_grid.json`).

task	$\hat{\sigma}_{\text{init}}$	$\hat{\sigma}_{\text{order}}$	$\hat{\sigma}_\varepsilon$ (run noise)	$\hat{\sigma}_c$ (interaction)	share ρ [95%]	bundled, n=10
`blimp`	0.0036	0.0053	0.0098	0.0029	0.08 [0, 0.54]	0.0081
`lambada_openai`	∼0	∼0	0.0003	0.0002	0.29 [0, 0.70]†	0.0004
`social_iqa`	0.0001	∼0	0.0054	0.0008	0.02 [0, 0.48]	0.0038
`arc_easy`	∼0	∼0	0.0086	∼0	0.00 [0, 0.31]	0.0067
`arc_challenge`	0.0007	0.0022	0.0082	∼0	0.00 [0, 0.41]	0.0073
`piqa`	∼0	0.0016	0.0054	0.0015	0.08 [0, 0.53]	0.0062

The same 20M runs calibrate the proxy at a ten-seed budget, on the two tasks with checkpoint-dense evaluations in the grid (`blimp`, `lambada_openai`). The last-five-checkpoint proxy under-reads the true 10-seed bundled noise by a median 4.0× (`blimp`) / 2.4× (`lambada_openai`) — with run-level bootstrap 95% intervals $[2.2, 5.6]$ / $[0.9, 3.8]$ (resampling the ten bundled runs, recomputing the median each time; `apph_proxy_boot.json`) — the intervals support the direction more strongly than the precise magnitude — worse than the 1.6× on DataDecide's transferred panel. At 20M the proxy is miscalibrated in magnitude even at a ten-seed budget. At 160M one task admits the same read: on `blimp`, the only task with checkpoint-dense evaluations of eight of the ten bundled PolyPythias runs, the last-five-checkpoint proxy under-reads the cross-seed noise by a median 5.1× (Table 23).

Table 23: **Three decomposition points read differently.** At 1B the two sources are comparable in magnitude and the additivity price is 1.62; at 160M the interaction is not detectably nonzero but the init:order ratio drifts by task; at 20M the main effects are not detectable and the joint cell dominates. Whether this is a scale trend or an asset difference is undecidable from three points on three different assets. The df= 2-fragile `social_iqa` cell (additivity factor 12.97) is included in the 160M median; excluding it shifts the median to 1.52.

asset / point	decomposition $\sqrt{\sigma_i^2+\sigma_o^2}/\sigma_i$	interaction	proxy calibration
S&N 1B arms	median 1.62 (IQR 1.40–1.68)	not measurable	calibrated at home (R^2 0.81/0.88); the advertised 0.9 is uncertifiable at n=3 (Prop. 1)
PolyPythias 160M decoupled	median 1.66 (IQR 1.19–2.28)	unresolvable at n=3 (intervals contain 0, 6/6 tasks)	$y/x = 5.1$ (`blimp`)
20M crossed (this work)	main effects ≈ 0	interaction share 0.00–0.29 (median 0.05); run-noise-dominant on 3/6 tasks (4×4×3)	$y/x = 2.4$–4.0 at n=10

Declared deltas and limits. Recipe: the dolmino mix's high-quality Common-Crawl portion (DCLM-Baseline itself is not retrievable through our mirror); trainer: a minimal re-implementation of the published 20M configuration, verified to the documented parameter count; evaluator: a batched log-likelihood evaluator with the task YAMLs' scoring rules replicated verbatim (stage-1 PolyPythias evals used lm-eval-harness 0.4.12, with two tasks rebuilt from canonical sources because the HF mirror's LFS is unreachable from our cluster). One run per grid cell: to bound the run-level nondeterminism this leaves, we first re-ran five cells end-to-end with identical seeds (the four diagonal cells and one off-diagonal cell); the per-task final scores move by 0.0002–0.023 in the task's score units (median 0.005; up to 2.3 percentage points on the accuracy tasks). We then completed the grid to $4\times4\times3$ (three full replicates per cell): the residual decomposes into the measured run noise $\hat{\sigma}_\varepsilon$ plus an interaction term whose share of the residual is 0.00–0.29 with model-based F-inversion intervals (median 0.05; no task interaction-dominant, three run-noise-dominant; Table 22). The 20M residual is therefore *consistent with run-level nondeterminism and carries no measurable interaction-dominant signal*; the headline reading — that init and order main effects are not detectable at this scale's power — is unaffected, since those are estimated from the row/column margins, not the cells.