WHAT DOES A TRAINING TOKEN BUY? CONTEXTUAL INFORMATION GAIN BEYOND PERPLEXITY ACROSS OPEN MODEL SUITES

Anonymous authors
Paper under double-blind review

ABSTRACT

A lower perplexity on held-out text can come from two sources. The model may use the context better, or it may make a better guess about the text before any context is given; perplexity adds the two into one number, so a falling perplexity cannot say which improved. We separate them. For every held-out token we score the model twice, with its context and under its own unconditional next-token distribution, obtained by averaging its predictions over reference contexts. The difference, contextual information gain (CIG), is the pointwise mutual information between context and token under the model, so a higher value means the context did more of the work; the first term is the usual held-out cross-entropy, and the second tracks how the model's unconditional expectations move during training. We measure both on 790 public checkpoints from seven model families (14M to 12B parameters, up to $1071\times$ the Chinchilla-optimal data budget) on 206 held-out texts from ten domains, in bits per byte, and judge every trend against the model's own checkpoint-to-checkpoint noise. In the three training runs whose learning rate and training-data composition stay constant, held-out cross-entropy keeps falling to the end of the window, at up to $173\times$ the optimal budget, while the model's prior moves away from held-out text by 18 to 29% of the loss improvement without becoming more peaked. The additive scaling law predicts that after the same number of training tokens every model size improves at the same rate per factor of two in tokens; instead, larger models improve faster, at all 58 matched comparisons on four model ladders across three corpora. Only the 70M to 160M Pythia models get worse late in training, whereas models trained far beyond their optimal budget keep improving. In warmup-stable-decay schedules the decay phase lowers held-out loss $17\times$ faster per factor of two than the stable phase in three runs and $36\times$ faster in a fourth; every documented decay phase coincides with a change of training data, but within the one decay phase with 13 checkpoints, at unchanged data, the rate rises $3.5\times$ as the learning rate falls to zero. The measurement thus turns the model's own noise into a stopping diagnostic, and tells, from checkpoints and held-out text alone, whether the next factor of two in training tokens still improves the use of context or only shifts the model's expectations.

1 INTRODUCTION

Everyone measures training loss; almost no one asks what a marginal token of training buys on text the model has never seen - and when, if ever, it stops buying anything. Compute-allocation studies answer a different question: given a budget, how to split it between parameters and data (Kaplan et al., 2020; Hoffmann et al., 2022). They are fit to final losses of completed runs, and they do not say whether the next doubling of data still improves how *this* model uses context on held-out text, or only rearranges what it expects before seeing a word; a loss curve cannot say either.

Part of the difficulty is the instrument. A model can improve its held-out loss in two distinct ways, by extracting more from context or by having a better unconditional expectation of the text, and perplexity sums the two. We measure them separately. For each held-out token x_t with context $x_{<t}$,

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.

contextual information gain is the difference of two surprisals,

$$g(t) \;=\; s_{\mathrm{ref}}(t) - s_{\mathrm{ctx}}(t) \;=\; \log_2 \frac{p_\theta(x_t \mid x_{<t})}{\bar{q}(x_t)} \quad [\text{bits}], \qquad \bar{q}(v) \;=\; \frac{1}{N}\sum_{i=1}^{N} p_\theta(v \mid c_i), \tag{1}$$

where $\bar{q}$ is the model's *own implied unconditional prior*, estimated by averaging its predictive distributions over N reference contexts c_i drawn from held-out documents. With this reference, g is exactly pointwise mutual information between context and token under the model, and its expectation is the model-relative predictive information $I(X_t; X_{<t})$ (Bialek et al., 2001; Crutchfield & Feldman, 2003); it is the inference-only autoregressive analogue of pointwise $\mathcal{V}$-information (Ethayarajh et al., 2022). Because the probe texts are fixed, the reference term obeys an exact decomposition,

$$\mathbb{E}_P[-\log_2 \bar{q}] \;=\; H(P) \;+\; \mathrm{KL}(P \,\|\, \bar{q})\,, \tag{2}$$

with $H(P)$ constant across checkpoints: *every* movement of the reference term is KL motion of the implied prior relative to held-out text, measured against the reference sample it is estimated on. One sweep tracks both terms with no extra forward passes. The *reading gain* used below is the decrease of held-out cross-entropy per doubling of training tokens, so the duration, scaling, corpus and schedule results need no reference at all; the decomposition adds the second term - the motion of the model's own prior, which cross-entropy sums in and hides - and a per-token quantity for domain analysis. A reference is a choice; §3 and §4 test its effect.

We point this instrument at seven open model families that release intermediate checkpoints - 790 checkpoints spanning 14M - 12B parameters and about $1\times$ to $1071\times$ the Chinchilla-optimal budget (Figure 1) - and score all of them on the same 206 domain-stratified held-out probes, in bits per byte (bpb) so that seven tokenizers are comparable, with a per-model noise floor and two-level bootstrap intervals on every slope. Each result uses the subset of checkpoints that can support it: duration claims use the three long windows in which nothing but the token count changes; the scaling test compares sizes after the same number of training tokens within 20 series; the corpus and repetition results use Pythia's twin runs; the schedule results use the four warmup-stable-decay runs. The first three findings below are the contribution; the stopping diagnostic of §8 is offered as a diagnostic, not a rule.

Findings. (1) Held-out loss keeps falling. In the three windows where learning rate and data mixture are verifiably constant, held-out cross-entropy falls by 0.006 to 0.009 bits/byte per doubling of training tokens (a factor of two more tokens, never a repeated pass over the data), every interval excluding zero, out to 37 - $173\times$ the Chinchilla budget; two constant-LR windows with unverified mixtures agree to $556\times$ (§4). **(2) The prior drifts, and does not sharpen.** Against the fixed 12-document reference, the implied prior moves away from held-out text at 18 - 29% of that improvement, re-assigning its mass at constant entropy (§4). **(3) Larger models improve faster at equal data.** After the same number of training tokens, larger models gain more per doubling at all 58 matched comparisons on four fp32 ladders across three corpora, which the additive scaling law rules out; a pair of sizes sharing peak learning rate and data order shows the same interaction late in training (§5). **(4) Late decline is a size band, not an overtraining threshold.** Only 70 - 160M models get worse late in training, while models at up to $11.6\times$ higher training multiples keep improving; a repetition penalty (the 1.45-epoch deduplicated run's shortfall behind the 0.9-epoch standard run) is negative in all ten domains, largest in code, and within one floor of zero from 1B parameters on. The advantage on verifiably seen text shrinks with size (§6). **(5) The decay phase improves 17 - 36$\times$ faster per doubling**, and in the three schedules that anneal to zero the rate climbs into the anneal (§7). **(6)** Domain order is a plateau property that dissolves late, and a model's own noise floor gives a stopping diagnostic we are developing (§8). The instrument measures; it does not explain: no causal claim about *why* larger models gain more per token, and no allocation claim.

2 Background and related work

The instrument. Estimating the information content of text from predictive models goes back to Shannon (1951), and the mutual information between past and future symbols is the predictive information of Bialek et al. (2001) and Crutchfield & Feldman (2003). The choice in Eq. 1 that carries the weight is the reference $\bar{q}$. The popular shortcut, the model's own beginning-of-sequence

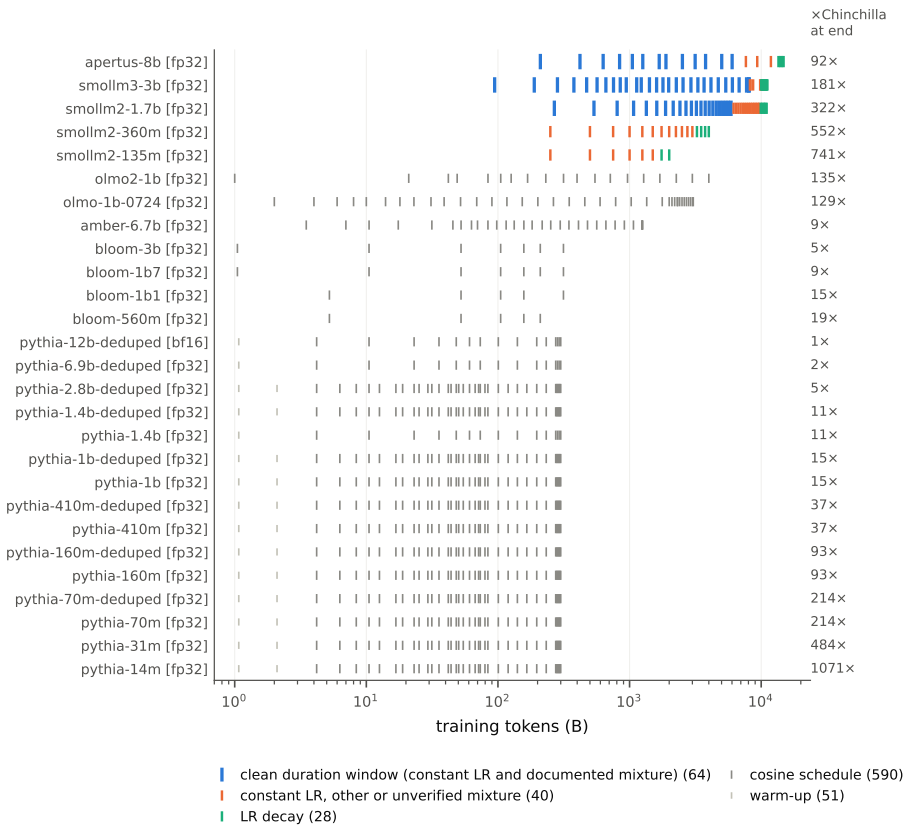

Figure 1: The 790 checkpoints of the 27 series, placed in training regimes from documented boundaries only: blue = clean duration windows (constant LR and documented mixture: Apertus 13, SmolLM2-1.7B 22, SmolLM3 29 checkpoints), orange = constant LR with another or unverified mixture, green = LR decay, grey = cosine schedules, faint = warm-up. Pythia-2.8B on the standard Pile is excluded: its public intermediate branches serve other weights (Appendix A).

token as "empty context", is not a neutral reference: the first position behaves as an attention sink (Xiao et al., 2024), and its next-token distribution at released checkpoints is far from any marginal, which is why we average over reference contexts instead. A separate methods paper, under review (Anonymous, 2026), quantifies that failure and the stability of the marginal-by-averaging reference across architectures, corpora and tokenizers. The present paper takes the definition as given, reuses none of that paper's results as a contribution, and asks what the instrument measures *across training*: duration, scale, corpus, contamination, schedule and domain, including what the reference term's motion over training does and does not mean (§4).

Scaling laws. Kaplan et al. (2020) fit joint power laws in parameters N and data D with an interaction; Hoffmann et al. (2022) fit the additive form $L = E + AN^{-a} + BD^{-b}$, whose replications and reconciliations (Besiroglu et al., 2024; Porian et al., 2024; Pearce & Song, 2024) concern the compute-optimal frontier, and recent work asks which component of the cross-entropy scales (Yan et al., 2025). Our question is orthogonal: not where the loss-minimising (N, D) lies, but what a marginal token buys on held-out text along real training runs - including regimes those fits exclude: data repetition (Muennighoff et al., 2023), heavy over-training (Gadre et al., 2024; Springer et al., 2025), and schedule effects (Hu et al., 2024; Hägele et al., 2024).

Checkpoint suites. Public intermediate checkpoints make training dynamics observable (Biderman et al., 2023; Groeneveld et al., 2024; OLMo Team et al., 2024; Liu et al., 2023; Ben Allal et al., 2025; Hugging Face SmolLM team, 2025; Swiss AI Initiative, 2025; BigScience Workshop et al., 2022); prior trajectory studies track loss and benchmark curves (Xia et al., 2023), word acquisition (Chang & Bergen, 2022), internal state (Kim et al., 2025) or token-level loss for data selection (Lin et al., 2024), and Paloma (Magnusson et al., 2023) establishes bits per byte as the cross-tokenizer unit. New here is the decomposition, not the checkpoints.

3 MEASUREMENT SETUP

Probes. 206 held-out texts: 200 domain-stratified - Wikipedia 6×20 (biology, economics, history, philosophy, physics, sociology), 20 PubMed Central articles (biomedicine), 20 permissively-licensed Python files (software), 20 Project Gutenberg passages (literature-classics), 20 Federal Register rules (government-regulation) - plus 6 legacy probes. Each is truncated at 1800 tokens under each model's own tokenizer; the set is English-only and frozen, so every "domain" claim is about these probes. Each probe carries two pre-registered annotations: an epistemic-axis label (nomothetic 80 / mixed 80 / idiographic 40; Windelband, 1980; Allport, 1937; Biber, 1988) and a contamination prior (150 *certain* in pre-2020 corpora; 30 verifiably *post-cutoff*, 2026; 20 *likely*).

References. The headline prior $\bar{q}_B$ is estimated from 12 register-balanced held-out documents (21,600 contexts; two of them code, both from one Python library); a legacy 3-document Wikipedia prior $\bar{q}_A$ (5,400 contexts) is kept as a sensitivity check. Context-term results do not involve either; every drift number below names its reference (§4).

Model pool. Seven families - Pythia (both Pile variants, identical data order, 14M - 12B), SmolLM2 (135M/360M/1.7B), SmolLM3-3B, OLMo/OLMo2-1B, Amber-6.7B, Apertus-8B, and BLOOM (560M - 3B) - giving 27 model@precision checkpoint series, 790 checkpoints and 162,740 probe measurements (full pool and verified regime provenance in Table 4, Appendix A). All measurements are in 32-bit floating point (fp32, TF32 off); the single exception, Pythia-12B, is measured in 16-bit bf16 and confined to a labelled sensitivity arm. All cross-family quantities are in bits per byte, $\mathrm{bpb} = (\text{bits/token}) \times n_{\mathrm{tokens}}/n_{\mathrm{bytes}}$, the only scale comparable across tokenizers with 32k - 250k vocabularies (Magnusson et al., 2023).

Terms. The *context term* s_{ctx} is the surprisal of a held-out token given its context; averaged over probes it is the held-out cross-entropy. The *reference term* s_{ref} is the surprisal of the same token under the implied prior $\bar{q}$ of Eq. 1; its slope over training is the *drift* of the prior. A *doubling* is a factor of two in training tokens, one unit on a $\log_2$ axis, never a repeated pass over the data. The *reading gain* is the decrease of held-out cross-entropy per doubling, $-\,d\,s_{\mathrm{ctx}}/d\log_2 D$, in bpb per doubling. A model's *noise floor* is its own checkpoint-to-checkpoint scatter: the robust (MAD) SD of the second differences of its probe-mean curve, divided by $\sqrt{6}$. A slope is *resolved* when its bootstrap CI excludes zero and its magnitude exceeds the floor.

Estimators. A local gain is an OLS slope over ≥ 4 checkpoints spanning ≥ 0.8 doublings, and every slope carries a two-level bootstrap 95% CI (probes $\times$ checkpoints). Second differences annihilate a locally linear trend, so the floor measures checkpoint-to-checkpoint churn. The additivity statistic of §5 is a regression across models, resolved on its CI only. BLOOM's 6 - 7-point curves give no usable floor; its slopes are CI-tested only and labelled.

Clean windows. Only clean windows (constant LR, which in a warmup-stable-decay, WSD, schedule is the stable phase, and a constant, documented data mixture) support duration claims. Three models have one: SmolLM2-1.7B stage 1 and SmolLM3-3B stage 1, from their reports, and Apertus-8B stage 1. Apertus's window ends at 7,038 B, where its batch size doubles at constant learning rate and, by its report, its first data stage ends (Swiss AI Initiative, 2025); the 7,652 B checkpoint beyond it is kept only as a sensitivity row. SmolLM2-135M/360M have constant-LR windows with unverified mixtures ("suggestive"). Cosine families are schedule-confounded throughout and are used only for cross-model comparisons at the same token counts.

Precision. The fp32 anchor (one fixed checkpoint, six legacy probes) reproduces to $\leq 5 \times 10^{-5}$ bits/token across five runs, three GPUs and five days, and is re-run in every scoring session; bf16 corrupted the conditional term by $+0.76$ bits at 1800-token contexts and was not reproducible, so it never pools with fp32, and Pythia-12B @ bf16 is kept only as a labelled sensitivity arm (Appendix B).

4 HELD-OUT LOSS KEEPS FALLING, AND THE PRIOR DRIFTS

Table 1 is the duration result. All three clean windows lie above the 70 - 160M band where §6 finds held-out loss can rise again, so the result is a statement about these long windows and about models above the band. In all three the reading gain is positive and resolved, $+0.006$ to $+0.009$ bits/byte per doubling, and none is decelerating into its noise floor: over each window the held-out gain totals

Table 1: Every constant-LR window (bpb per doubling of training tokens; two-level bootstrap 95% CIs, joint across terms). Drift = slope of the reference term against $\bar{q}_B$ (Eq. 2). The first three rows are the clean windows and total 64 checkpoints ($13 + 22 + 29$); Apertus's uncapped window is a sensitivity row; n.r. = not resolved.

model	window	n / dbl.	×Chin	gain [95% CI]	drift vs. $\bar{q}_B$	drift/gain
Apertus-8B	stage 1, ≤7.04T	13/4.84	37×	$+\mathbf{0.0087}$ $[0.0061, 0.0108]$	$+0.0026$ $[0.0013, 0.0034]$	29.3% [20.0, 36.8]
SmolLM2-1.7B	stage 1 (web)	22/4.46	173×	$+\mathbf{0.0066}$ $[0.0051, 0.0075]$	$+0.0012$ $[0.0004, 0.0016]$	17.6% [7.2, 23.3]
SmolLM3-3B	stage 1	29/6.43	132×	$+\mathbf{0.0060}$ $[0.0048, 0.0072]$	$+0.0014$ $[0.0010, 0.0019]$	24.1% [19.2, 28.1]
Apertus-8B	≤8T	14/5.19	48×	$+\mathbf{0.0094}$ $[0.0064, 0.0114]$	$+0.0028$ $[0.0016, 0.0037]$	29.9% [21.4, 37.4]
SmolLM2-360M	mix. unverified	12/3.58	414×	$+\mathbf{0.0042}$ $[0.0026, 0.0049]$	$+0.0012$ $[0.0007, 0.0016]$	29.8% [22.0, 46.1]
SmolLM2-135M	mix. unverified	6/2.58	556×	$+\mathbf{0.0037}$ $[0.0027, 0.0042]$	$+0.0005$ $[-0.0019, 0.0012]$ n.r.	12.2% [−60.6, 31.8]

0.03 - 0.05 bits/byte, at 5 - 11 noise floors per doubling, and held-out loss is still falling at 173× the Chinchilla-optimal budget. The two unverified-mixture windows agree to 556× ("suggestively"). Three short mid-training constant-LR windows (SmolLM2-1.7B stages 2 and 3, SmolLM3 stage 2; 5 - 7 checkpoints over about a quarter-doubling each) are *not* resolved and stand as the counterweight (Appendix C).

The prior drifts away while held-out loss falls. In the same windows the reference term *rises*: measured against the fixed 12-document reference, the implied prior $\bar{q}_B$ moves away from the held-out probes at 17.6 - 29.3% of the reading gain, with every interval excluding zero (envelope 7 - 37%); SmolLM2-135M, whose drift is unresolved (12.2%), is excluded by name. In every one of the 27 series the reference term rises over its longest regime, with a CI excluding zero in 26 (Appendix C), though outside the clean windows this is schedule-confounded corroboration only.

The drift is not sharpening. A prior that merely became peakier would also raise the reference term. It does not happen here. We compare each checkpoint's prior $\bar{q}_t$ with the window's final prior (0.05 - 0.075 bits apart at the window start). Matching the entropy of $\bar{q}_t$ to the final prior's removes none of that divergence: 99 - 102% is left at every checkpoint of all three clean windows (above 100% because the entropy-matching temperature is not the divergence-minimising one), and the best temperature removes at most 1.7% (prediction 28, Appendix I). The prior's entropy changes by only 0.2 - 0.3% of its 10.1 - 10.6 bits over each window, a trend the data do not resolve: the mass is re-assigned at nearly constant entropy, not concentrated.

Whose drift is it? $\bar{q}$ is an average over a finite reference sample, so part of its motion could be the model predicting those documents better rather than moving away from held-out text; the drift is indeed larger against the narrower $\bar{q}_A$ (1.15 - 1.89× in point estimate; the ratio's interval excludes 1 only for SmolLM3). A test registered before it ran (prediction 27, Appendix I) asks whether that extra drift concentrates in the four domains $\bar{q}_A$ does not sample, which would put part of the drift's profile in the reference sample. It does not: the contrast is small and unresolved in all three clean windows (exact one-sided $p = 0.06, 0.31, 0.66$; Appendix C). As far as this test can tell the motion is the model's; the context-term results are unaffected either way.

5 Larger models improve faster at equal data: additivity fails

The additive family $L(N, D) = E + AN^{-a} + BD^{-b}$ (Hoffmann et al., 2022) makes a sharp prediction: the gain per doubling of data at matched D does not depend on N, i.e. the mixed derivative $\partial^2 L / \partial \log_2 N \, \partial \log_2 D$ vanishes. The Kaplan form, with its N - D interaction, permits a dependence (Kaplan et al., 2020), as do recent coupled forms (Videau et al., 2026; Xu et al., 2026). Intermediate checkpoints let us test this directly on held-out text: at each training-data point D shared by a ladder of sizes, we regress every size's local reading gain on $\log_2 N$; the slope $\beta(D)$, in bpb per doubling of data per doubling of parameters, should be zero.

Table 2 and Figure 2 give the result: $\beta > 0$ **and resolved at all 58 matched comparisons on four fp32 ladders across three corpora** - deduplicated Pile (7 sizes, median $+0.0050$), standard Pile (6 sizes, $+0.0011$), BLOOM/ROOTS (4 sizes, $+0.0214$: multilingual, a different positional mechanism) and SmolLM2's stable phase (3 sizes, $+0.0013$: no anneal anywhere in the window). Adjacent comparisons share overlapping local windows, so these are not 58 independent tests. The mixed-precision arm (adding 12B @ bf16 to the deduplicated ladder) is positive and resolved at all 13 comparisons, and under the non-embedding parameter count the medians shrink but no resolved

Table 2: The additivity test. +/−res = comparisons whose bootstrap CI excludes zero, by sign. †The ladder behind the registered abstract. ‡Mixed-precision sensitivity arm, never pooled with fp32.

ladder	sizes	matched on	comps.	med. β	+res	−res
Pythia, deduped Pile	7 (70M - 6.9B)	steps	13	+0.00498	13	0
Pythia, standard Pile†	6 (14M - 1B)	steps	31	+0.00109	31	0
BLOOM, ROOTS	4 (560M - 3B)	interp. grid	8	+0.02137	8	0
SmolLM2, stable only	3 (135M - 1.7B)	tokens ±12%	6	+0.00127	6	0
Pythia, dedup., mixed prec.‡	8 (incl. 12B)	steps	13	+0.00429	13	0

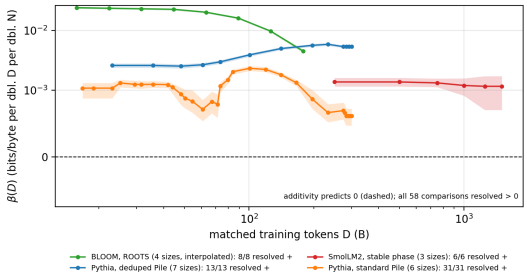

Figure 2: $\beta(D)$, the slope of the per-doubling reading gain against $\log_2 N$ at matched data D, with bootstrap bands on the four fp32 ladders (symmetric-log axis); additivity predicts zero (dashed). Every comparison is positive and resolved.

comparison changes sign. BLOOM's β, the largest in the study, replicates the *direction* on an independent corpus, not a magnitude: its ladder is interpolated and its token accounting unverified.

Not a learning-rate artifact. Peak LR falls with size across Pythia, so $\beta > 0$ could be an LR effect. Pythia-410M and 1B share a peak LR (3×10^{-4}), data order and corpus - one contrast over 1.29 doublings of N, which can corroborate the ladders, not replace them. Past 100 B tokens the larger model's per-doubling gain exceeds the smaller's at every matched checkpoint on both corpora (13/13 resolved, from 9 distinct local windows), by $1.2\times$ (deduplicated) and $3.8\times$ (standard) the corresponding ladder's median β over the same range and by 7.8 and 3.5 noise floors late in training; a β at fixed LR that is *larger* than the ladder's is the opposite of what an LR artifact predicts. Just after the warm-up guard (16.8 - 29.4 B tokens) the pair has the opposite sign - early, the larger model gains less per doubling - so additivity fails in both directions. The 6.9B/12B pair mixes precisions and is within one floor; we do not interpret it (Appendix D).

The sign of the effect is the same on every ladder; its size is not understood: the deduplicated ladder's median β exceeds the standard ladder's by $4.6\times$ ($6.3\times$ under non-embedding N, so not an artifact of how parameters are counted); we flag it unexplained.

6 Where improvement stops: a size band, repetition, seen text

Pythia's deduplicated and standard Pile runs share sizes, schedule and data order - a controlled size × corpus design (1.45 vs. 0.9 epochs); Figure 3 shows the late-window reading gain (tables in Appendix E).

Degradation is a size band. Two models reverse: 160M on *both* corpora (-0.0105 standard, -0.0151 deduplicated) and 70M under repetition only (-0.0068 vs. $+0.0026$). Everything smaller (484 - $1071\times$ the Chinchilla budget) and everything larger is still improving: 14M at $1071\times$ improves while 160M at $93\times$ degrades. On this instrument "overtraining" is not a threshold in D/N but a band around 70 - 160M, present on both corpora, with its minimum at 160M.

The repetition penalty is separable, present in every domain, largest in code, and gone by 1B. The corpus effect (deduplicated minus standard) is -0.0094 at 70M, -0.0045 at 160M, -0.0069 at 410M, then $+0.0004$ at 1B (0.4 floors) and -0.0010 at 1.4B (1.0 floor), both within the floor and so unresolved, as registered in advance. Where it exists it is negative in 10 of 10 domains at every size

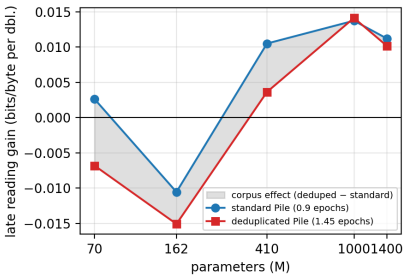

Figure 3: Late reading gain (bpb per doubling; late window = tokens $\geq 0.55\times$ the final count, the last $\approx$0.9 doublings) vs. size, five Pythia sizes on both Pile variants with identical data order. Late decline is a 70 - 160M band; the shaded corpus effect, the shortfall of the 1.45-epoch deduplicated run behind the 0.9-epoch standard run, is the repetition penalty, within one floor of zero from 1B.

Table 3: The harvest factor: the decay-phase reading gain divided by the same model's early stable-phase gain, each a least-squares slope over its whole window, with intervals from a paired probe bootstrap. f is the fraction of the model's own LR decay that its checkpoints cover. Acceleration is the second-half block factor minus the first-half one, in units of its checkpoint-noise σ, with the per-checkpoint noise measured in each model's own stable window and the shared middle checkpoint accounted for (other conventions in Appendix B). *Apertus with the capped stable window (§3); with the uncapped window, $15.88\times$ $[12.54, 19.26]$.

model	decay ckpts	factor [95% CI]	decay covered	data switch at onset	accel. (σ, calibrated)
SmolLM2-360M	4	$16.82\times$ $[15.9,\ 18.0]$	$f = 0.07$ - 1.00	unverified	4.6σ
SmolLM2-1.7B	4	$16.72\times$ $[12.9,\ 20.8]$	$f = 0.28$ - 1.00	yes (stage 4)	6.6σ
Apertus-8B*	4	$17.14\times$ $[13.6,\ 20.8]$	$f = 0.12$ - 0.88	yes (stage 5)	0.8σ
SmolLM3-3B	13	$35.77\times$ $[33.7,\ 38.0]$	$f = 0.01$ - 1.00	yes (stage 3)	11.5σ
three sparse, pooled		$16.84\times$ $[15.88,\ 17.80]$	inverse-variance; $Q = 0.03$, $p = 0.98$, $I^2 = 0$		

(28 of 30 cells resolved) but not uniform: a domain-label permutation test, the verdict we registered, detects structure at 70M and 410M ($p = 0.0005$ and $p < 0.001$; not at 160M, $p = 0.38$; Bonferroni threshold 0.0167), and code carries $1.9\times$, $2.8\times$ and $2.7\times$ the median domain's penalty at 70M, 160M and 410M. A code-concentrated excess is a content signature as much as a repetition one - deduplication removes near-duplicate files, and our code probes carry the "likely" contamination prior - and this instrument does not separate the two (Muennighoff et al., 2023). The Pythia authors found no benchmark evidence that the second epoch hurt (Biderman et al., 2023); held-out predictive information sees it at up to 8.4 noise floors.

Seen text: the advantage shrinks with scale. The probe set embeds a verifiable contamination split with genre fixed: PubMed Central articles from $\leq$2019, verifiably in the Pile (Gao et al., 2020), against 2026 articles no Pile model has seen. Across the fifteen verifiable Pythia series the seen-text advantage at the final checkpoint shrinks nearly monotonically with size, -0.056 bpb (14M) $\rightarrow$ -0.0014 (6.9B) (Appendix E), and late in training unseen text improves essentially as fast as seen text. Duration gains are generalisation, not recall. On memorisation we claim only what we measured: per-token statistics on the Pile-trained models show no concentration of the seen-text advantage onto individual tokens; the four BLOOM models show the opposite, reported as an open observation.

7 The decay phase improves faster, and accelerates as LR falls

In warmup-stable-decay runs the final phase improves much faster (Table 3): the per-doubling reading gain in the decay phase is 16.7 - $17.1\times$ the same model's early stable gain in three models (pooled $16.8\times$ $[15.9, 17.8]$) and $35.8\times$ $[33.7, 38.0]$ in SmolLM3; we call that ratio the *harvest factor* (its denominator is the early stable gain, not the gain just before the decay). Its level is confounded with the curriculum: in every run with a documented curriculum the decay begins together with a new data stage - SmolLM2-1.7B's stage 4 (Ben Allal et al., 2025), SmolLM3's stage 3 (Hugging Face SmolLM team, 2025), Apertus's final stage (Swiss AI Initiative, 2025) - so we speak of the

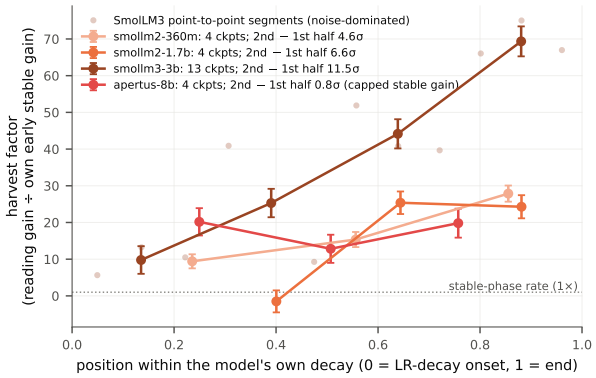

Figure 4: Harvest factor vs. position within each model's own decay: quarter-window block factors from their two endpoint checkpoints, with 1σ bars from each model's measured per-checkpoint scatter (Appendix B); Apertus on its capped stable gain. Faint dots: SmolLM3's noise-dominated point-to-point values (for the 4-checkpoint models the blocks are the segments). Three curves rise into the anneal; Apertus, the one schedule that stops at an LR floor, does not.

decay phase, not of the learning-rate decay. What is not confounded is the shape inside a decay at constant mixture, the third qualification below. An evaluation at the end of training silently averages this phase in, which is why cosine suites support no duration claim here.

Three measured qualifications. **First**, the three sparse ratios are indistinguishable within their uncertainty, not equal: the paired test could only have detected differences above $\approx 30\%$ (Appendix F). **Second**, SmolLM3's excess is real, $2.12\times$ $[1.86, 2.41]$ the mean of the other three (Apertus on its capped window; $2.17\times$ $[1.91, 2.46]$ uncapped), and it is unexplained. It is not checkpoint coverage: when SmolLM3's dense decay is sub-sampled to each sparse model's coverage, its factor rises to 34.9 - $43.7\times$, away from the sparse models, not toward them. Nor is it the curriculum (SmolLM2-1.7B also switches mixture at its decay onset), the log axis, the denominator window, or the terminal learning rate: SmolLM2's schedules, like SmolLM3's, anneal linearly to zero (Appendix F). **Third**, the harvest has a shape, and it is not the same in every model (Figure 4). Within SmolLM3's decay, where the mixture is constant, the factor climbs from $17.6\times$ over the first third to $61.9\times$ over the last (11.5σ between halves; the shrinking $\log_2$ axis accounts for $\sim 4\%$) - an acceleration *into* the anneal, the direction reported for loss (Hu et al., 2024; Hägele et al., 2024), and the opposite of our own registered prediction. A mixture switch would predict the opposite shape, fast adaptation first. The two SmolLM2 decays rise as well, from ≈ 0 to $25\times$ and from 9 to $21\times$ between halves (6.6σ and 4.6σ), while Apertus shows no rise (0.8σ) and excludes a SmolLM3-sized one at 8.6σ. So the three schedules that anneal linearly to zero accelerate into the anneal and the one that stops at $0.1\times$ its peak does not; with one such schedule, whose decay also follows faster-improving data stages (Appendix F), that is an observation, not a test. Because the SmolLM2 curves accelerate, their late-starting windows over-report their whole-window factors (adding SmolLM2-1.7B's one earlier decay checkpoint lowers $16.7\times$ to $12.3\times$), so the gap to SmolLM3 is, if anything, larger.

8 Domain structure, and a stopping diagnostic

Domain order is a plateau property. At the training plateau (deduplicated Pythia ladder) all ten domains gain at every size, and *which* gain most is nearly the same across a $99\times$ span of parameters (Kendall $W = 0.77$). That concordance dissolves late in training ($W = 0.31$; the final window is excluded as a suspected step-grid artifact, Appendix G). *How much* each domain benefits from scale differs throughout: software is the most size-sensitive (plateau gain $0.0137 \rightarrow 0.0425$ bpb/doubling from 70M to 6.9B) and collapses hardest when small models degrade; literature-classics is its mirror. Our pre-registered epistemic-axis contrast is null in all three windows ($p = 0.25$ - 0.67).

A stopping diagnostic. We are developing a stopping strategy from each model's own noise floor. It reads the end-of-run reading gain in units of that floor and calls a run *improving* above two floors (a two-sigma-like convention we have not tuned), *degrading* below minus two, and *at floor* in between; and it records D^*_{read}, the first checkpoint, after at least one doubling of data, at which the gain falls within two floors of zero. Across 22 series: 18 improving, 2 degrading (exactly the band), 2 at

floor (both BLOOM, whose floors are unusable). Evaluated inside their clean windows, all three clean-window models are still improving at their last clean checkpoint - 5.0, 4.8 and 3.7 floors for SmolLM2-1.7B, SmolLM3 and Apertus - and none reaches D^*_{read} within the window, as registered in advance (Appendix H). Since the per-doubling value of data rises with N (§5), a fixed stopping value implies token budgets that grow with model size; the floor gives each run its own criterion. We offer this as a diagnostic, not yet a rule.

9 LIMITATIONS

Provenance is inherited, not controlled. These are observational trajectories. Peak LR varies across Pythia sizes (one same-LR pair mitigates); SmolLM2-135M/360M mixtures are unverified ("suggestive"); every documented decay coincides with a data-stage switch, and Apertus changes its mixture at 7T, 12T and 13.5T and doubles its batch at 7,038 B, so only its first stage is a clean window; BLOOM's token accounting is unverified and its ladder interpolated. A public checkpoint name is not a checkpoint: one series we scored, Pythia-2.8B on the standard Pile, serves other weights on its intermediate branches and is excluded; a fingerprint audit of every series found two more such cases: a Pythia-1.4B-deduped branch carrying its final weights (its effect on every result is reported and is nil except for one D^*_{read}) and a BLOOM-1.7B checkpoint duplicating another, whose provenance is being traced (BLOOM's results are already labelled direction-only or floor-limited; Appendix A).

The estimator has stated edges. Decay windows are 0.11 - 0.32 doublings with 4 - 13 checkpoints, and their shape statistics rest on a per-checkpoint noise measured in stable windows (in the one decay dense enough to check, the scatter is slightly larger: 1.32 vs. 1.18 floors). Comparison counts are not independent tests (§5); the same-LR control is a single contrast. The additivity slope β is a regression across models, so it has no noise floor of its own and is resolved on its bootstrap confidence interval alone; a floor needs $\gtrsim$8 checkpoints, which BLOOM's series lack.

Scope. Probes are 206 fixed English texts; cosine families support no duration claims; bf16 never pools with fp32. Seen-text verification exists only for Pile-trained models, and the per-token memorisation statistics are the least developed measurement here. The drift is measured against a reference sample whose code is two files from one library; whether its motion belongs to the model or to the sample is tested but not settled. The deduplicated-vs-standard β gap and the SmolLM3 decay excess are robust in our arms and unexplained.

10 CONCLUSION

Contextual information gain splits a held-out loss curve into the part that comes from using context and the part that comes from expecting the text unconditionally, by scoring each token under its context and under the model's own unconditional prediction. Wherever nothing but the token count verifiably changed, held-out cross-entropy keeps falling while the prior moves away from held-out text by re-assignment, not sharpening; after the same number of tokens, larger models improve faster, against the additive scaling law; what stops improving at 70 - 160M is a property of that size band, not of the budget; and the decay phase improves held-out loss many times faster per token, accelerating as the learning rate falls. Within the limits of observational trajectories (§9), the decomposition answers what a loss curve cannot: whether the next factor of two in tokens still improves the use of context, and when that improvement falls within the model's own noise, which bears on when to stop and on how much data a larger model is owed.

AI USE STATEMENT

We used generative AI tools (Claude, Anthropic; Gemini, Google) substantively and record it as ICLR policy asks. They contributed to the experimental design (probe set, model pool, regime rules, estimators, pre-registered predictions), the implementation (measurement runners, integrity tooling, analysis code), the interpretation of results, and the drafting and auditing of this text, including simulated reviews used to stress-test the presentation. They produced no measurement or result on their own. Every run was launched, every number checked against its results file, every reference verified against its source and every decision taken by the authors; assistant errors found on the way

(a prediction registered in the wrong direction, an omitted warm-up guard, an error bar that neglected a shared checkpoint) were fixed and re-verified. We take responsibility for the final content of this work.

Ethics statement

The probe set consists of published text used under its licences (CC BY / CC BY-SA, US public domain, MIT/Apache-licensed source code, and US federal public-domain documents), with per-probe licence strings shipped in the provenance manifest. No human subjects, personal data, or annotation labour were involved. The models evaluated are public releases; we foresee no direct misuse pathway from a measurement method.

Reproducibility statement

The supplementary material contains the measurement library with its self-test and frozen anchor bands, the probe set and both reference priors with their build manifest, per-probe provenance (licences and contamination priors), the per-checkpoint results tables, the exact model revision lists, and an analysis package (scripts and a master notebook) that produces every number and figure of this paper from those tables, with a self-test on synthetic data, a regression against the analysis notebooks of record, a keyed comparison of the final re-run against the results of record, and the per-checkpoint fingerprint audit. All measurements are fp32 with TF32 disabled; the anchor checkpoint is re-scored in every session.

References

Gordon W. Allport. *Personality: A Psychological Interpretation*. Henry Holt, New York, 1937.

Anonymous. The baseline is the question: Degenerate empty contexts in token-level information measures for language models, 2026. Under review.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. SmolLM2: When smol goes big — data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Tyler A. Chang and Benjamin K. Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.

James P. Crutchfield and David P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1):25–54, 2003.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 5988–6008, 2022.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *arXiv preprint arXiv:2405.18392*, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Hugging Face SmolLM team. SmolLM3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`, 2025. Model and intermediate checkpoints: HuggingFaceTB/SmolLM3-3B.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. LLM360: Towards fully transparent open-source LLMs. *arXiv preprint arXiv:2312.06550*, 2023.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, et al. Paloma: A benchmark for evaluating language model fit. *arXiv preprint arXiv:2312.10523*, 2023.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16264.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Tim Pearce and Jinyeop Song. Reconciling Kaplan and Chinchilla scaling laws. *Transactions on Machine Learning Research*, 2024. arXiv:2406.12907.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. *arXiv preprint arXiv:2406.19146*, 2024.

Claude E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30 (1):50–64, 1951.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. *arXiv preprint arXiv:2503.19206*, 2025.

Swiss AI Initiative. Apertus: Democratizing open and compliant LLMs for global language environments. *arXiv preprint arXiv:2509.14233*, 2025.

Mathurin Videau, Badr Youbi-Idrissi, David Lopez-Paz, and Kartik Ahuja. Skaling: Chinchilla's exponents meet Kaplan's coupling. *arXiv preprint arXiv:2608.07222*, 2026.

Wilhelm Windelband. History and natural science. *History and Theory*, 19(2), 1980. Rectorial address, Strasbourg, 1894; translated by Guy Oakes.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Zhiwei Xu, Shihao Wu, Hanseul Cho, Wei Hu, and Yixin Wang. Data-constrained language model pretraining: Improved regularization and scaling laws. *arXiv preprint arXiv:2606.06888*, 2026.

Junxi Yan, Zixi Wei, Jingtao Zhan, Qingyao Ai, and Yiqun Liu. What scales in cross-entropy scaling law? *arXiv preprint arXiv:2510.04067*, 2025.

A Pool, provenance, and exclusions

Table 4: The pool. Chin. = training multiple relative to the Chinchilla-optimal $D \approx 20N$. The last column lists the verified regime boundaries used to define analysis windows.

family	sizes (checkpoints)	schedule	max Chin.	verified regime provenance
Pythia, deduped Pile + 12B bf16 (17)	70M - 6.9B fp32 (41/41/41/41/41/41/17) *1.45 epochs; identical data order*	cosine	214×	peak LRs vs. paper Table 1; identical batch & order
Pythia, standard Pile	14M - 1.4B (41×6, 17) *2.8B excluded (below)*	cosine	1071×	14M/31M non-suite;
SmolLM2	135M (8), 360M (16), 1.7B (41)	WSD	556×	1.7B stages verified; 135M/360M mixture unverified
SmolLM3	3B (47)	WSD	181×	stage-3 decay: 13 checkpoints
OLMo / OLMo2	1B-0724 (38); 1B (20)	cosine	135×	OLMo2 has a 4.4-dbl grid gap
Amber (LLM360)	6.7B (31 of 360)	cosine	9×	1.26T tokens
Apertus	8B (20)	WSD	92×	batch 2× at 7,038 B; stages 7T/12T/13.5T; cooldown from 13.35T, 1-sqrt to 0.1× peak
BLOOM	560M, 1.1B, 1.7B, 3B (6/6/7/7)	cosine	19×	multilingual (EN ≈30%); tokens/step unverified

The seven families comprise 27 model@precision checkpoint series (790 checkpoints; 162,740 probe measurements). OLMo and OLMo2 are counted as one family. Other excluded series: `pythia-160m-deduped@bf16` (single checkpoint; two scoring passes disagreed by 0.1425 bits) and `pythia-410m-deduped@bf16` (precision calibration only). `pythia-12b-deduped@bf16` appears only in the labelled sensitivity arm.

Excluded: Pythia-2.8B on the standard Pile. In the Hugging Face repository `EleutherAI/pythia-2.8b`, the intermediate branches we evaluated either carry a copy of the final checkpoint's weight file (`model.safetensors`, which `transformers` loads by default; identical SHA-256 on `main` and six early branches we checked) or weights whose scores on our probes match Pythia-2.8B-deduped to within 10^{-4} bits per byte at all seven checkpoints from 140 B tokens (0.08 of the deduplicated twin's noise floor). The issue is also reported publicly on the model page.[1] A per-checkpoint fingerprint audit of every evaluated series (checkpoints whose scores agree to 10^{-4} bits on every probe) found the same signature twice more: Pythia-1.4B-deduped's step-26000 branch scores identically to its final checkpoint (step 143000), one checkpoint of 41, whose effect on every analysis is reported in Appendices B and H and which a GPU re-score confirmed to be what the released branch gives; and one BLOOM-1.7B checkpoint that duplicates another, whose provenance is being traced. The registered prediction that the audit would flag only Pythia-2.8B (prediction 25) is therefore refuted. Anyone who loaded intermediate `EleutherAI/pythia-2.8b` checkpoints through `transformers` with default file selection received the final model on the early branches. The exclusion removes an eight-size standard ladder, one corpus-effect cell and one contamination row from an earlier version of our analysis; no result above depends on it.

Verified regime boundaries. Pythia peak LRs match the suite paper's Table 1 (10^{-3} to 1.2×10^{-4}, all decaying to $0.1\times$). SmolLM2-1.7B stages from its report (web 0 - 6T; 6 - 8T; 8 - 10T; decay with a new mixture 10 - 11T, LR linearly to zero over the last 10% of steps) (Ben Allal et al., 2025). SmolLM3 stages from its report and branch names (stage 3, 10 - 11.1T, is the decay: linear to zero over the final 10%, one mixture throughout) (Hugging Face SmolLM team, 2025). Apertus from its launch script (peak 1.1×10^{-4}, minimum 1.1×10^{-5}, 1-sqrt cooldown of 403,905,806 samples $=$ 1,654 B tokens, onset 13,346 B; batch 1024 $\rightarrow$ 2048 sequences after 1,718,272,000 samples $=$ 7,038 B, consistent with the final checkpoint's step count) and report (first data stage to 7T, a further switch at 12T, and a cooldown coinciding with a switch to the highest-quality sources at 13.5T; the report's text gives "8T" for the batch change, which its own step count contradicts) (Swiss AI Initiative, 2025). BLOOM revision names give token counts, with tokens-per-step unverified.

B ESTIMATORS, FLOORS, AND PRECISION

Local gain. OLS of the probe-mean context term on $\log_2$(tokens) over windows of ≥ 4 checkpoints spanning ≥ 0.8 doublings (decay phases: plain OLS over the window). Local windows start at ≥ 15.7 B tokens in every additivity test, clear of the warm-up$\rightarrow$cosine kink at 3 B. **Floor.** $1.4826 \cdot \mathrm{median}\,|\Delta^2 y - \mathrm{median}\,\Delta^2 y|/\sqrt{6}$ on the probe-mean curve. **Bootstrap.** Two-level (resample probes, resample checkpoints), 95% percentile intervals, joint across terms so ratios such as drift over gain carry their own intervals; the harvest ratios use a paired probe bootstrap. **Harvest shape.** Point-to-point factors divide a checkpoint difference by a tiny $\log_2$ step and are noise-dominated; block factors over quarter- or half-windows depend only on their two endpoint checkpoints and carry an analytic checkpoint-noise error bar, taken in §7 as $\sqrt{6}$ floors per checkpoint. **Noise calibration.** The floor already estimates the per-checkpoint scatter: for independent checkpoint noise the second differences have $\sqrt{6}$ times its standard deviation. Measured directly, as the residual scatter about a smooth (quadratic) fit inside each smooth window, the per-checkpoint scatter is 0.66, 1.18 and 0.59 floors in the three clean windows (SmolLM2-1.7B, SmolLM3, Apertus), 1.32 floors in SmolLM3's decay, and 0.18 - 0.76 floors in Pythia's plateau windows; the pooled median over stable windows is 0.44 (registered as ≤ 1.6, prediction 22, Appendix I), with no sign of correlation between neighbouring checkpoints. So the floor itself overstates the per-checkpoint scatter, and the $\sqrt{6}$ convention overstates it by 2 - $4\times$ in the WSD windows. The one exception, Pythia-1.4B-deduped's plateau window at 9.9 floors, is a single checkpoint: its step-26000 branch sits 35 floors below its neighbours on every probe, a GPU re-score with the reference runner reproduces the stored scores to $< 5 \times 10^{-4}$ bits (the weights, not our measurement), and the fingerprint audit shows those scores are identical to its final checkpoint's - the same repository failure as Pythia-2.8B's, confined to one checkpoint (Appendix A). Blanking it changes no β of the deduplicated ladder (all 13 steps unchanged), moves the plateau Kendall W by 0.001, leaves the late window of §6 untouched, and removes only the D^*_{read} that checkpoint had created (Appendix H); the pooled median above includes that window. The acceleration statistics of §7 therefore use each WSD model's own measured scatter and account for the

[1] `https://huggingface.co/EleutherAI/pythia-2.8b/discussions/5`

checkpoint the two half-windows share (which the original error propagation neglected, overstating significance by $\approx 1.2\times$): on that basis SmolLM2-360M, SmolLM2-1.7B, SmolLM3 and Apertus accelerate at 4.6, 6.6, 11.5 and 0.8σ, and Apertus (capped window) excludes a SmolLM3-sized acceleration at 8.6σ; with one floor per checkpoint instead, 2.7, 4.3, 13.6, 0.5 and 5.1σ; with the $\sqrt{6}$ convention, 1.1, 1.8, 5.5, 0.2 and 2.1σ. Floors taken by checkpoint index on unevenly spaced grids also leak part of the real trend into the floor; a leak-free floor is $0.70\times$ the published one for Apertus, 1.00 - $1.10\times$ for the SmolLM windows and 0.32 - $0.75\times$ on Pythia's plateau grid. Both effects make every floor-based verdict in this paper conservative. **Prior entropy test.** From each checkpoint's stored $\bar{q}_B$ we compute $H(\bar{q}_t)$ and $\mathrm{KL}(\bar{q}_t \| \bar{q}_{\mathrm{final}})$, re-temper $\bar{q}_t \propto \bar{q}_t^{\beta}$ to the final entropy, and report the share of KL that remains. **Precision.** bf16 damaged the 1800-token conditional term by $+0.76$ bits at the anchor while leaving the reference in band - position-dependent damage invisible to evaluations that only read short contexts - and two bf16 passes over the same checkpoint disagreed by 0.1425 bits. The 410M bf16-vs-fp32 *slope* difference was below that model's floor ($0.0075 < 0.0208$ bits/token), which is why a bf16 sensitivity *ladder* is meaningful although bf16 *levels* never pool with fp32. We include Pythia-12B @ bf16 deliberately: bf16 is how models this size are typically deployed, and an instrument proposed for practical use should have its behaviour there on record, larger noise included. Practical reading: at the precision most large models are served in, per-checkpoint levels of a long-context information measure are not trustworthy, but cross-model slope structure survives.

C ALL DURATION AND SCHEDULE SLOPES

The full per-regime slope table (every model, warmup/stable/cosine/decay, context and both reference terms, with CIs and resolution flags) ships as machine-readable output. Figure 5 shows the reference term over training, Figure 6 each series' drift in its longest regime, and Figure 7 every checkpoint's local reading gain on the training-multiple axis. **Unresolved constant-LR windows.** Three short mid-training windows outside the clean set are not resolved: SmolLM2-1.7B stages 2 and 3 ($+0.0071$ $[-0.0008, +0.0190]$ and -0.0167 $[-0.0372, +0.0013]$ bpb per doubling) and SmolLM3 stage 2 (-0.0027 $[-0.0277, +0.0082]$), each 5 - 7 checkpoints over about a quarter-doubling, which the estimator cannot resolve. **Reference-sample mechanism.** As a model reads its reference documents better, $\bar{q}$ moves toward what those documents continue with, and away from held-out text unlike them, through the reading gain itself. Two observations fit this reading: the drift is larger against the narrower $\bar{q}_A$ than against $\bar{q}_B$ in every clean window (1.15 - $1.89\times$), and in SmolLM3 the ratio of drift to gain is the same in the stable window and in the decay (0.241 vs. 0.234) while the gain rises $35.8\times$. The registered test (prediction 27) is the one in §4. **Per-domain drift (descriptive).** Against $\bar{q}_B$, a permutation test over domain labels finds domain structure in the drift in Apertus and SmolLM3 ($p < 0.001$, the test's resolution) and not in SmolLM2-1.7B ($p = 0.09$); with the software domain excluded it finds structure in all three ($p < 0.001$ each; registered as prediction 26). The prior moves its mass away from code tokens fastest ($3.2\times$ and $3.0\times$ the median domain in Apertus and SmolLM3; $1.1\times$, unresolved, in SmolLM2-1.7B). Domain structure does not exclude sharpening - a temperature change moves each domain at its own rate - so the sharpening answer rests on the entropy test of §4 alone; and since $\bar{q}_B$'s code is two files from one library while the software probes come from eight others, the code excess may reflect how narrowly the reference samples code. We claim that the drift differs by domain, not which domains drift fastest.

D ADDITIVITY: ROBUSTNESS DETAIL

Standard ladder without the excluded series. With Pythia-2.8B excluded and 1.4B added, the standard-Pile ladder at the 13 matched steps (7 sizes, 14M - 1.4B) is resolved positive at 13 of 13 with median $+0.0014$, as registered (prediction 24); the 6-size dense ladder of Table 2 is unchanged. **Non-embedding** N (the Pythia authors' recommendation): medians $+0.00408$ (deduplicated, 13/13) and $+0.00065$ (standard, 31/31); no resolved comparison changes sign. **Corpus gap.** Deduplicated over standard median β: $4.6\times$ on total N, $6.3\times$ on non-embedding N. **Same-peak-LR pairs.** At every matched checkpoint we form $\beta_{\mathrm{pair}} = (g_{\mathrm{large}} - g_{\mathrm{small}}) / \log_2(N_{\mathrm{large}}/N_{\mathrm{small}})$ with a paired probe bootstrap. Resolved-positive/negative contrasts and medians (per doubling of N; the 6.9B/12B pair mixes precisions):

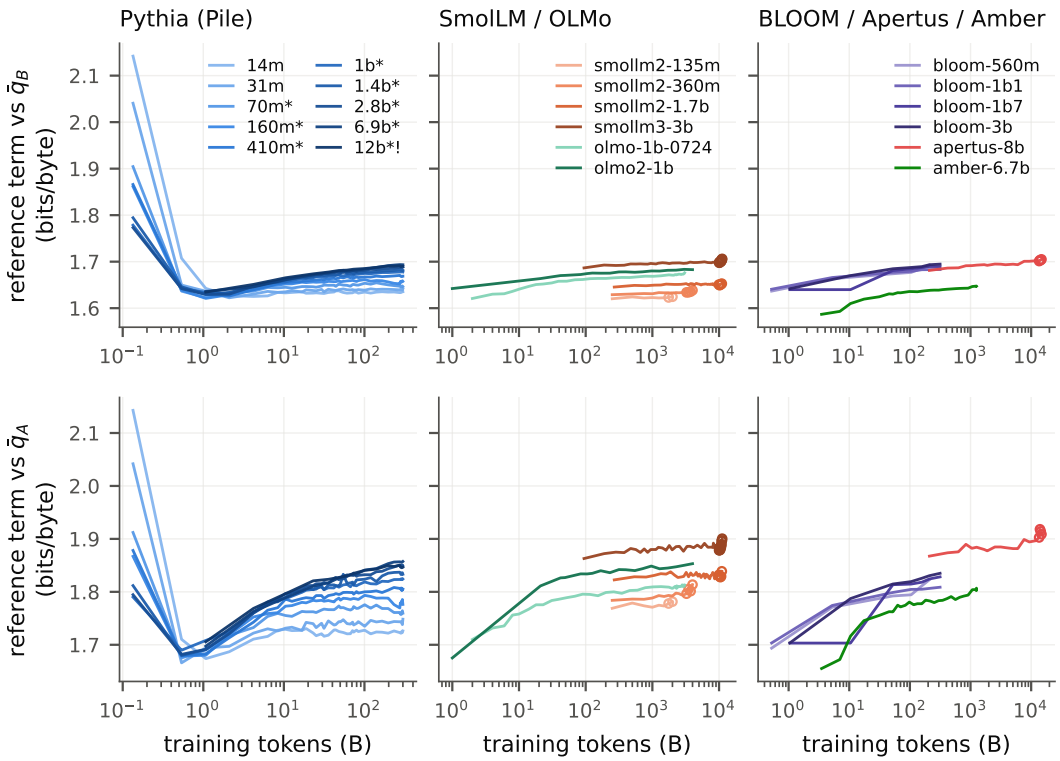

Figure 5: Probe-mean reference term over training against $\bar{q}_B$ (top; 12 documents, headline) and $\bar{q}_A$ (bottom; 3 Wikipedia documents), by family; hollow rings mark WSD decay checkpoints. Standard-Pile Pythia twins are hidden except 14M and 31M.

pair	as registered ($\geq$1.07 B)	guarded ($\geq$15.7 B), post hoc	past 100 B
410M/1B, deduped	18/11 of 38, med. $+0.0005$	18/4 of 31	13/0 of 13, med. $+0.0065$
410M/1B, standard	26/10 of 38, med. $+0.0010$	26/3 of 31	13/0 of 13, med. $+0.0025$
6.9B fp32/12B bf16‡	1/8 of 16, med. -0.0015	1/8 of 13	0/7 of 8, med. -0.0026

The registered criteria, and their outcomes as coded: prediction 13 (majority resolved-positive over all matched checkpoints, both corpora) was not confirmed - the deduplicated pair is positive at 18 of 38. The evaluation as coded started at 1.07 B tokens and omitted the warm-up guard every other additivity test applies ($\geq$ 15.7 B); with the guard (a post-hoc correction of that porting defect) both pairs are majorities (18/31, 26/31; medians $+0.00099$, $+0.00152$). Prediction 14 (median within $3\times$ of the ladder) was not confirmed either way, because the registered test compared medians over two different token ranges. Past 100 B - the scope of the headline, fixed before this control ran - both fp32 pairs are 13/13 resolved positive, from 9 distinct local windows. Just after the guard, the first local window (16.8 - 29.4 B) is resolved negative on both corpora (-0.0008 and -0.0018 per $\log_2 N$). The mixed-precision 6.9B/12B pair is negative past 73 B (7 of 8 past 100 B, median -0.0026) but only -0.9 noise floors late and confounded with bf16's non-reproducibility; it is disclosed, not interpreted, while the mixed-precision *ladder* over eight sizes remains 13/13 positive because its regression is dominated by the 70M$\rightarrow$6.9B span.

E Degradation, repetition, and seen text: tables

Late gains and corpus effect (bpb/doubling; late window = tokens $\geq 0.55\times$ final; floor ratios use the deduplicated twin's floor):

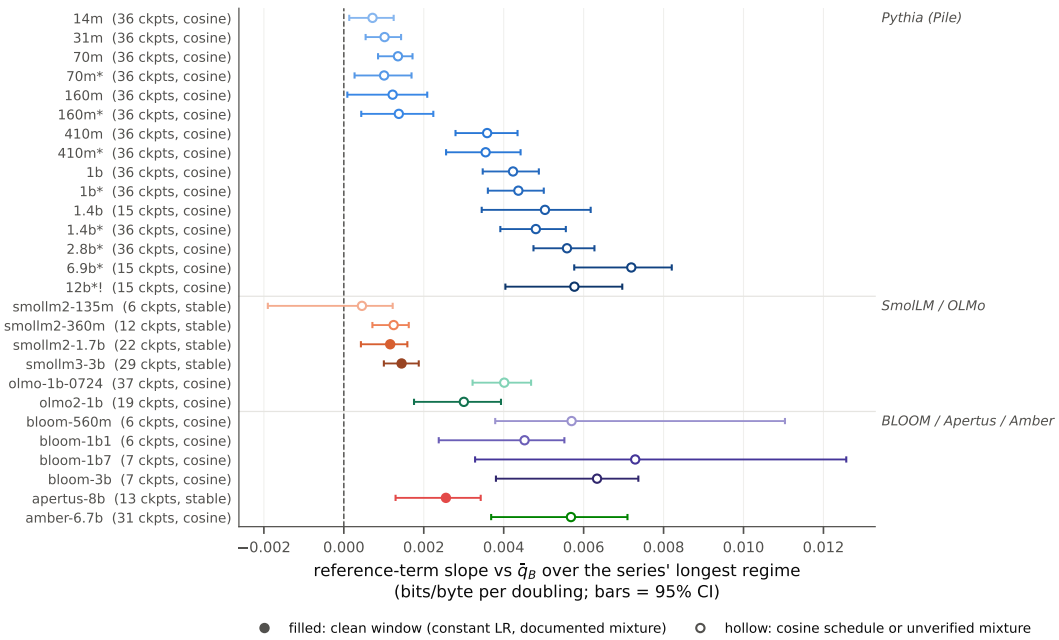

Figure 6: Reference-term slope against $\bar{q}_B$ over each series' longest regime (Apertus: capped window) with 95% CIs; filled = clean window. All 27 slopes are positive; 26 intervals exclude zero (not SmolLM2-135M).

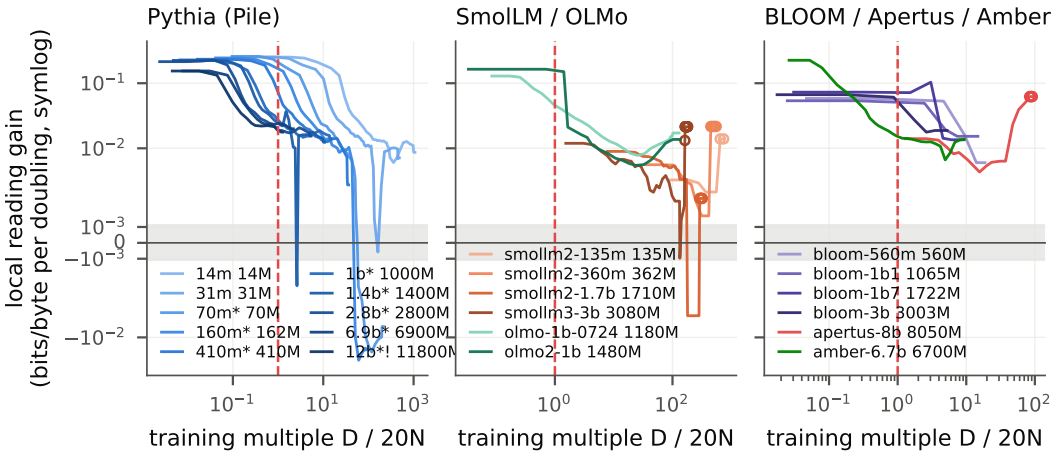

Figure 7: Local reading gain per doubling against the training multiple $D/20N$ for every checkpoint, by family (log-symlog). A universal curve would collapse the lines; they do not. WSD local gains are fitted inside each checkpoint's own regime, with the whole-curve window where a regime is too short for the estimator (every decay phase; SmolLM3's stage 2; Apertus's post-cap stable checkpoints); hollow rings = LR-decay checkpoints; grey band = median noise floor.

N	standard (0.9 ep)	deduped (1.45 ep)	corpus effect	÷ dedup floor
70M	+0.0026	−0.0068	−0.0094	8.38
162M	−0.0105	−0.0151	−0.0045	3.56
410M	+0.0105	+0.0036	−0.0069	5.11
1000M	+0.0137	+0.0141	+0.0004	0.38
1400M	+0.0112	+0.0102	−0.0010	0.99

Per-domain corpus effect (bpb/doubling; paired probe-bootstrap SE in parentheses; ° marks the two CIs that include zero):

domain	70M	160M	410M
biology	$-0.0103\,(0.0012)$	$-0.0056\,(0.0021)$	$-0.0080\,(0.0008)$
biomedicine	$-0.0084\,(0.0011)$	$-0.0022\,(0.0009)$	$-0.0043\,(0.0006)$
economics	$-0.0101\,(0.0011)$	$-0.0046\,(0.0008)$	$-0.0066\,(0.0008)$
government-regulation	$-0.0105\,(0.0013)$	$-0.0023\,(0.0012)^{\circ}$	$-0.0059\,(0.0007)$
history	$-0.0087\,(0.0011)$	$-0.0051\,(0.0012)$	$-0.0068\,(0.0009)$
literature-classics	$-0.0073\,(0.0010)$	$-0.0027\,(0.0008)$	$-0.0042\,(0.0010)$
philosophy	$-0.0090\,(0.0012)$	$-0.0048\,(0.0017)$	$-0.0052\,(0.0006)$
physics	$-0.0084\,(0.0014)$	$-0.0042\,(0.0016)$	$-0.0061\,(0.0008)$
sociology	$-0.0054\,(0.0010)$	$-0.0029\,(0.0007)$	$-0.0059\,(0.0006)$
software	$-0.0167\,(0.0027)$	$-0.0119\,(0.0076)^{\circ}$	$-0.0162\,(0.0033)$

Negative in 10/10 domains at each size. The registered verdict is a domain-label permutation test, calibrated at 8.3% false alarms where Cochran's Q gives 13.3%: it detects structure at 70M ($p = 0.0005$) and 410M ($p < 0.001$, the test's resolution) and not at 160M ($p = 0.38$); the Bonferroni threshold over three sizes is 0.0167. As a supplementary, anti-conservative description, $Q = 27.6$, 12.0, 29.3 on 9 df at 70M/160M/410M. Code's excess over the median domain has $z = -3.0$, -1.1, -3.1. Without code the residual heterogeneity is within Q's miscalibration; we do not claim a second structured domain. **Verifiable contamination** (level gap at final checkpoint, seen minus unseen; late-gain difference is reported only for Pythia):

model	level gap (bpb)	late-gain diff
Pythia-14m	-0.0559	-0.0016
Pythia-31m	-0.0561	-0.0007
Pythia-70m	-0.0444	-0.0009
Pythia-70m-deduped	-0.0450	$+0.0007$
Pythia-160m	-0.0297	$+0.0012$
Pythia-160m-deduped	-0.0323	$+0.0002$
Pythia-410m	-0.0185	$+0.0027$
Pythia-410m-deduped	-0.0197	$+0.0018$
Pythia-1b	-0.0125	$+0.0011$
Pythia-1b-deduped	-0.0111	$+0.0001$
Pythia-1.4b	-0.0078	$+0.0000$
Pythia-1.4b-deduped	-0.0083	$+0.0012$
Pythia-2.8b-deduped	-0.0035	$+0.0026$
Pythia-6.9b-deduped	-0.0014	$+0.0050$
Pythia-12b-deduped	-0.0025	$+0.0111$

Apertus's late-gain difference is excluded wherever gains are compared: it is the decay leaking through a provenance-blanking rule, not contamination. Recency gaps (positive = 2026 text easier): SmolLM2-1.7B $+0.004$, SmolLM3 $+0.011$, Apertus $+0.011$, Amber $+0.012$. BLOOM Gini differences (seen − unseen concentration; *not* a verifiable split): $+0.038$, $+0.025$, $+0.026$, $+0.041$ for 560M/1.1B/1.7B/3B; the Pythia series show no such excess.

F Decay phase: measurements behind §7

Terminal learning rates (documentation, not explanation): SmolLM3, linear to 0 over the final 10%; SmolLM2-1.7B, linear to 0 over the final 10%; SmolLM2-360M, linear to 0 over the final 20% (released configuration); Apertus, 1-sqrt to $0.1\times$ peak over the final 11%. SmolLM3 and the two SmolLM2 models share schedule form and depth, yet differ by $2.1\times$ in level, so the decay floor cannot explain the SmolLM3 excess; a hypothesis that it did, written down before the schedules were checked, is recorded as falsified (Appendix I). **Coverage resampling.** SmolLM3's 13-point decay resampled at each sparse model's fractional window and checkpoint count: like SmolLM2-360M $\rightarrow$ $38.9\times$ $[36.1, 42.0]$; like SmolLM2-1.7B $\rightarrow$ $43.7\times$ $[40.5, 47.0]$; like Apertus $\rightarrow$ $34.9\times$ $[32.0, 38.0]$. All move away from the sparse cluster (median gap closure -17%). **Window definition matters.** Starting SmolLM2-1.7B's decay at the LR onset (9,900 B, 5 checkpoints) instead of the mixture onset (10,000 B, 4) gives $12.25\times$ $[8.75, 15.72]$ instead of $16.72\times$: one checkpoint moves the ratio by a quarter, so every factor carries its window. **Half-window factors** (second minus first, $\pm 1\sigma$ from each model's measured per-checkpoint scatter, shared checkpoint accounted for): SmolLM3

Table 5: The stopping diagnostic per series: end-of-run reading gain in units of the model's own floor, its status, the training multiple at the end, and D^*_{read} where it is reached. ^{w} evaluated inside the clean window; † see the text.

model	status	gain/floor	end ×Chin	D^*_{read}
pythia-14m	improving	+12.5	1071.0×	-
pythia-31m	improving	+5.1	483.7×	84 B (135×)
pythia-70m-deduped	degrading	-6.1	214.2×	73 B (52×)
smollm2-135m	improving	+16.8	555.7×	-
pythia-160m-deduped	degrading	-11.8	92.6×	-
smollm2-360m	improving	+16.8	414.4×	2250 B (311×)
pythia-410m-deduped	improving	+2.7	36.6×	-
bloom-560m	at floor	+0.3	18.7×	105 B (9×)
pythia-1b-deduped	improving	+13.8	15.0×	-
bloom-1b1	at floor	+0.6	14.8×	52 B (2×)
olmo-1b-0724	improving	+19.2	129.2×	-
pythia-1.4b-deduped	improving	+11.8	10.7×	71 B (3×)†
olmo2-1b	improving	+20.4	135.2×	-
smollm2-1.7b^{w}	improving	+5.0	172.6×	-
bloom-1b7	improving	+2.7	9.1×	-
pythia-2.8b-deduped	improving	+13.4	5.4×	-
bloom-3b	improving	+2.2	5.2×	-
smollm3-3b^{w}	improving	+4.8	131.8×	-
amber-6.7b	improving	+25.9	9.4×	-
pythia-6.9b-deduped	improving	+7.9	2.2×	-
apertus-8b^{w}	improving	+3.7	37.4×	-
pythia-12b-deduped	improving	+7.1	1.3×	-

$17.4 \to 56.6$ ($+39.2 \pm 3.4$); SmolLM2-1.7B $-1.5 \to 24.8$ ($+26.3 \pm 4.0$); SmolLM2-360M $9.4 \to 21.4$ ($+12.0 \pm 2.6$); Apertus $20.2 \to 16.3$ (-3.9 ± 5.0; capped stable gain). **Pairing.** Pairing the probe resample across models narrows the pooled-difference interval by only $1.06\times$, so probe error is nearly independent across models. **SmolLM3 over the mean of the three** uses the same 2,000 shared probe resamples with either Apertus window; only Apertus's stable window differs between $2.12\times$ $[1.86, 2.41]$ (capped) and $2.17\times$ $[1.91, 2.46]$. **Log axis.** SmolLM3's uniform-in-tokens grid shrinks $\log_2$ intervals by 10.3% across the window; the observed rise is 252%. **Apertus before its cooldown.** Apertus's last pre-cooldown segment (11,852→13,532 B, which spans its 12T data switch and ends 186 B into the cooldown) already reads at the decay-phase rate: 18.3 ± 1.5 times the capped stage-1 gain (one floor per checkpoint, conservative for Apertus), against 20.2 ± 6.3, 12.8 ± 6.5 and 19.8 ± 6.7 for the three segments inside the cooldown. Even crediting the cooldown with an instantaneous response, the pre-onset part of that segment reads 13.4 ± 2.2 times the stage-1 gain (17.0 ± 1.6 if the cooldown is taken to start at the report's 13.5T), and it alone is worth 48% (67%) of the whole stage-1 gain. The local gain at the end of the capped window already shows it: fitted inside the window it is 0.0063 bpb per doubling, fitted over neighbouring checkpoints that extend past the batch doubling and data switch it is 0.0207. SmolLM2-1.7B's and SmolLM3's last pre-decay stages show no resolved gain. In Apertus the decay-phase factor therefore cannot be separated from its preceding data stages.

G Domains: concordance, axis test, and one excluded window

Kendall W across the deduplicated ladder, by window: plateau 0.77 → late 0.31 → tail 0.11. The tail window (final $\sim 8\%$ of training) is excluded from claims: it contains a uniform sign inversion between adjacent sizes (160M positive in 10/10 domains while 410M is negative in 10/10), which we treat as a suspected step-grid artifact. Epistemic-axis medians (nomothetic vs. idiographic, bpb/doubling): plateau 0.0212 vs. 0.0204 (p=0.67); late 0.0101 vs. 0.0125 (p=0.25); tail 0.0036 vs. 0.0068 (p=0.34).

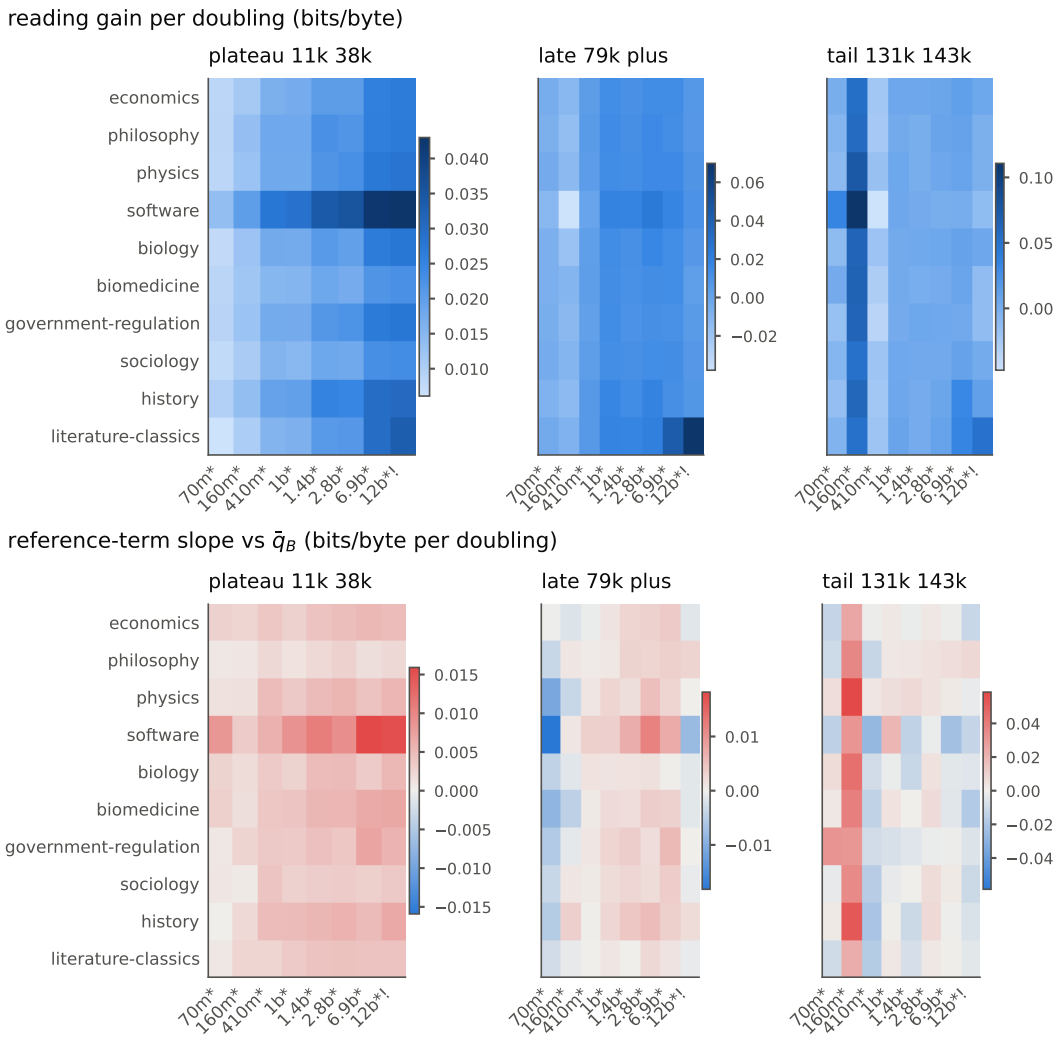

Figure 8: Per-domain reading gain (top) and reference-term slope against $\bar{q}_B$ (bottom) across the deduplicated ladder, by window; colour = value, the tables are in the supplement.

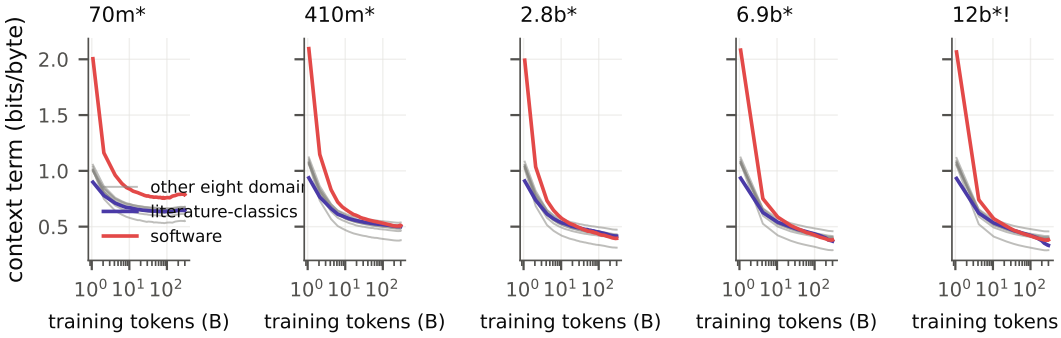

Figure 9: Domain trajectories (context term, bpb) for five sizes: software (red) and literature-classics (violet) against the other eight domains (grey).

H STOPPING DIAGNOSTIC

Table 5 lists 22 series: the 27 less the five standard-Pile twins of deduplicated models, which share their grids. Statuses as defined in §8: improving above two floors, at floor within two floors of zero, degrading below minus two; D^*_{read} is the first checkpoint after one doubling at which the gain is

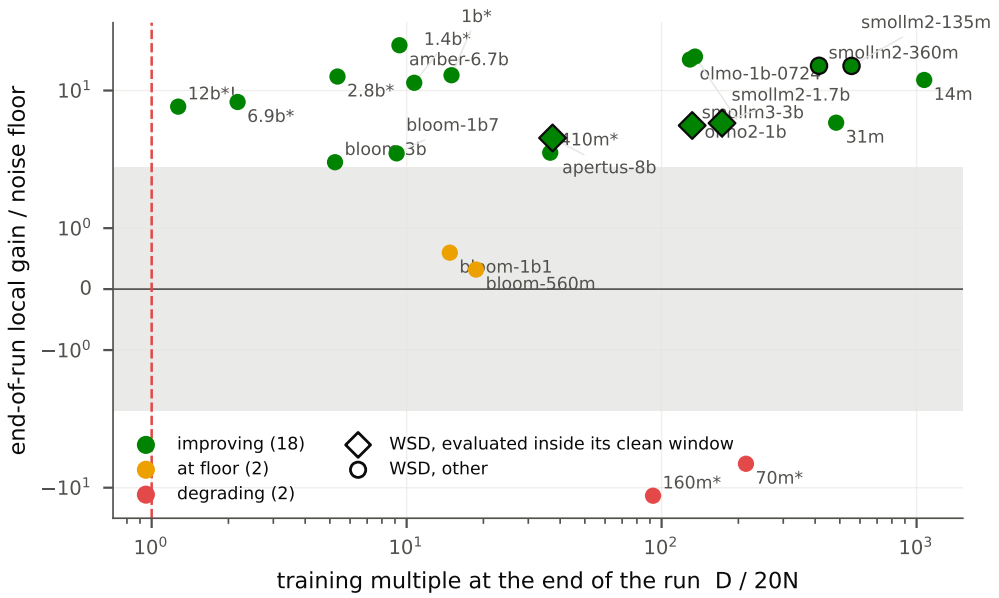

Figure 10: End-of-run gain in units of each model's own floor vs. training multiple $D/20N$ (grey $= \pm 2$ floors). Red: the band. Orange: BLOOM's floor artifact. Models at 484 - 1071$\times$ still improve; the three clean-window models (diamonds) are evaluated inside their windows.

within two floors of zero. †Pythia-1.4B-deduped's D^*_{read} is created by its step-26000 checkpoint, which carries the final model's weights (Appendix A); without it, D^*_{read} is not reached and the end-of-run status is unchanged. The three clean-window models (^{w}) are judged at the last checkpoint of their clean window, with the local gain fitted inside the window only; the last 1 - 5 in-window points share one saturated local window, so the end value is the window's final local gain. If the local window is instead allowed to reach across the next stage boundary, the end gain changes in either direction: 6.0 floors for SmolLM2-1.7B, 0.4 for SmolLM3 (its stage 2 reads slower) and 12.1 for Apertus (its next stage reads faster, §F). An earlier evaluation pooled every constant-LR regime of these models into one window (ending at 9,657, 9,912 and 9,332 B) and gave 3.5, 12.9 and 22.9 floors with D^*_{read} at 6,974 and 8,024 B; those values mix data stages and are superseded. BLOOM floors are lower bounds, so its two AT FLOOR labels are artifacts of floor quality, not saturation.

I PRE-REGISTRATION LEDGER

Predictions were written down, dated, before the analysis that tests them. A prediction can pass while the reasoning written beside it fails; we grade the two separately.

#	prediction	verdict
1	corpus effect within ± 2 floors at 1.4B and 2.8B	confirmed at 1.4B (0.99 floors); the 2.8B cell is void (excluded series)
2	$\beta > 0$ at every step of an 8-size standard ladder	partly refuted as run; the ladder contained the excluded series and is void; prediction 24 retests
3	if SmolLM3's CI excludes ~ 16.7, drop "constant"	triggered ($35.8\times$)
4	Apertus ratio moves $> 20\%$ across assumed onsets	refuted (4.9%); moot once the onset was found
5	$\beta > 0$ on BLOOM	confirmed (8/8)
6	no late BLOOM degradation	confirmed, with the floor caveat
7	harvest factor varies $\geq 2\times$ within a decay	confirmed
8	harvest steepest at the onset	refuted, reversed: it rises into the anneal
9	coverage explains $\geq 30\%$ of the sparse - dense gap	refuted with the wrong sign (-17%)
10	three sparse ratios indistinguishable	passes without power (detects $\geq 29.5\%$)
11	pairing narrows CIs $\geq 2.5\times$	refuted ($1.06\times$)
12	SmolLM3's factor over the mean of the other three excludes 1	confirmed, $2.17\times$ $[1.91, 2.46]$ as registered; $2.12\times$ $[1.86, 2.41]$ with Apertus capped
13	same-LR β majority-positive	not confirmed as coded (App. D)
14	same-LR β within $3\times$ of the ladder's	not confirmed as coded (App. D)
15	domain structure at any size removes "uniform"	fired at 70M and 410M
16	joint CIs reproduce; drift-to-gain CIs exclude 0	confirmed
17	capped Apertus gain within $\pm 10\%$, resolved	confirmed (-7.3%)
18	domain-structured drift in ≥ 1 clean window	confirmed (2 of 3); the rationale ("so not sharpening") fails
19	prior entropy falls and rotation dominates	passes as registered; entropy change unresolved, sharpening not detected
20	no single checkpoint explains Pythia-2.8B's flat window	passes; the cause is the repository (App. A)
21	terminal LR explains SmolLM3's excess	falsified by the SmolLM2 schedules
22	pooled per-checkpoint scatter inside smooth windows (median over stable windows) is at most 1.6 floors, i.e. the $\sqrt{6}$ convention overstates it	confirmed (0.44; 0.59 - 1.32 in the WSD windows); the one 9.9-floor window is a duplicated checkpoint (App. B)
23	all three clean-window models still improve at the end of their window	confirmed (5.0, 4.8, 3.7 floors in-window)
24	standard ladder without 2.8B (7 sizes, 13 steps) has no resolved negative	confirmed (13/13, median $+0.0014$)
25	the fingerprint audit flags only Pythia-2.8B	refuted: Pythia-1.4B-deduped step 26000 = its final checkpoint; one BLOOM-1.7B duplicate, being traced
26	drift structure survives with code excluded, in all three clean windows	confirmed ($p < 0.001$ each)
27	extra $\bar{q}_A$ drift sits in the domains $\bar{q}_A$ lacks, ≥ 2 of 3 windows	refuted (one-sided $p = 0.06, 0.31, 0.66$; contrasts $+0.0007, +0.0003, -0.0002$)
28	no temperature removes $> 5\%$ of the prior's motion	confirmed (at most 1.7%)
29	Apertus's three unscored stage-4 checkpoints read $\geq 8\times$ its stage-1 gain before the cooldown	registered, not run (fp32 needs an A100)