WHAT IS A REPEATED TOKEN WORTH? BEYOND TOKEN COUNTS IN LANGUAGE MODEL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data reuse in language-model pretraining is usually summarized by token counts: processed tokens, unique tokens, passes and tokens per parameter. We ask whether these counts determine what a repeated token is worth, and find that they do not, using single-seed sweeps over model sizes up to 2B parameters and pre-registered multi-seed paired studies on three data sources. First, the value of a repeat depends on the comparison: relative to stopping after one pass, a second pass is worth most of a fresh one (an estimated 85%, and at least 63% within measurement resolution), yet at a fixed compute budget, replacing fresh tokens with repeats raises loss for every model in a five-size sweep and mean loss for every pre-registered allocation on all three sources. Second, equal counts do not give equal loss: with all counts matched, concentrating repeats on fewer samples raises loss slightly but consistently, and replaying each shard consecutively rather than in spaced passes raises loss at all three sizes tested, by up to 0.46 bpb. Third, scaling depends on which count is held fixed: at fixed unique data, across eight model sizes, larger models reach their loss minimum after fewer passes (about 15 down to 4 over a 16-fold size range), whereas at matched tokens per parameter the larger of two models loses less to repetition. A count-based loss surface fitted to smaller models predicts larger ones where reuse helps, but below its fitted range of unique data it predicts a benefit from repetition where every measured case gets worse. Finally, in a 60-model study, re-tokenizing repeats with BPE-dropout lowers loss at high repetition but not at matched compute. Reuse claims should therefore report the comparison, the allocation and order of repeats, and the tokenization, not only aggregate counts.

1 INTRODUCTION

Training datasets are growing faster than the stock of public human-generated text (Villalobos et al., 2024), so pretraining increasingly reuses data. Reuse is usually described by token counts—processed tokens D, unique tokens U, passes R and tokens per parameter (TPP)—which effective-data formulations map to a diminishing effective dataset size (Muennighoff et al., 2023). We ask whether these counts are sufficient statistics for what a repeated token is worth.

We find that the counts leave four properties of reuse undetermined: (i) *The comparator*: *fixed-corpus repetition* compares extra passes with stopping at fixed model size N and U (adding processed tokens and compute), whereas *fresh-token replacement* compares repeats with fresh tokens at fixed N and D (Figure 1). (ii) *Allocation*: two allocations can match processed tokens, unique content and total extra exposures, yet concentrate those exposures on different subsets. (iii) *Order*: when during training the repeats are presented. (iv) *Tokenization*: whether identical bytes are presented as identical or varied tokens. Our findings cover these properties and how their effects change with model size (paired studies give intervals; sweeps are descriptive; Appendix Table 1):

- **Whether repetition helps depends on the comparator (§4).** For a 59M model and relative to early stopping, a second pass is worth an estimated 85% as much as the same number of fresh tokens (63–112% within the 0.002-bpb resolution), yet at a fixed compute budget, replacing fresh tokens with repeats raises mean loss in all nine pre-registered allocations on each of three data sources, with penalties that grow during training.

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.

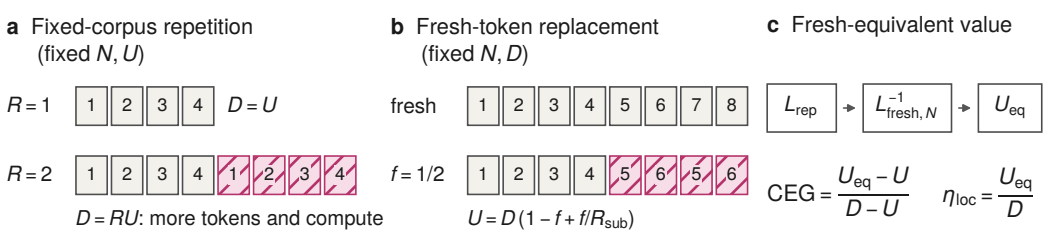


Figure 1: **A repeated token can be compared with stopping or with a fresh token.** (a) Fixed-corpus repetition keeps model size N and unique tokens U fixed, so extra passes add processed tokens D and compute. (b) Fresh-token replacement keeps N and D fixed; a fraction f of processed tokens comes from a repeated subset seen R_{sub} times. (c) A fresh-equivalent budget U_{eq} inverts a same-model fresh-data reference curve; it defines CEG and the whole-run fraction η_{loc} (Eq. 1).

- **Equal counts, unequal loss (§5).** With counts matched, concentrating the extra exposures rather than spreading them raises loss with a consistent sign across sources and replicates; replaying each shard consecutively rather than in spaced passes raises loss by up to 0.46 bpb.
- **Scaling trends depend on what is held fixed (§6).** At fixed unique data U, larger models reach their loss minimum after fewer passes (about 15 at 127M, about 4 at 2B); at matched TPP, the heavy-replacement penalty is smaller at 184.6M than at 45.1M. A count-based surface fitted to smaller models predicts larger ones where reuse helps, but not at unique budgets below its fitted range.
- **Tokenization matters (§7).** Re-tokenizing repeated byte sequences with BPE-dropout lowers the 16-pass loss by 0.154 bpb but costs 0.021 bpb at one pass and is not compute-matched.

2 Related Work

Reuse under data constraints. Compute-optimal scaling allocates compute between parameters and fresh tokens (Kaplan et al., 2020; Hoffmann et al., 2022). Data-constrained scaling assigns repeated tokens a diminishing effective value (Muennighoff et al., 2023), and multi-epoch degradation has been reported to worsen with model size (Xue et al., 2023). Lovelace et al. (2026) already add a penalty proportional to $(R-1)N/U$ (with embeddings in N), and Qin et al. (2026) model token effectiveness across model size, fresh tokens per parameter and repetition; their saturation parameter is an effective-data ceiling, not a loss argmin. Yan et al. (2026) analyze reuse efficiency in linear regression, where N denotes dataset size. These works mostly compare repeated passes over a fixed corpus against a fresh-data reference; we separate that comparison from fixed-budget fresh-token replacement and report paired uncertainty.

Partial repetition and exposure structure. Repeating a data subset can degrade models (Hernandez et al., 2022), and deduplication improves them (Lee et al., 2022). The closest fixed-budget comparator (Chudnovsky et al., 2026) replays 10% of tokens up to 20,000 times and studies the repeat count that *maximizes* loss. Kazdan et al. (2026) relate semantic redundancy to replay in finite-corpus pretraining. Our grid covers lower repeat counts across sources and fractions and adds a matched-content control. We do not measure semantic redundancy, and our seen/unseen gaps do not identify a memorization mechanism (Carlini et al., 2023; Tirumala et al., 2022).

Recipes and mixtures. Regularization, ensembling and data augmentation change the attainable loss under a fixed corpus (Kim et al., 2026; Xu et al., 2026; Chen et al., 2026), as do sparsity (Wu et al., 2026) and memorization-guided epoch and replay schedules (Zuo et al., 2026); we evaluate specified recipes, not a tuned loss envelope. Mixture studies address source composition and per-source repetition directly (Liu et al., 2026; Sedova et al., 2026; Zhou et al., 2026); our sources are trained separately and do not yield a mixture optimum.

Stochastic tokenization. Subword regularization (Kudo, 2018) and BPE-dropout (Provilkov et al., 2020) were introduced for translation; BPE-dropout is used in autoregressive pretraining (Hu et al., 2024), serves as a baseline for stochastic token splitting (Sims et al., 2026), appears in masked-LM studies (Visser et al., 2026) and mitigates LLM sensitivity to typographical and format variation (Chai et al., 2024). We do not claim BPE-dropout or late shutoff as new; we measure loss across horizons and schedules with explicit token and padding costs. Partial repetition, BPE-dropout and the $(R-1)N/U$ coordinate thus all have precedents; we contribute controlled, paired evidence that separates the comparisons.

3 Questions and Setup

Effective-data scaling laws assign each repeated token a single discounted value set by counts. We test this assumption with three primary research questions. **RQ1 (comparator; §4):** does a repeated token's value depend on whether it adds processed tokens (fixed-corpus repetition) or displaces fresh tokens at a fixed budget (fresh-token replacement)? **RQ2 (allocation; §5):** at fixed processed, unique, and extra-exposure counts, do concentration and order change loss? **RQ3 (scale; §6):** do repetition effects across model sizes depend on whether unique data or TPP is held fixed, and do fits to smaller runs predict larger ones?

Allocations and excess loss. We write D for processed target tokens and U for unique target positions. Fixed-corpus repetition shows every sample R times ($D = UR$). Fresh-token replacement exposes a fraction f of processed targets to a repeated subset seen R_{sub} times and the rest to fresh tokens, yielding $U = D(1 - f + f/R_{\text{sub}})$. For source s and paired replicate j, the replacement excess loss is $\Delta_{s,j}(f, R_{\text{sub}}) = L_{s,j}(f, R_{\text{sub}}) - L_{s,j}(0, 1)$, where the fresh baseline shares the source, architecture, budget, and recipe.

Study families and paired designs. Appendix Table 1 lists what each study family holds fixed and varies. To isolate repetition effects from initialization noise and data order, we complement broad single-seed sweeps (using the $12Ld^2$ proxy on SlimPajama; Soboleva et al., 2023) with prespecified *paired studies*. Paired studies use exact non-embedding parameter counts and 3 or 6 paired replicates—each a shared data-order and initialization seed pair across the compared allocations. The source study trains 90 models (45.1M parameters, $D \approx 805$M targets) across three sources, with 18 runs extending one matched-content contrast to 6 replicates; the scale study trains 108 runs and the tokenization study 60. Models use Muon (Jordan et al., 2024) for hidden layers and Adam (Kingma & Ba, 2015) elsewhere; every run is separately annealed to its final horizon (Appendix A).

What matching TPP means. The scale study matches processed tokens per parameter $q_D = D/N_{\text{nonemb}} \approx 4.465$ in its replacement design and unique tokens per parameter $q_U = U/N_{\text{nonemb}} \approx 1.116$ in its full-corpus design. This controls parameter-relative exposure, not FLOPs, optimizer steps or batch schedules (Appendix A).

Evaluation. We report bits per byte (bpb) on per-source test sets of 2^{20} tokens, after removing training windows that exactly match an evaluation window; recipes and repeat counts are chosen on separate dev sets. Paired studies report 95% t intervals over replicates (Appendix B); in single-seed sweeps, differences below 0.002 bpb are treated as unresolved (the resolution threshold).

4 Two Quantities, Not One

4.1 Additional passes have large local value

For a fixed model, let U_{eq} be the fresh budget whose reference loss matches a repeated endpoint, and define the *credited extra-token gain* (CEG), the fresh-token credit per extra processed token, and the local whole-run fraction as

$$\text{CEG} = \frac{U_{\text{eq}} - U}{D - U}, \qquad \eta_{\text{loc}} = \frac{U_{\text{eq}}}{D}. \tag{1}$$

At the 59M sweep configuration, two passes over 25B unique tokens reach 1.0308 bpb, against 1.0296 for one pass over 50B fresh tokens (Figure 2a); the 0.0012 bpb gap is below the 0.002 bpb resolution

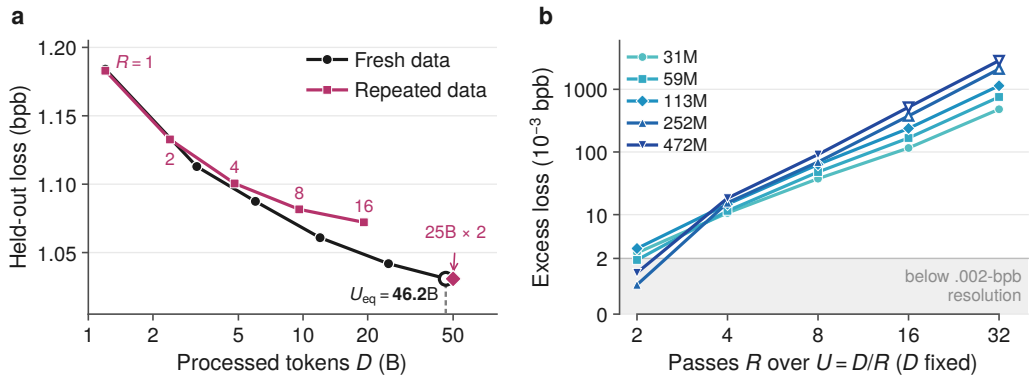


Figure 2: **At 59M, a second pass is worth most of a fresh pass relative to stopping (a), yet at a fixed budget every sweep model loses from repetition (b).** Single-seed sweeps; post-quantization endpoints; $12Ld^2$ size proxy; Muon. (a) Fixed-corpus repetition at 59M: measured one-pass fresh-data endpoints (circles, joined) and a fixed $U = 1.2$B ladder labeled by R (squares); both start from one pass at 1.2B. The diamond is 25B repeated twice; dashed guides map it to $U_{\text{eq}} = 46.2$B on the reference $L = E + AD^{-\alpha}$ fitted to the 6–50B endpoints (hollow circle), not extrapolated below 6B. (b) Excess loss of R passes over $U = D/R$ versus one pass at the same model and D (3.2B; 4B for 252M/472M at $R \leq 8$; open: added runs, App. G); linear below 0.002 bpb (gray).

threshold. Inverting a four-anchor fresh curve assigns the repeated run 46.2B fresh-equivalent tokens, so the second pass contributes about 21.2B (CEG $= 0.85$, $\eta_{\text{loc}} = 0.92$). Perturbing the repeated endpoint by ± 0.002 bpb moves CEG to 0.627–1.120, a sensitivity range rather than a confidence interval. On the fixed $U = 1.2$B ladder, the two-pass point lies on the fresh curve, and loss keeps falling through 16 passes while separating increasingly from it.

4.2 Replacing fresh tokens costs loss

At a fixed processed budget, every sweep model has higher loss with R passes over $U = D/R$ than with one pass over D fresh tokens (Figure 2b). At two passes the excess is at the resolution threshold (two independent one-pass 59M runs differ by 0.0011 bpb), so its order across sizes is not interpretable. It then grows with each doubling of R and, from $R = 4$, with model size: it stays below 2% of the fresh loss through $R = 4$, reaches 10–50% at $R = 16$, and at $R = 32$ the loss is 1.4 (31M) to 3.8 (472M) times its fresh value; at 252M and 472M training loss keeps falling (overfitting, not divergence). Replacing only part of the budget does not escape this: in the single-seed 59M five-source grid ($D = 2$B) and a 252M CommonCrawl panel ($D = 8$B), no partial allocation beats fresh data by more than 0.0023 bpb, barely above the threshold, whereas drawing the whole budget from 32 passes raises loss to 1.6–3.7 times its fresh value (Figure 12).

In the paired 45.1M study, the replicate mean exceeds its paired fresh baseline in all 27 non-fresh source–allocation combinations (Figure 3a–c). Intervals for light replacement often include zero (on Books, $(f = 1/8, R_{\text{sub}} = 2)$ costs only 0.00025 bpb), but the mean penalty rises monotonically with R_{sub} at fixed f. Each source is trained and evaluated separately and fresh losses differ (GitHub 0.594 bpb; CommonCrawl and Books 1.099 bpb and 1.156 bpb), so we compare sources by excess loss as a share of their own fresh loss: GitHub is the most sensitive in 8 of 9 allocations (38% versus 12% at $(f = 1, R_{\text{sub}} = 8)$). No source-independent safe repeat count follows.

Early checkpoints can misrank allocations. Development loss shows that intermediate checkpoints need not preserve the final ranking (Appendix Figure 15a). At the midpoint, $(f = 1/2, R_{\text{sub}} = 32)$ has excess loss 0.116 bpb versus 0.068 bpb for $(f = 1, R_{\text{sub}} = 8)$. During the final 20% learning-rate decay, the full-corpus penalty grows from 0.113 to 0.169 bpb while the partial-replacement penalty stays near 0.14 bpb, so the two swap order only in the last tenth of training.

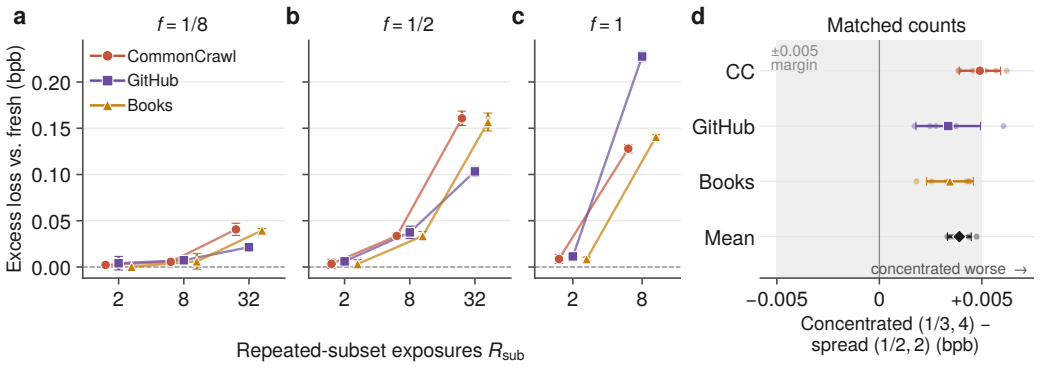


Figure 3: **All 27 non-fresh allocation means exceed their fresh baselines (a–c), and matched counts do not give matched loss (d).** Paired study; 45.1M non-embedding parameters, $D = 805{,}306{,}368$. (a–c) Excess loss relative to the same-source fresh baseline per allocation (f, R_{sub}): lines join source means with pointwise 95% t intervals ($n = 3$, df 2); dots are paired replicates. (d) Concentrated $(f = 1/3, R_{\text{sub}} = 4)$ minus spread $(f = 1/2, R_{\text{sub}} = 2)$ at equal processed, unique and extra-exposure counts (positive: concentration is worse); per-source and equal-source means with 95% t intervals over six replicates (df 5); gray shading marks the prespecified ± 0.005-bpb margin.

5 Equal Counts, Unequal Loss

5.1 Matched content does not give matched loss

Allocations $(f = 1/2, R_{\text{sub}} = 2)$ and $(f = 1/3, R_{\text{sub}} = 4)$ match every count: model, processed budget, unique samples and extra exposures, which fall twice on a larger subset or four times on a smaller one. Concentrating them raises loss by 0.0039 bpb (95% CI $[0.0033, 0.0045]$; Figure 3d), or 0.3–0.6% of each source's fresh loss; all 18 source–replicate differences are positive (sign-flip $p = 0.031$, the minimum attainable with six replicates). The mean is within the prespecified ± 0.005 bpb margin, but the CommonCrawl interval ($[0.0039, 0.0059]$) crosses it. The gap opens after the first fifth of training (Appendix Figure 15b); its consistent sign shows that total counts do not determine loss. Models that discount repeats per sample rather than per corpus can separate the two (Appendix F).

5.2 Repeat order and mixture

Holding (N, U, R) fixed, we reorder the same tokens three ways (Figure 4a): *spaced* presents R successive full passes (the default), *shuffled* draws one global permutation of all RU tokens, and *consecutive* repeats each shard R times in a row. Consecutive repetition raises held-out loss by up to 0.46 bpb relative to spaced repetition, whereas global shuffling stays within 0.013 bpb (Figure 4b); all three orders have identical token counts. Consecutive order also shrinks the unseen-minus-seen loss gap (held-out evaluation minus a training-probe split) from 0.038–0.186 bpb to 0.004–0.019 bpb (Figure 4c), so a small gap need not mean good held-out loss.

Across seven sources, unigram entropy correlates negatively with the 16-pass loss penalty ($r = -0.89$), yet an entropy-proportional replay mixture is worse than uniform repetition, by 0.06 bpb in mixture loss and 0.09 bpb in mean domain loss, so the per-source correlation does not by itself give a mixture rule (Appendix D).

6 Scaling Trends Depend on What Is Held Fixed

6.1 Fixed unique data: larger models peak after fewer passes

At $U = 1$B on CommonCrawl, the three smallest sweep models (14M–59M) do not degrade by 32 passes. The five larger ladders turn upward inside the grid; a parabola in $\log R$ through each ladder's

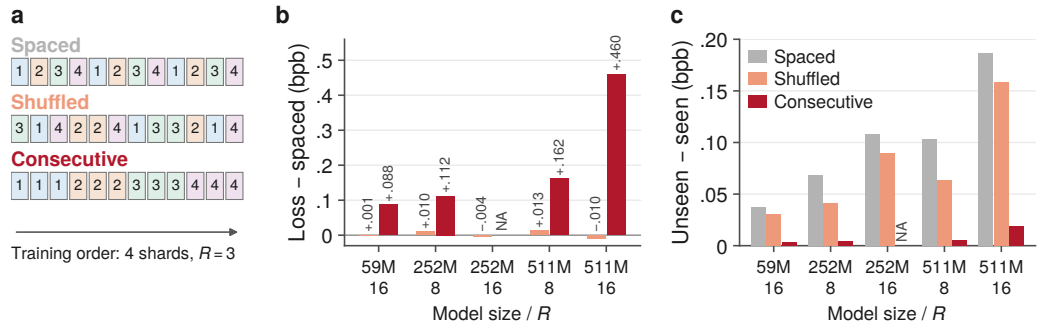


Figure 4: **Consecutive repetition, not shuffling, drives the large loss increase at fixed (N, U, R).** Sweeps on CommonCrawl at $U = 1$B; $12Ld^2$ size proxy. (a) Schematic of the three orders of the same data, drawn at shard level. (b) Loss of shuffled and consecutive runs minus that of the spaced run with the same (N, U, R). (c) Unseen-minus-seen loss gap: consecutive order nearly removes the gap while raising loss, so a smaller gap does not by itself indicate better generalization. NA: run not available.

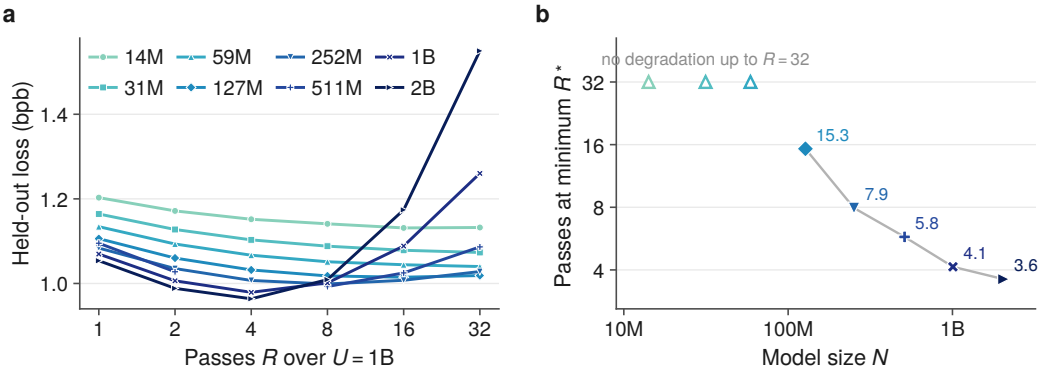


Figure 5: **At fixed U, loss minima occur after fewer passes at larger sizes.** CommonCrawl sweep at $U = 1$B. (a) Held-out loss of separately annealed runs; the 1B and 2B ladders reach 1.260 and 1.550 bpb at $R = 32$. (b) Passes at the loss minimum versus size: the minimum of a parabola in $\log R$ through each ladder's lowest point and its neighbors, for the five ladders that later rise by over 0.002 bpb; open triangles mark three ladders with no such degradation by $R = 32$. The excluded 511M/$R = 4$ record leaves a gap in (a); its later rerun (App. G) is not used here. Sizes use the $12Ld^2$ proxy.

lowest point and its two neighbors places the loss minimum at 15.3, 7.9, 5.8, 4.1 and 3.6 passes for the 127M, 252M, 511M, 1B and 2B proxy parameters (Figure 5).

Past the minimum, loss also rises faster at larger sizes: from the minimum to 32 passes it rises by 0.004 bpb at 127M and 0.59 bpb at 2B, and only the 1B and 2B models end worse than one pass (by 0.19 bpb and 0.50 bpb). A power law through the five minima gives $R^* = 4.51(U/N)^{0.51}$ (leave-one-out exponents 0.39–0.61). It does not extend to smaller models: it places the 31M and 59M minima at 26 and 19 passes, yet both ladders still improve at 32. With one shared U, U/N varies only through N, so the fit cannot separate a data effect from a parameter effect.

6.2 Matched tokens per parameter: the larger model loses less to heavy replacement

At matched processed TPP ($q_D \approx 4.465$), the prespecified primary contrast—the larger-minus-smaller excess loss difference at $(f = 1/2, R_{\text{sub}} = 8)$, averaged across two sources—is -0.0138 bpb (95% CI $[-0.0192, -0.0083]$; Figure 6a). The penalty is smaller at the larger size but does not vanish: 0.069 / 0.062 bpb (CommonCrawl / GitHub) at 45.1M versus 0.054 / 0.049 bpb at 184.6M.

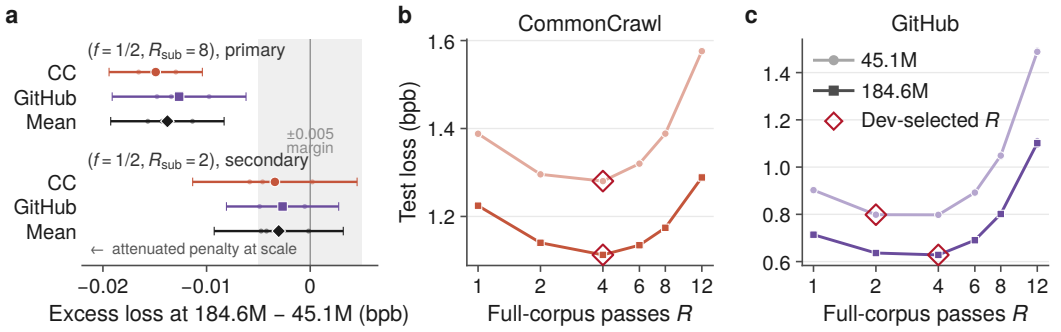


Figure 6: **At matched processed TPP, the larger model has a smaller heavy-replacement penalty; at matched unique TPP, the optimum is not established.** Exact non-embedding parameters. (a) Matched processed TPP ($D/N = 4.4655$ vs. 4.4641): 184.6M minus 45.1M replacement excess loss, with the primary ($f = 1/2, R_{\text{sub}} = 8$) and secondary ($f = 1/2, R_{\text{sub}} = 2$) allocations; negative values indicate that the larger model incurs a smaller replacement penalty. 95% t intervals over three replicates; gray shading marks the prespecified ± 0.005-bpb margin. (b,c) Matched unique TPP ($U/N = 1.1164$ vs. 1.1160): test loss of separately annealed full-corpus runs; lines join replicate means and are not fitted curves. Diamonds mark the repeat count selected on development data before testing.

On development loss the contrast turns negative after the first 20% of training (Appendix Figure 15c), so it is not produced only by the final learning-rate decay. For light replacement, $(f = 1/2, R_{\text{sub}} = 2)$, the contrast is -0.0030 bpb (95% CI $[-0.0093, 0.0032]$): neither a sign nor equivalence within ± 0.005 bpb is established.

At matched unique TPP ($q_U \approx 1.116$), mean test loss is lowest at four passes in all four source–scale strata (at 45.1M/GitHub two passes are within 0.0006 bpb; Figure 6b,c), and unlike on fixed-U ladders the development-selected count does not fall with size (4 and 4 on CommonCrawl, 2 and 4 on GitHub). These optima are descriptive: no selected count has both neighbors worse under the prespecified simultaneous criterion (Bonferroni over 20 adjacent contrasts, df 2). That criterion has low power—at 184.6M/CommonCrawl all three replicate differences from four to six passes lie in $[+0.020, +0.025]$ bpb, yet its interval includes zero (Appendix Figure 9)—so this is not evidence of a flat optimum.

Reconciling the two designs. Along a fixed-U ladder N/U grows with size, so the penalty coordinate $z = (R-1)N/U$ (§6.3) is larger at every R and earlier minima are expected without any size-specific effect. Matching TPP removes this growth, and the larger model then has the smaller penalty (Figure 6a), so the penalty is not a function of TPP alone. Neither design isolates parameter count: batch size, shape, steps and learning rates also change with size.

6.3 Do fits to smaller runs predict larger ones?

A count-based reuse surface is useful for planning only if a fit to smaller runs predicts larger or out-of-range ones. We fit 73 fixed-U sweep endpoints (one architecture-mismatched record excluded) with

$$L = E + \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{eff}}^{\beta}} + \lambda z^{p}, \quad D_{\text{eff}} = U\Big\{1 + R_D\big[1 - e^{-(R-1)/R_D}\big]\Big\}, \quad z = (R-1)\frac{N}{U}. \tag{2}$$

The saturating benefit follows Muennighoff et al. (2023), and a penalty in $(R-1)N/U$ appears in Lovelace et al. (2026); we test this combination rather than propose the coordinate. Saturation alone cannot produce a loss minimum at finite R; the penalty can (Table 8).

Closer where reuse helps than where it hurts. Fit only to the 57 sweep configurations with $N \leq 511$M, the surface predicts the 16 held-out 1B and 2B endpoints with a mean absolute error of 0.0311 bpb (Appendix Figure 10): 0.0157 bpb on the ten beneficial points but 0.0567 bpb on the six high-reuse points, with a maximum error of 0.202 bpb.

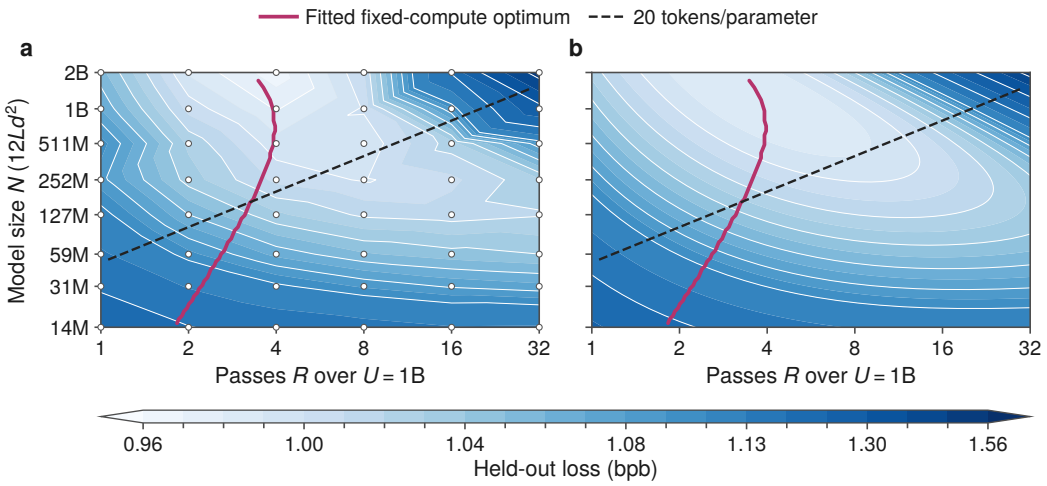


Figure 7: **The count-based surface captures the shape of the measured landscape at $U = 1$B, but not its detail.** (a) Piecewise-linear interpolation of the measurements (circles) in log coordinates; 511M/$R = 4$ is a later rerun not used by the fit (App. G). (b) The 73-point surface of Eq. (2), evaluated at its reported parameters on the same color scale and ending at the measured $R = 32$ boundary. Solid curves minimize fitted loss along fixed compute $6NUR$ where the minimum is interior; the dashed line is a 20-tokens-per-parameter fresh reference, not a refitted Chinchilla law. Contours are not additional observations; $12Ld^2$ sizes.

These held-out sizes also guided the choice of functional form, so this is model-selection validation rather than an independent test. In sample the fit looks close ($R^2 = 0.982$, MAE 0.0094 bpb), yet 63 of 73 residuals exceed the 0.002 bpb resolution threshold: a high R^2 does not imply calibration at the scale of the paired effects. The fit also places minima near 22, 18 and 15 passes for the three smallest models, whose ladders show no degradation by 32 (Appendix C).

Below the fitted unique budgets. Evaluated post hoc at the scale study's unique budgets, 5–20 times below the smallest fitted budget, the frozen surface predicts that 12 passes still beat one; instead, all four measured strata get worse, by 0.06–0.59 bpb (Appendix F). We therefore read the surface as a conditional description of these sweeps, not a prescriptive law. The measured unseen-minus-seen gap tracks its penalty term ($r = 0.956$ with z^p, 15 endpoints), but both vary along the designed axes, so the association does not identify a memorization mechanism (Appendix C).

7 Can Re-tokenization Make a Repeat Fresher?

A repeated pass shows the model the same bytes, but not necessarily the same tokens. Stochastic byte-pair encoding (BPE-dropout; Provilkov et al., 2020; Chai et al., 2024) re-segments each occurrence while preserving bytes, so a repeat can present new tokens without new data. We cross always-off, always-on ($p = 0.05$), final-20%-off and final-50%-off schedules with five full-corpus horizons $R \in \{1, 2, 4, 8, 16\}$ over three replicates (60 models; Figure 8). Positions, occurrences, order and bytes match within replicate and R; targets, padding and compute do not, and always-on uses about 1.9 times as many padded positions as always-off (Table 9).

Constant renewal helps only at high reuse. Relative to always-off, always-on is 0.021 bpb worse at $R = 1$ and 0.010 bpb worse at $R = 4$, but 0.022 bpb better at $R = 8$ and 0.154 bpb better at $R = 16$ (pointwise 95% CI $[0.144, 0.164]$). From 4 to 16 passes, always-off degrades by 0.218 bpb whereas always-on degrades by 0.054 bpb, so renewal softens the high-reuse penalty without eliminating it. Late shutoff trades between the two regimes: final-20%-off improves by 0.015 bpb over always-on at $R \leq 4$ but lags by 0.038 bpb at $R = 16$; extending shutoff to 50% adds 0.013 bpb and 0.055 bpb at $R = 8$ and 16; after shutoff, development loss need not change monotonically (Appendix H).

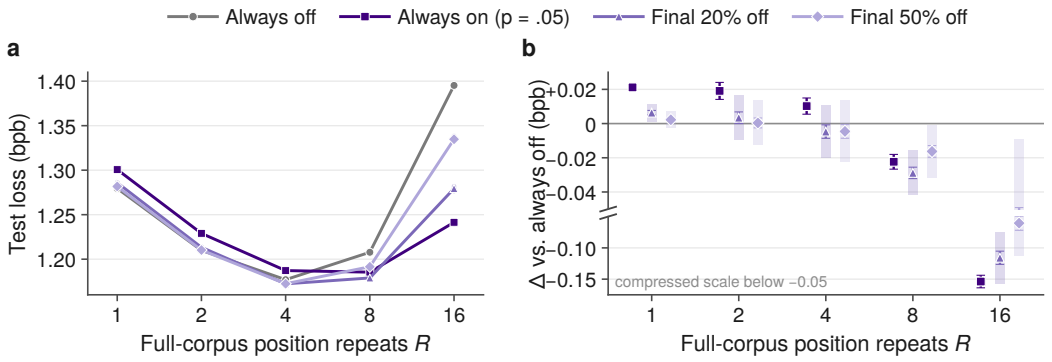


Figure 8: **Re-tokenizing repeats costs loss at short horizons and reduces it at high reuse.** (a) Final test loss: means and replicates. (b) Paired difference from always-off with pointwise 95% t intervals; light bars show simultaneous Bonferroni intervals for frozen secondary families. Both axis segments are linear; the scale is compressed below -0.05 bpb (break marks). Policies match position exposures and updates, not actual tokens or padded compute; all rankings are descriptive.

What the tokenization study does not show. The frozen primary of the original 45 models, final-20%-off minus always-on averaged over horizons, is -0.0028 bpb (95% CI $[-0.0042, -0.0014]$), short of the prespecified -0.005 bpb benefit; the separately frozen primary of the development-motivated 15-model extension, final-50%-off minus final-20%-off, is $+0.0122$ bpb ($[+0.0114, +0.0129]$). Horizon-specific contrasts are descriptive and, since renewal and compute move together, do not isolate a renewed-view mechanism (Appendix H).

8 Discussion and Conclusion

Implications for pretraining on finite data. Within the tested scales, the results suggest the following. When fresh data runs out, a second pass was worth most of a fresh pass at 59M, but at fixed U larger models reach their minimum after fewer passes. When fresh data is available, substituting repeats never lowered mean paired loss, and at a fixed share the cost grows with R_{sub}. Repeats are better spread across samples and passes than concentrated or run consecutively. Allocations should be compared at their final, separately annealed horizon, since early checkpoints can misrank them, and count-based surfaces should not be extrapolated outside their fitted range.

Reporting reuse. A reuse claim should state its comparator (stopping or fresh-token replacement), its allocation (repeated fraction, per-sample exposure distribution and the scale quantity held fixed), its presentation (order, mixture and tokenization), the checkpoint it is read at and its evidence status: here paired studies support bounded contrasts, while sweeps and their fitted surface are descriptive.

Conclusion. Token counts alone do not determine what a repeated token is worth. Relative to stopping, a second pass is worth most of a fresh one, yet at fixed compute, replacing fresh tokens with repeats raised mean loss in every pre-registered allocation. With all counts matched, allocation and order still change loss: slightly when exposures are concentrated on fewer samples, and by up to 0.46 bpb when shards are replayed consecutively. How repetition interacts with model size depends on which count is held fixed, and re-tokenizing repeats lowers loss at high repetition but not at matched compute. Within the tested scales, the value of a repeated token is set by its comparator, allocation, presentation and evaluation horizon, not by its count. Extending these measurements to larger models, longer training and downstream evaluations is a natural next step.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, plotting code, numerical consistency checks between reported values and frozen result files. All AI-assisted text, code and checks were reviewed by the authors, and every reported number is traceable to frozen experimental records. We take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A specify the architectures, exact parameter counts, token budgets, optimizers, learning-rate schedules, holdout construction and evaluation metric for the paired studies. Appendix B gives the seed pairs, test statistics and critical values, Appendix C the fitted surface parameters, and Appendix H the tokenizer and schedule details of the tokenization study.

REFERENCES

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

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1582–1599, 2024.

Michael K. Chen, Xikun Zhang, Fan Bai, Zhengding Hu, and Zhen Wang. Demystifying training-time augmentation for data-constrained language model pretraining. *arXiv preprint arXiv:2606.16246*, 2026.

Jessica Chudnovsky, Joshua Kazdan, Noam Levi, Rylan Schaeffer, Yegor Denisov-Blanch, Bo He, Mehmet Donmez, Sanmi Koyejo, and David Donoho. Internal data repetition destroys language models. *arXiv preprint arXiv:2606.24998*, 2026.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

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, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30016–30030, 2022.

Yiwen Hu, Huatong Song, Jia Deng, Jiapeng Wang, Jie Chen, Kun Zhou, Yutao Zhu, Jinhao Jiang, Zican Dong, Wayne Xin Zhao, and Ji-Rong Wen. YuLan-Mini: An open data-efficient language model. *arXiv preprint arXiv:2412.17743*, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, https://kellerjordan.github.io/posts/muon/, 2024.

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.

Joshua Kazdan, Noam Levi, Rylan Schaeffer, Jessica Chudnovsky, Abhay Puri, Bo He, Mehmet Donmez, Sanmi Koyejo, and David Donoho. Scale dependent data duplication. *arXiv preprint arXiv:2603.06603*, 2026.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. In *International Conference on Learning Representations (ICLR)*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, 2018.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Fengze Liu, Weidong Zhou, Binbin Liu, Ping Guo, Zijun Wang, Bingni Zhang, Yifan Zhang, Yifeng Yu, Xiaohuan Zhou, and Taifeng Wang. InfoLaw: Information scaling laws for large language models with quality-weighted mixture data and repetition. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.02364.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Justin Lovelace, Christian Belardi, Srivatsa Kundurthy, Shriya Sudhakar, and Kilian Q. Weinberger. Prescriptive scaling laws for data constrained training. *arXiv preprint arXiv:2605.01640*, 2026.

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)*, volume 36, pp. 50358–50376, 2023.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1882–1892, 2020.

Tian Qin, Kimia Hamidieh, and David Alvarez-Melis. Bridging compute- and data-optimal pretraining. *arXiv preprint arXiv:2607.25271*, 2026.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints. *arXiv preprint arXiv:2605.12715*, 2026.

Anya Sims, Thomas Foster, Tuan-Duy H. Nguyen, Klara Kaleb, Joseph Lee, Jakob N. Foerster, Yee Whye Teh, and Cong Lu. StochasTok: Improving fine-grained subword understanding in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R. Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. Dataset release, https://huggingface.co/datasets/cerebras/SlimPajama-627B, June 2023.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 38274–38290, 2022.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of LLM scaling based on human-generated data. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49523–49544, 2024.

Ruan Visser, Trienko Grobler, and Marcel Dunaiski. Pretraining language models with subword regularization: An empirical study of BPE dropout in low-resource NLP. *arXiv preprint arXiv:2605.13436*, 2026.

Boqian Wu, Qiao Xiao, Patrik Okanovic, Tomasz Sternal, Maurice van Keulen, Mykola Pechenizkiy, Elena Mocanu, Torsten Hoefler, and Decebal Constantin Mocanu. When data is scarce: Scaling sparse language models with repeated training. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.01155.

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.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling LLM under token-crisis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 59304–59322, 2023.

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. In *International Conference on Learning Representations (ICLR)*, 2026.

Kevin Zhou, Lisa Alazraki, Kris Cao, and Marek Rei. Repetition mismatch: Why data mixture experiments don't scale and how to fix them. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026. To appear; arXiv:2606.07597.

Jingwei Zuo, Cong Zeng, Ilyas Chahed, Maksim Velikanov, Dhia Eddine Rhaiem, Pasquale Balsebre, Abhay Kumar, Younes Belkada, and Hakim Hacid. Train smarter, not longer: Memorization-guided data reuse for efficient LLM training. *arXiv preprint arXiv:2607.04969*, 2026. 3rd DATA-FM Workshop at ICLR 2026.

A Essential Configurations and Evaluation Protocols

Table 1: **Sweeps describe; paired studies compare.** Sweeps use $12Ld^2$ proxy sizes and one seed per configuration; paired studies use exact non-embedding parameter counts and 3 or 6 paired replicates. CC: CommonCrawl.

Family	Models	Controlled	Varied	Sources
Single-seed sweeps (descriptive)				
Fixed budget	31M–472M	N, D	R ($U = D/R$)	mixed
Fresh reference	59M	N	U; R at 1.2B	mixed
Fixed unique data	14M–2B	U per ladder	N, $R \leq 32$; $U \in \{1, 2, 4, 8\}$B	CC
Replacement grid	59M, 252M	N, D	f, R_{sub}	5 sources; CC
Repeat order	59M–511M	N, U, R	order	CC
Multi-seed paired studies (interval estimates)				
Source	45.1M	N, D	fresh + 9 allocations	CC, GitHub, Books
Scale	45.1M, 184.6M	q_D or q_U	N; allocation or R	CC, GitHub
Tokenization	45.1M	N, U	schedule, $R \leq 16$	CC

Table 2: Architectures used in the new paired experiments. F counts analytical training forward/backward FLOPs per target, not complete system cost.

Field	45.1M model	184.6M model
Exact non-embedding parameters	45,084,900	184,623,360
Tied embedding parameters	32,164,480	51,463,168
Exact total parameters	77,249,380	236,086,528
Rounded total size	77.2M	236.1M
Width / depth	640 / 10	1024 / 16
Width/depth ratio	64	64
Query / KV heads	10 / 5	16 / 8
Head dimension	64	64
MLP width / activation	2560 / ReLU^2	4096 / ReLU^2
Vocabulary / context length	50,257 / 1024	50,257 / 1024
F per target	502,644,480	1,516,738,560

The paired models use grouped-query attention, rotary positions (base 10,000), RMS normalization, tied embeddings initialized at standard deviation 0.005, query/key gain 1.5 and logit softcap 30. Hidden matrices use Muon with five Newton–Schulz steps; scalar/vector and tied-embedding parameters use Adam, $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. Weight decay, dropout and gradient clipping are zero. Master parameters, gradients and optimizer states are FP32; activations are BF16. The recorded scientific environment is Python 3.12.4, NumPy 2.4.5, PyTorch 2.9.1/CUDA13.1, Triton 3.5.1 on eight H800 GPUs. The path uses full-graph compilation with TF32 and comprehensive padding disabled; the Muon update is not compiled. These are recorded recipes, not claims of globally optimal hyperparameters. Table 3 lists the per-study batch, accumulation and update settings.

Table 3: Essential paired-run settings. B is global targets/update; G is accumulation. Each GPU receives 16,384 targets per microbatch on eight GPUs. LR order is matrix/scalar/tied embedding. Replacement rows have fixed D; full-corpus rows have fixed U and $D = RU$, with $R \in \{1, 2, 4, 6, 8, 12\}$.

Study/model	LR tuple	B	G	Fixed budget / updates
Source / 45.1M	0.01/0.01/0.05	131,072	1	$D = 805{,}306{,}368$; $K = 6{,}144$
Processed TPP / 45.1M	0.02/0.02/0.05	131,072	1	$D = 201{,}326{,}592$; $K = 1{,}536$
Processed TPP / 184.6M	0.02/0.02/0.05	524,288	4	$D = 824{,}180{,}736$; $K = 1{,}572$
Unique TPP / 45.1M	0.02/0.02/0.05	131,072	1	$U = 50{,}331{,}648$; $K = 384R$
Unique TPP / 184.6M	0.02/0.02/0.05	524,288	4	$U = 206{,}045{,}184$; $K = 393R$

Twenty compilation warmup updates restore model, optimizer, RNG and loader state; they do not contribute to scientific D and are not LR warmup. Paired-run momentum increases from 0.85 to 0.95

over $\lceil 0.1K \rceil$ updates, and LR decays linearly over the final $\lceil 0.2K \rceil$, with $K = D/B$. In contrast, the primary broad sweeps use $(0.04, 0.04, 0.05)$ and ramp momentum over 500 updates, with recorded global batches in Table 4. Architecture sensitivity families can use different rates or proportional ramps, so they are not pooled with the main scale comparison.

Calibration and token counts. The 12 source-study calibration runs compare $(0.01, 0.01, 0.05)$ at $B = 131{,}072$ with $(0.02, 0.02, 0.05)$ at $B = 131{,}072$ and 262,144 on CommonCrawl/GitHub, for fresh data and for $(1/2, 8)$. The lower-rate setting is retained by the prespecified 0.002-bpb near-best rule, not because it has the smallest regret. Sixteen scale calibration runs compare the two LR tuples on both sizes and sources, for fresh data and 12 full-corpus passes. Mean regrets are 0.0119 and 0 bpb; only $(0.02, 0.02, 0.05)$ is within the band. Calibration runs are not scientific replicates. The two studies process 96,636,764,160 and 85,629,861,888 targets including calibration, or 86,973,087,744 and 69,221,744,640 science-only targets. Complete calibration, budget and schedule rows are in the supplementary material.

Matched TPP is not matched compute or matched steps. At 45.1M/184.6M, processed TPP is 4.4655/4.4641 and unique TPP is 1.1164/1.1160, a -0.031% relative difference. The larger unique budget was first rounded to a 524,288-target grid. Parameter and token budget ratios are 4.0950 and 4.0938; the batch ratio is four, so updates differ by 393/384 ($+2.34\%$). Counting total parameters instead gives processed TPP 2.606/3.491 and unique TPP 0.652/0.873, which are not matched. Analytical training FLOPs per target in Table 2 include the implemented model terms, not complete hardware cost, and differ from the sweep proxy $6ND$.

Table 4: Main CommonCrawl ladder and the separate mixed-corpus ladder. N is the rounded $12Ld^2$ proxy; N_{count} is the exact non-embedding count in millions, rounded here to three decimals. Batch is global target tokens per optimizer update. All rows have head dimension 64 and MLP width $4d$, but d/L is not fixed. The 59M row comes from the single-size fixed-corpus sweep; the other rows come from the model-size sweep. The mixed-corpus ladder of the fixed-budget sweep (Figure 2b) shares the 31M and 59M rows; its 113M and 472M models are not the 127M and 511M models above, and its 252M model is also used in the mixture sweep and the 252M replacement panel. The excluded architecture-mismatched/$R = 4$ architecture is described in App. C.

d	L	d/L	heads (KV)	N	N_{count} (M)	batch (tok)
Main CommonCrawl model-size sweep						
384	8	48.00	6 (3)	14M	12.990	262,144
512	10	51.20	8 (4)	31M	28.859	524,288
640	12	53.33	10 (5)	59M	54.102	524,288
768	18	42.67	12 (6)	127M	116.848	524,288
1024	20	51.20	16 (8)	252M	230.779	524,288
1280	26	49.23	20 (10)	511M	468.733	1,048,576
2048	20	102.40	32 (16)	1.0B	922.932	1,048,576
2560	26	98.46	40 (20)	2.0B	1874.630	1,048,576
Mixed-corpus ladder						
768	16	48.00	12 (4)	113M	100.719	524,288
1024	20	51.20	16 (4)	252M	220.293	1,048,576
1280	24	53.33	20 (5)	472M	413.016	1,048,576

Sweep count, metric and holdout boundaries. The common proxy $N = 12Ld^2$ differs from exact non-embedding counts and does not fix architecture. At the 252M proxy label, four-KV replacement/mixture runs count 220,293,440 non-embedding parameters while the eight-KV ladder counts 230,779,200. The original byte-normalized metrics and their specified pre/post-int8 stages are preserved, not converted to the paired ordinary-target metric. The same validation routine and split were recovered for pre/post-quantization in the mixed fresh-reference family, but the five checkpoints were not recovered for paired checkpoint-level verification. The older 1.2B ladder overlay does not establish cross-family holdout identity. Exact holdout exclusion does not eliminate shorter or semantic near-duplicates; exposure of the new holdouts to earlier models is unknown.

Single-size optimizer comparison. Twelve completed 59M-proxy Muon/AdamW (Jordan et al., 2024; Loshchilov & Hutter, 2019) runs share the recorded architecture, mixed corpus, $D = 3.2$B and batch 524,288; each repeat count uses $U = D/R$. Both ladders worsen relative to their own one-pass baseline (Table 5). At $R = 32$, Muon/AdamW absolute losses are 1.867810/1.760766 bpb, despite smaller Muon's one-pass loss (1.112899 versus 1.152695). These post-int8 endpoints therefore establish neither a tuned optimizer ranking nor optimizer independence. The results are single-seed and do not transfer to other model sizes.

Table 5: **Optimizer sensitivity at one model size and token budget.** Absolute post-quantization mixture losses L are in bpb; $\Delta L = L(R) - L(1)$ is computed relative to each optimizer's own single-pass endpoint, in 10^{-3} bpb. Values use the recorded final byte-normalized validation endpoint, not a streaming checkpoint. L and ΔL are rounded independently from unrounded endpoints, so ΔL can differ in its last digit from the difference of displayed L values. Decimal precision is for traceability, not statistical inference.

R	L_{Muon}	L_{AdamW}	ΔL_{Muon}	ΔL_{AdamW}
1	1.112899	1.152695	0.000	0.000
2	1.114845	1.153794	1.945	1.100
4	1.124383	1.157711	11.484	5.016
8	1.160769	1.189120	47.869	36.426
16	1.279985	1.283226	167.086	130.531
32	1.867810	1.760766	754.911	608.072

Paired evaluation construction. Each source dev/test split contains 1,024 non-overlapping raw windows of 1,025 tokens, providing 1,024 prediction targets each. NumPy PCG64 uses `SeedSequence([20260917, source_id])`, with CC/GitHub/Books IDs 0/1/2. Complete 1,024-token matches to prior holdouts and reserved technical regions are excluded; new splits and training sample collections are screened by selected IDs, full target windows and full contexts. The scale tests use exactly the same CC/GitHub splits. This does not imply document/repository/semantic independence. Each test has 1,048,576 target slots; EOT counts are 615, 524 and 3 and raw byte counts are 4,698,208, 2,524,586 and 4,204,392 for CC, GitHub and Books. EOT contributes to training D but not the ordinary-target NLL or byte numerator/denominator. All-target metrics are separately retained.

B Paired Estimates and Local-Optimum Criteria

The paired unit is a replicate: one training data-order/initialization seed pair. Scientific seed pairs are (73001,1337), (73002,1400), (73003,1401); matched-content comparisons add (73004,1402), (73005,1403), (73006,1404). Calibration uses (73999,1399). Data draws can overlap across replicates. Source-grid and scale contrasts use three replicates and $t_{0.975,2} = 4.30265$; the six-replicate matched-content contrast uses $t_{0.975,5} = 2.5705818356$ and all 64 sign patterns. Only the equal-source matched-content mean and heavy-replacement scale contrast are their respective primary quantities; source-specific and light-replacement results are secondary. Development data alone fixed recipes and repeat candidates; no test outcome selected a checkpoint, seed, repeat count or fit. The 0.002-bpb sweep threshold is operational, not a confidence bound.

Full-corpus candidates are selected on development means before test access: 45.1M/CC four, 45.1M/GitHub two, and both 184.6M sources four passes. The 20 adjacent contrasts share simultaneous 95% Bonferroni paired-t intervals (critical value 19.962, df 2). No candidate has both neighbors significantly worse and no lower nonadjacent mean. Failure to bracket is not equivalence, and the 0.0006-bpb lower test mean at four versus two passes for 45.1M/GitHub does not reselect the candidate. Figure 9 shows all 20 intervals and 60 replicate differences; every endpoint is in the supplementary material.

Prespecified rules for the scale study. The sole primary statistic is the two-source equal mean of the larger-minus-smaller replacement excess loss at $(f, R_{\text{sub}}) = (1/2, 8)$; the $(1/2, 2)$ contrast is secondary. Practical equivalence requires the whole interval inside ± 0.005 bpb. For full-corpus

Table 6: Selected paired estimates retained without reanalysis. Matched-content rows use six replicates (df5); scale rows use three (df2). Only the source-mean matched-content and heavy-replacement scale contrasts are primary. Full rows, replicate values and exact numerical sources remain in the standalone tables.

Contrast	Stratum	Mean (bpb)	95% CI
Matched content	Source mean	0.003899196	[0.003320199,0.004478193]
	CommonCrawl	0.004898479	[0.003893582,0.005903375]
	GitHub	0.003356595	[0.001780624,0.004932565]
	Books	0.003442514	[0.002303969,0.004581059]
Scale, $(1/2, 8)$	Source mean	-0.013778684	$[-0.019248520, -0.008308849]$
Scale, $(1/2, 2)$	Source mean	-0.003038765	$[-0.009265014, 0.003187484]$

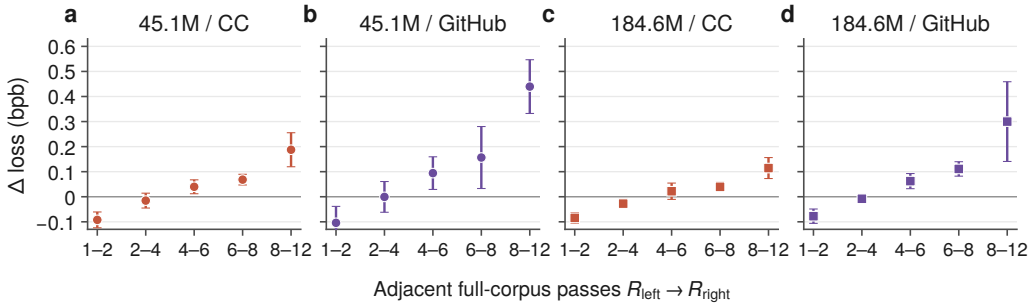


Figure 9: **Adjacent full-corpus contrasts at matched unique TPP.** Test loss at R_{right} minus R_{left} for each source–scale stratum: means with simultaneous 95% Bonferroni paired-t intervals over all 20 contrasts (df 2, critical value 19.96), and the three replicate differences (dots). Model sizes are exact non-embedding counts.

repetition, the three-replicate development mean selects a candidate from $\{1, 2, 4, 6, 8, 12\}$, with ties favoring fewer passes, before any test access. A local bracket requires both neighboring losses to be higher under the simultaneous intervals and no lower nonadjacent mean; even that would not establish a continuous or global optimum.

C Sweep Fit Parameters, Errors and Robustness

The retained surface uses the original 73-endpoint subset, excluding the misgrouped 511M/$R = 4$ entry without a replacement. That run used a different architecture (width 1536, depth 18, 16 query/4 KV heads, head dimension 96), not the width-1280/depth-26/20-query/10-KV family. Its observed loss 0.994897 remains in the source records but does not enter the fit. Three later $U = 4$B endpoints are not added. This exclusion and all fit corrections already preceded the paired experiments; they are not discoveries from the new tests.

The fixed-U ladders use separately annealed runs at each horizon, not intermediate checkpoints of one schedule. The 14M ladder rises by only 0.0010 bpb from 16 to 32 passes, and the 31M and 59M ladders improve through 32. The 511M parabola uses $R = 2, 8, 16$ after the architecture-mismatched $R = 4$ exclusion. Treating the ladders without an upturn as operational bounds gives a separate 0.708 sensitivity exponent, not a confidence interval.

Same U and R across model sizes. Table 7 gives each fixed-U ladder relative to the same model's one-pass run at the same U, so every column compares sizes at an identical (U, R). At both unique budgets the second-pass gain grows with size: from 0.031 to 0.063–0.067 bpb between 14M and 511M–2B at $U = 1$B, and from 0.022 to 0.036 bpb between 59M and 1B at $U = 4$B. At $U = 1$B only the 1B and 2B models end above their one-pass loss, from $R = 16$ onward. At $U = 4$B, where $U/N \geq 3.97$ for every size, all four ladders still improve at $R = 8$, consistent with the later minima that $R^* \propto (U/N)^{0.51}$ implies. The reference here is stopping after one pass; fresh data at $D = UR$ were not trained for these sizes. The shared-(U, R) comparison against fresh data is Figure 2b,

where D is fixed and $U = D/R$, on the mixed corpus. All entries are single-seed; the 511M one-pass endpoint lies above the 252M endpoint, so its gains are measured from a comparatively weak reference.

Table 7: **Fixed-U ladders relative to each model's own one-pass run** (CommonCrawl sweeps, pre-quantization CommonCrawl held-out bpb, $12Ld^2$ proxy sizes). $L(1)$ is the one-pass loss; other columns give $L(R) - L(1)$ in 10^{-3} bpb, negative when the extra passes help. NA marks the excluded 511M/$R = 4$ record (rerun: App. G); dashes mark horizons not trained. Single seed per entry.

N	$L(1)$	$R = 2$	$R = 4$	$R = 8$	$R = 16$	$R = 32$
$U = 1$B						
14M	1.2029	−31.4	−51.1	−61.9	−71.6	−70.6
31M	1.1642	−36.6	−61.2	−76.1	−85.8	−90.9
59M	1.1343	−41.0	−67.5	−83.0	−90.0	−94.2
127M	1.1060	−45.9	−74.0	−88.2	−91.1	−87.3
252M	1.0837	−48.4	−76.5	−84.9	−76.1	−55.5
511M	1.0951	−67.4	NA	−102.3	−70.1	−8.6
1B	1.0696	−63.3	−90.5	−68.2	+18.9	+190.7
2B	1.0538	−65.3	−90.0	−44.7	+120.2	+496.0
$U = 4$B						
59M	1.0605	−22.4	−37.9	−48.0	–	–
252M	0.9955	−28.3	−46.7	−59.2	–	–
511M	0.9788	−34.4	−57.2	−68.5	–	–
1B	0.9553	−35.8	−57.0	−64.8	–	–

The fitted parameters are $(E, A, \alpha, B, \beta, R_D, \lambda, p) = (0.705, 7.44, 0.189, 4.45 \times 10^3, 0.485, 4.545, 5.199 \times 10^{-3}, 1.160)$, with N, U entered as raw proxy-parameter and token counts. Units change coefficients. The stationary condition retains separate U and N/U dependence, so it does not mathematically imply a ratio-only optimum law. The five-minimum exponent 0.509, surface-derived 0.410, and operational-bound sensitivity 0.708 (Figure 11) are distinct descriptive quantities, not confidence limits. The leave-one-out 0.386–0.613 range is not a confidence band.

Table 8: **Penalty forms on the 73 sweep endpoints.** R^2 is in-sample; MAE is over the 16 held-out-size 1B/2B endpoints (Figure 10). Parameter counts include the shared benefit terms; diagnostics are in Appendix C.

Form	Penalty	Parameters	R^2	MAE (bpb)
Benefit-only	none	6	0.360	0.0884
Separable	$\lambda(R-1)^p(N/U)^q$	9	0.985	0.1142
Load-augmented	λz^p	8	0.982	0.0311

The all-branch residual and held-out-size results remain those in Table 8. Global R^2 does not establish 0.002-bpb calibration: 63 of 73 residuals exceed that resolution. Omission of the entire 14M model–data group gives the largest extrapolation below the retained training range; group-omission prediction ranges are conditional optimizer outputs, not bootstrap, confidence or prediction intervals.

Seen/unseen gap. In the CommonCrawl $U = 1$B sweeps at 14M, 31M and 59M, the unseen-minus-seen bpb gap increases with R, faster for larger models at matched R, and its change precedes the held-out upturn. Across the 15 $R > 1$ endpoints the log–log correlation between z^p and the gap is $r = 0.956$; the three $R = 1$ points ($z = 0$) are omitted from that logarithmic comparison. The gap was not used in the fit. Because multiplying $\log z$ by any positive exponent leaves the correlation unchanged, the association validates neither p nor a causal memorization mechanism. All points are in the supplementary material.

Architecture reversal checks use counted active non-embedding parameters, not the dense proxy fit. Forty-eight re-evaluated endpoints plus five rounded 1.2B-ladder summaries give a nine-configuration best-shallower first-reversal slope 0.205 under a 0.010-bpb criterion, with a conditional profile set

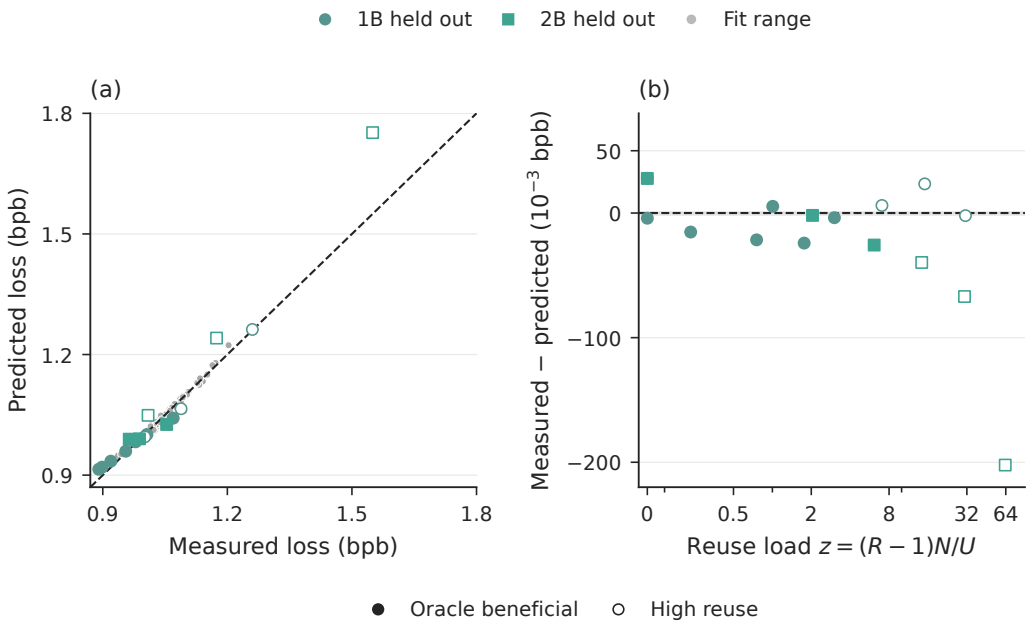


Figure 10: **Held-out-size model-selection check of the surface.** Eq. (2) is fit to the 57 configurations with $N \leq 511$M and predicts all 16 held-out 1B/2B endpoints. (a) Measured versus predicted loss. (b) Measured-minus-predicted residuals versus z on a symmetric-log axis that includes the three $R = 1$, $z = 0$ points. Filled symbols mark the ten oracle-beneficial points and hollow symbols the six high-reuse points; error summaries are given in §6.3.

[0.040,0.385]. The complete-checkpoint subset alone is weakly identified and has no unique positive slope. Every-shallower reversal estimates a different quantity. Within one proportional-ramp MoE recipe, final overshoot increases across expert count; microbatch, widths and tokens per expert differ. These checks neither validate the dense optimum law nor isolate pure expert count as a mechanism. Full cohort tables, alternative definitions and plots are in the supplementary material.

D Source, Fraction, Order and Mixture Boundaries

Across the six sweep panels, 21 of 24 fraction columns are nondecreasing in R at the 0.002 resolution; the three exceptions are Books, whose largest decline ($R = 16 \to 32$ at $f = 0.5$) is 0.015 bpb while remaining 0.115 bpb worse than fresh. Let $R_{\text{safe}}(f)$ be the largest R_{sub} whose excess over all-fresh training stays within a 0.01-bpb tolerance, log-interpolated along each f column; on CommonCrawl it follows $R_{\text{safe}} \approx \kappa f^{-|b|}$ with $(\kappa, |b|) \approx (2.7, 0.87)$. Changing the tolerance to 0.005 or 0.02 gives $(\kappa, |b|) = (1.95, 0.80)$ or $(3.93, 1.09)$. The CommonCrawl tolerance fit is based on log-interpolated 0.01-bpb crossings, not seed confidence. Across fractions and sizes, fR_{safe} ranges 1.65–4.07; alternative 0.005/0.02 thresholds give different coefficients. The positive low-penalty subset of 54 points gives $\Delta \propto (fR)^{1.06}$ and $R^2_{\log} = 0.684$, versus 0.353 for R alone. All 79 observed interior points, including points excluded from this positive fit, are in the supplementary material. The cross-scale contour in Figure 13 has different token budgets and architectures; the largest-recycling penalty rises from 0.83 to 3.39 bpb. This does not refute the separate matched-TPP absolute-loss contrast.

The entropy-informed mixture test uses seven source-specific 0.1B-token corpora, 28 nominal corpus passes, and 5,341 updates at batch 524,288 (2,800,222,208 targets). The displayed eight-run anti/entropy/uniform groups have 3/4/1 runs but only 2/2/1 distinct seeds, with mixed A100/H800 hardware; a ninth uniform run is not pooled. Error bars are population SD across these runs, not independent seed uncertainty; uniform SD is unavailable, not zero. Uniform reaches 1.211 own-domain and 1.217 mixture bpb; entropy is worse by 0.060 in mixture loss and by about 0.086 on

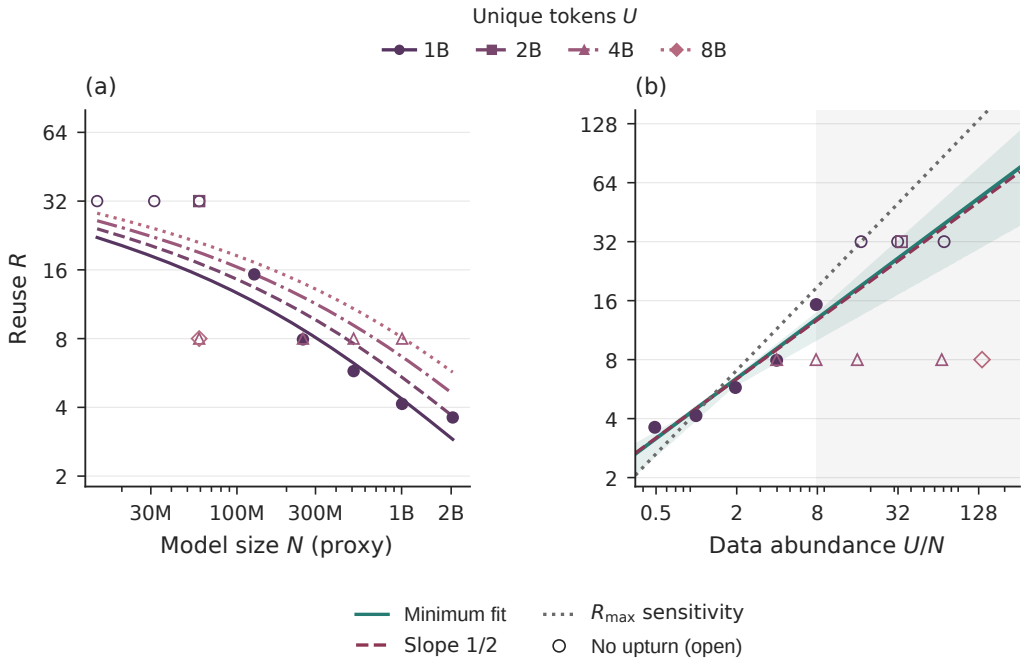


Figure 11: **Estimated minima and observed ranges.** (a) Filled points are the five minima estimated by a parabola in $\log R$; open symbols (legend: *No upturn*) mark $R_{\max}$ for ladders without an upturn, with shape and line style also identifying U. Curves are optima of the 73-point surface. (b) The fit to the five estimated minima (legend: *Minimum fit*) is $R^* = 4.508(U/N)^{0.509}$, with $R^2_{\log} = 0.937$; all five points have $U = 1$B. The shaded band is a leave-one-out envelope, not a confidence interval (slopes 0.386–0.613). Dashed and dotted lines show a slope-$1/2$ summary and the operational-bound sensitivity fit (0.708). Gray shading marks $U/N > 7.85$, beyond the range of the five estimated minima. Four groups without an upturn were measured through 32 epochs and five through 8; these observed limits do not prove a mathematical optimum beyond them.

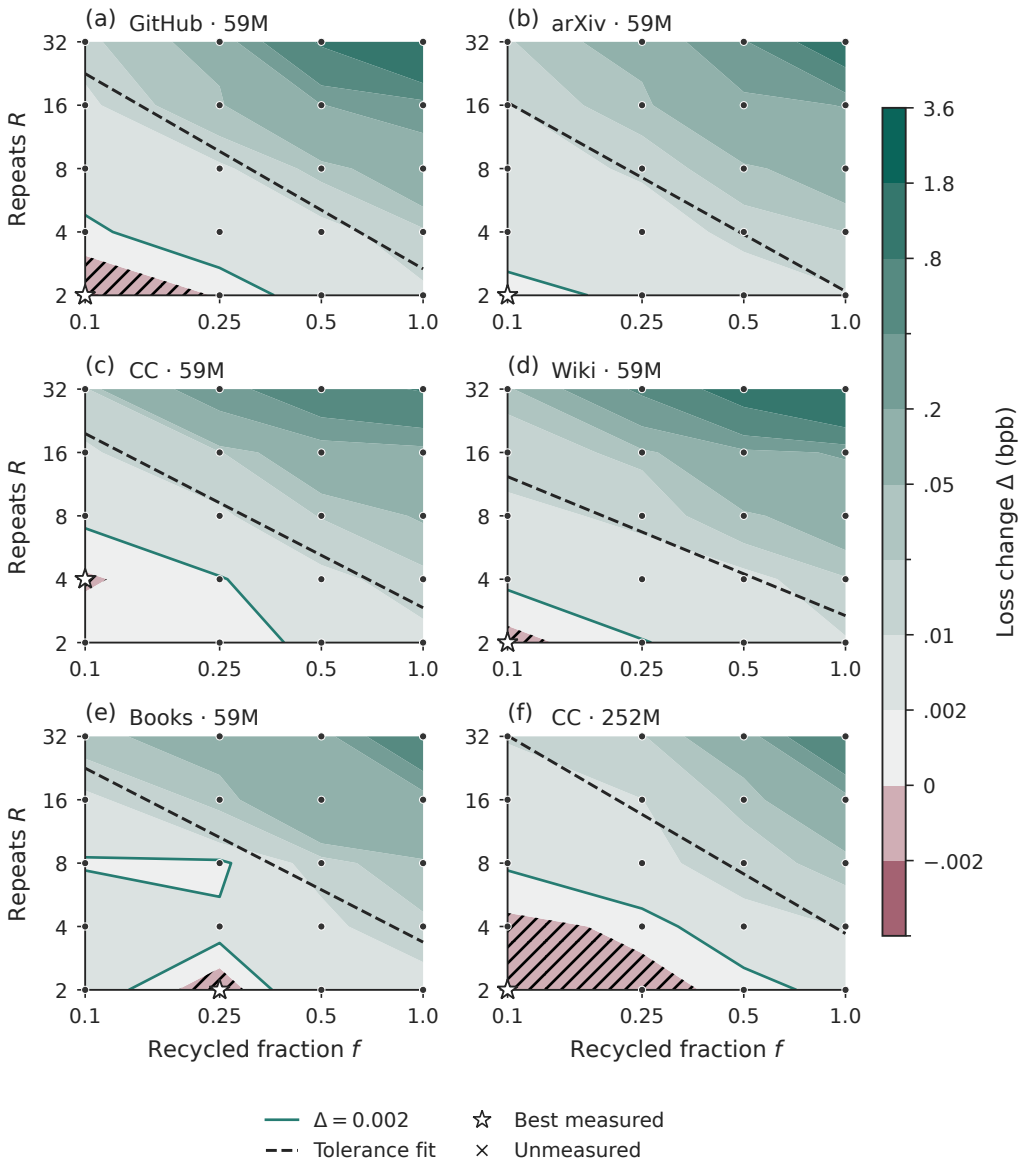


Figure 12: **Fresh-token replacement by domain at fixed processed budget (sweeps).** Signed excess loss $L(f, R_{\text{sub}}) - L(0, 1)$, with the axis R denoting R_{sub}, at 59M/$D = 2$B and, for the additional CommonCrawl panel, 252M/$D = 8$B, each relative to the same domain, model and budget. Dots are measured configurations; stars mark the lowest observed interior loss, not an established optimum; crosses mark unmeasured configurations, whose color is display interpolation. The solid contour marks 0.002 bpb and dashed lines the per-panel 0.01-bpb tolerance fits. Contour intervals are nonlinear; no run-to-run uncertainty is estimated.

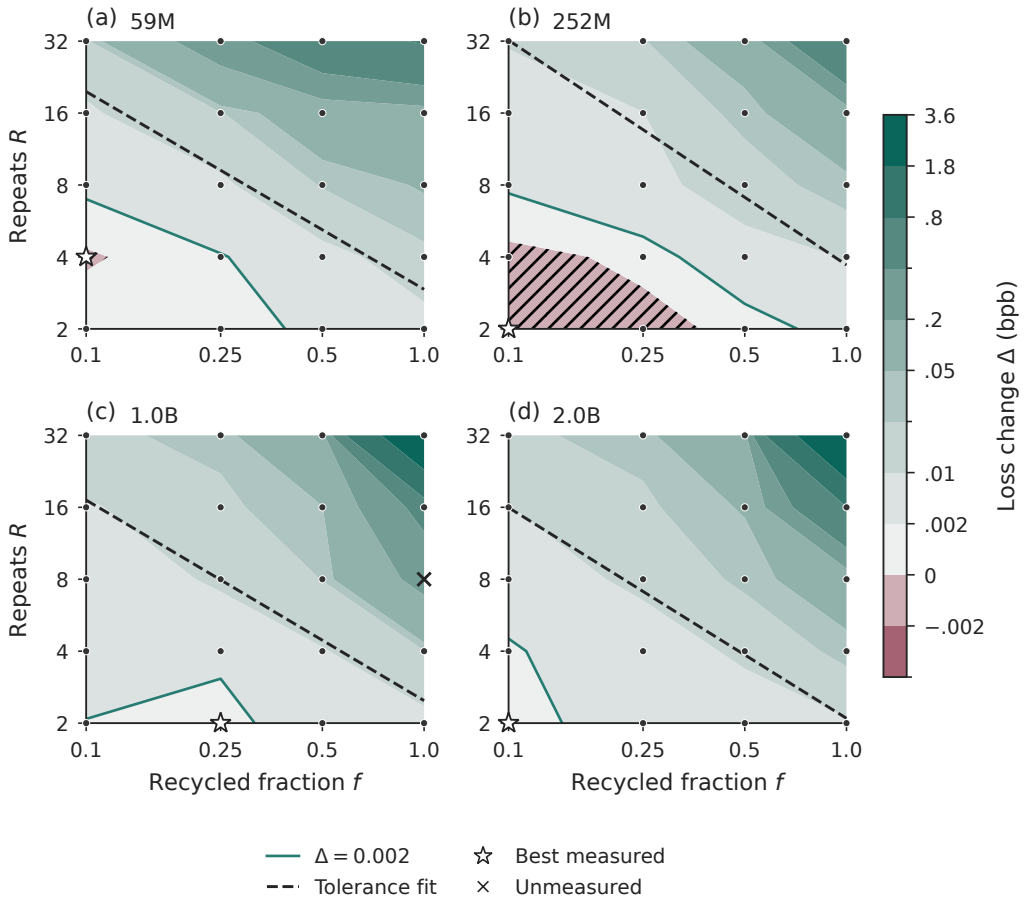


Figure 13: **Fresh-token replacement across model sizes at fixed processed budget.** CommonCrawl excess loss relative to same-model all-fresh training, at $D = 2$B for 59M and $D = 8$B for the larger models; the axis R denotes R_{sub}. Colors, measured dots, interior-minimum stars and logarithmic axes follow Fig. 12. Crosses identify unmeasured configurations, whose interpolated color is not a measurement. Solid tolerance contours mark 0.002 bpb; dashed curves fit the 0.01-bpb tolerance boundary separately at each size. High-recycling excess reaches 3.39 bpb at 2B. These panels have different token budgets and the 252M four-KV architecture differs from its main-sweep counterpart.

the primary equal-source metric, and anti has mixture loss 1.332 (Figure 14). The hypothesized anti>uniform>entropy order is not observed.

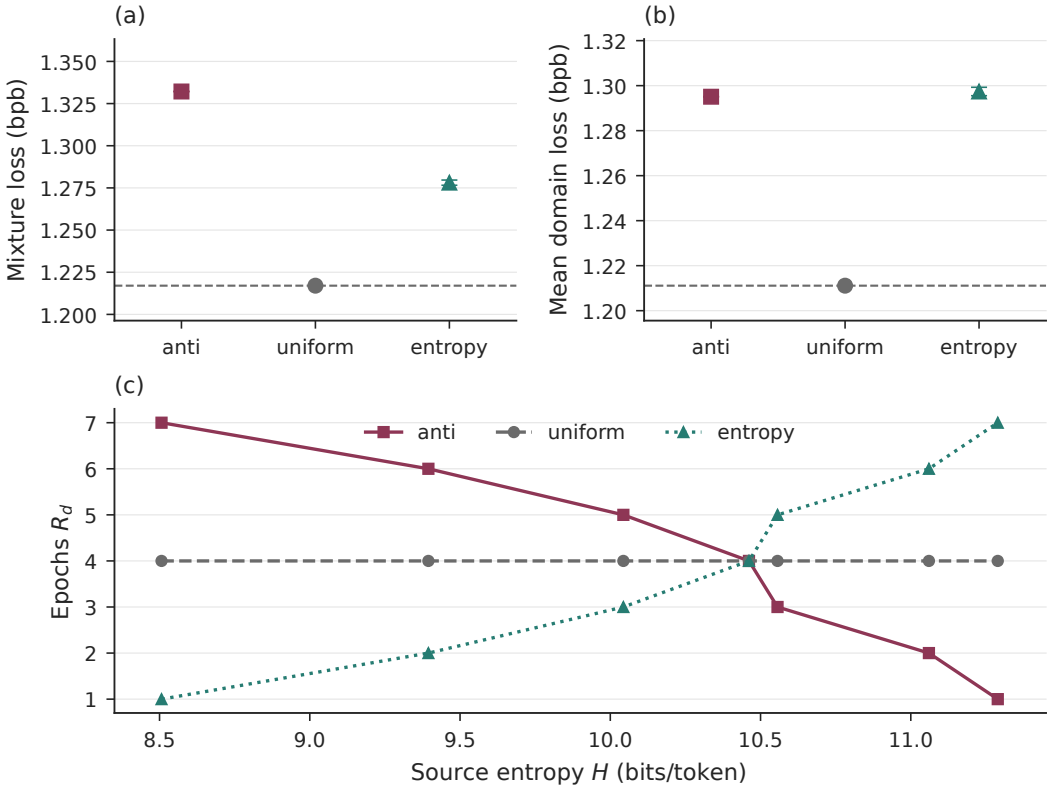


Figure 14: **Mixture allocation at matched token budget.** (a) Mixture loss and (b) the primary equal-weight mean of seven domain losses, before quantization. Points retain the original eight-run snapshot: anti/entropy/uniform have $3/4/1$ runs and $2/2/1$ distinct seeds. Anti combines two A100 and one H800 runs; entropy contains both seeds on both hardware types; uniform is one A100 run. Error bars are descriptive population SD across runs (`ddof` $= 0$), not independent-seed uncertainty. Uniform's SD is NA, not zero. Dashed horizontal lines denote uniform. (c) Every policy sums to 28 passes over seven source-specific 0.1B-token corpora, with Books at four passes. The later ninth completed uniform run is not pooled.

For the spacing intervention, the spaced, shuffled and consecutive runs share (N, U, R) and differ only in repeat order. The observed gap reduction with consecutive repetition (0.005 versus 0.10 at 511M/$R = 8$) does not establish better generalization; the endpoint probe does not isolate forgetting. The recommendation is limited to avoiding long consecutive repeats of one shard in the measured settings, not that schedules are universally interchangeable. Fourteen of the 15 runs completed; the 252M/$R = 16$ consecutive run did not, and Figure 4 shows it as NA.

Development-loss trajectories. The paired source and scale studies record development loss at every tenth of training (11 checks per run, without smoothing, downsampling or checkpoint reconstruction). Figure 15 subtracts the same-replicate fresh run at each point. Intermediate points are states of one run under its own schedule, not separately annealed endpoints, and development data also fixed recipes and repeat candidates, so these curves are descriptive and carry no intervals. In the source study (three original replicates), the $(1/2, 32)$ excess loss is 0.116 bpb at the midpoint and changes little during the final decay (0.142 to 0.140), whereas the $(1, 8)$ excess loss rises from 0.068 at the midpoint and from 0.113 to 0.169 during the decay; the two swap order only in the final tenth. The matched-content difference is within 0.0003 of zero at 10% and 20% and then rises steadily to 0.0039, close to the six-replicate test estimate. In the scale study, the 184.6M-minus-45.1M

$(1/2, 8)$ gap is negative from 20% onward, near -0.012 from the midpoint, and ends at -0.0136 (test -0.0138).

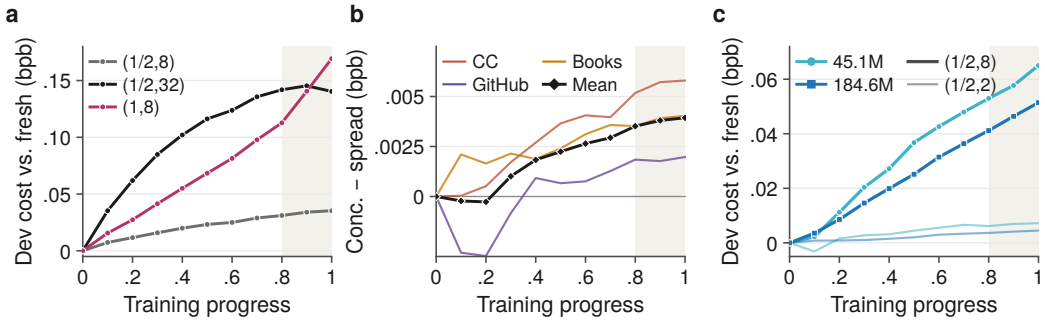


Figure 15: **Replacement excess loss builds up during training (development loss).** Each point subtracts the same-source, same-replicate fresh run at the same training progress; the learning rate decays linearly over the shaded final 20%. (a) Source study, 45.1M: mean over three sources and three original replicates. (b) Matched-content difference $(1/3, 4)$ minus $(1/2, 2)$ per source and mean, three original replicates. (c) Scale study at matched processed TPP, mean over two sources and three replicates; lighter 45.1M, darker 184.6M. Descriptive; no intervals.

E Conditional Compute-Allocation Summaries

The eight approximate IsoFLOP bins (Figure 16) group observations within half-octaves of compute, with up to 8.33% budget variation. Six interior and two boundary minima are retained. Their descriptive model-size slopes are 0.697 over the six minima and 0.995 over the three $R = 4$ minima; corresponding surface-derived slopes are 0.872 and 0.930. This is not a strict matched-compute intervention or a universal exponent. A fixed-model R minimum and a fixed-compute (N, R) allocation solve different problems.

Surface-derived fresh-equivalent fractions η_{surf} use $L_{\text{fresh}}(N, V) = E + A/N^{\alpha} + B/V^{\beta}$ at the same model size, unlike the local anchor-based η_{loc} of Eq. (1). CEG, the *credited extra-token gain*, is the fresh-token credit per extra processed token under fixed-corpus repetition; it is a token ratio and does not measure wall time or complete system FLOPs. The benefit-only reference is $\eta_{\text{surf}} = 0.949/0.799/0.571/0.336$ for $R = 2/4/8/16$, independent of N by construction. At $U = 1$B the full model gives second-pass $\eta_{\text{surf}} = 0.946/0.803$ at 59M/2B; at $R = 8$, 0.548/0.109. These are fit-derived values, not additional fresh-anchor measurements at each scale, and their scale dependence is a property of the parameterization. Under unique caps of 1/4/16B, analytic fresh-cap exhaustion estimates are 1.02×10^{16}, 1.43×10^{18} and 2.00×10^{20} FLOPs. The lowest transition is below the trained size range and the 16B cap exceeds the trained unique budget. Display markers requiring $R^* > 1.02$ are different from analytic transition points. The capped-loss gap uses fresh training at the same winning model size, not a separately reoptimized unrestricted frontier.

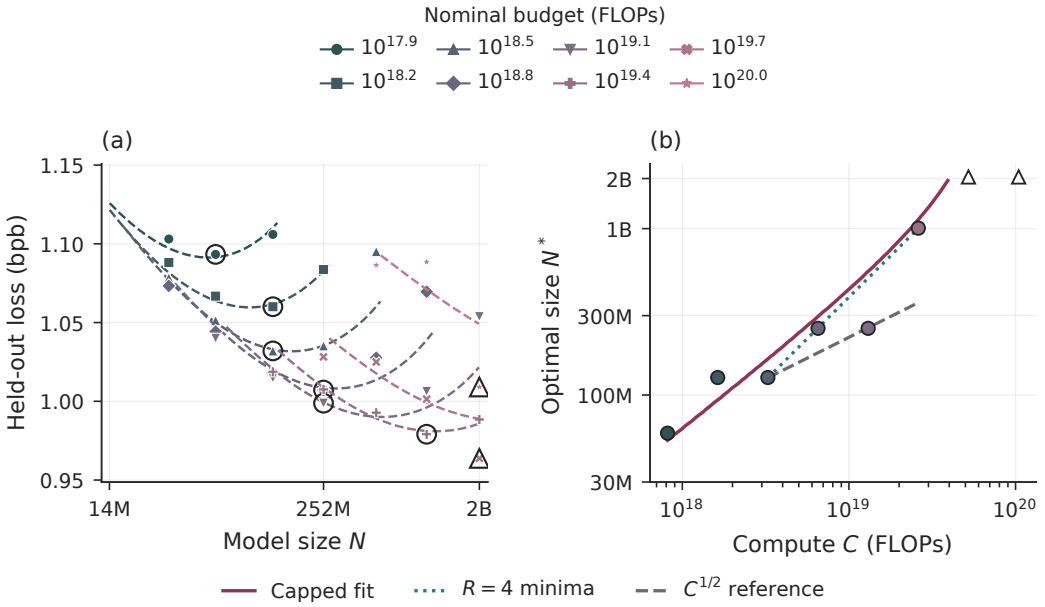


Figure 16: **Approximate IsoFLOP profiles at $U = 1$B.** Points are grouped by nearest half-octave in $\log_2 C$, with at least three per bin and up to 8.33% actual-budget variation. (a) Retained observations and fitted curves at bin-center compute. Outlined circles mark six interior grid minima and triangles two boundary-limited minima. (b) Selected model size versus bin-center compute. The six-point slope is 0.697 after excluding architecture-mismatched/$R = 4$. The dotted fit uses the three interior minima at $R = 4$ (slope 0.995); the dashed line is a slope-$1/2$ reference. The sparse-grid exponents are descriptive and have no estimated confidence interval.

F Post Hoc Checks of the Frozen Surface

The two checks below were added after all paired results were frozen. They evaluate the 73-point surface of Eq. (2) with the frozen parameters of Appendix C and the $12Ld^2$ size proxy; nothing is refit. The surface was fit to the sweeps' own byte-normalized CommonCrawl metric (before quantization), whereas the paired tests use ordinary-target bpb on their own holdouts, so the comparisons indicate direction and scale rather than calibrated agreement.

Per-sample effective data and the matched-content contrast. Applying the saturation $g(r) = 1 + R_D[1 - e^{-(r-1)/R_D}]$ to each 1,024-target sample instead of at corpus level gives $D_{\text{eff}} = 1024\,[196{,}608\,g(2) + 393{,}216] = 784.70$M targets for $(1/2, 2)$ and $1024\,[65{,}536\,g(4) + 524{,}288] = 751.36$M for $(1/3, 4)$. The benefit term $BD_{\text{eff}}^{-\beta}$ then predicts that $(1/3, 4)$ is worse by $+0.0046$ bpb, against the observed $+0.0039$ $[0.0033, 0.0045]$. The penalty term is defined by a full-corpus repeat count and is omitted. With the saturation constant $R_D = 15.4$ of Muennighoff et al. (2023) instead of the fitted 4.545, the prediction falls to $+0.0016$. Sample-wise diminishing returns thus account for the sign and order of magnitude of the matched-content effect, but the magnitude depends strongly on R_D; the check is consistent with, not a test of, effective-data descriptions.

The surface outside its range. The scale study's full-corpus design uses $U = 50.3$M and 206.0M unique targets for the 45.1M and 184.6M models (proxy sizes 49.2M and 201.3M), about 5–20 times below the smallest unique budget in the fitted sweep (1B). Over continuous $R \in [1, 32]$ the surface places the optimum at $R^* \approx 8.3$ and 6.3 (8.7 and 6.6 with exact non-embedding counts), with predicted improvements over one pass of 0.38 and 0.17 bpb, and predicts that 12 passes are still better than one by 0.37 and 0.14 bpb. The measured means are lowest at four passes in all four strata, with improvements of 0.086–0.112 bpb, and 12 passes are worse than one pass by 0.19 and 0.59 bpb at 45.1M (CommonCrawl/GitHub) and by 0.06 and 0.39 bpb at 184.6M. Because $z = (R - 1)N/U$ is the same at both sizes when N/U is matched, the fitted penalty is identical across the two models,

while the R-dependence of the benefit term scales as $U^{-\beta}$; the surface therefore predicts an earlier optimum at the larger model, but it overstates the benefit of continued reuse at small U by a wide margin. This out-of-range failure is why §6.3 treats the surface as a conditional description.

G Reproducibility Checks and Their Limits

Configuration, exposure, final-checkpoint and raw-metric identities were checked for all scientific runs in the two allocation studies, and checkpoint, development and configuration files were write-protected during testing. A cross-implementation one-update comparison and an independent cold-start continuation did not reproduce exactly. A continuation from the same checkpoint did, once deterministic settings were restricted to the technical workers. The scientific path separately passed its specified loss/gradient and distributed smoke checks. We therefore claim no bitwise equivalence of production training across implementations or cold starts.

One failed test startup produced no evaluation result but overwrote the process metadata of the first scale-replacement run, which cannot be recovered; that run's checkpoint, development metrics and configuration were unaffected, and later tests were isolated from training outputs. This limits provenance, not the reported paired arithmetic. Full hashes and machine identifiers are in the supplementary material.

Fixed-budget sweep completion. The 252M and 472M fixed-budget ladders of Figure 2b were run at $D = 4$B for $R \leq 8$. Their $R = 16$ runs expose whole 100M-token shards, so $U = 0.2$B and $D = 3.2$B, and had no same-budget one-pass reference. After the sweep was frozen we trained the two missing $D = 3.2$B one-pass runs and the two $R = 32$ runs ($U = 0.1$B) with the archived training script, configuration, data order and seed on one 8-GPU H800 node; only compiler padding was disabled for a newer PyTorch. A rerun of the frozen 252M/$R = 16$ endpoint gave 1.4319 against 1.4279 bpb ($+0.0041$), and the plotted $R = 16$ points remain the frozen runs. The completed excess losses are 0.376/0.519 bpb at $R = 16$ and 2.106/2.870 bpb at $R = 32$ for 252M/472M. Both $R = 32$ runs rise steadily in held-out loss while training loss keeps falling, with no loss spikes. These four runs enter only Figure 2b and the sentence reporting it; no fitted surface or other analysis uses them.

Fixed-unique sweep completion. The excluded 511M/$R = 4$ endpoint of the $U = 1$B CommonCrawl ladder (Appendix C) was retrained after the sweep was frozen, with the architecture and schedule of the other 511M ladder runs, the archived training script and the same seed, on the same node and with the same padding change. Its CommonCrawl held-out loss is 0.9940 bpb, between the $R = 2$ (1.0277) and $R = 8$ (0.9928) endpoints. The run enters only Figure 7a; the fitted surface, the ladder minima of Figure 5 and the power law fitted to them use the original records.

H Tokenization Intervention: Protocol and Inference

Table 9: **Primary tokenization contrasts and unequal actual budgets.** Primary effects average the paired policy difference over the five horizons within each replicate, with 95% paired-t CIs ($n = 3$). The prespecified benefit rule requires mean ≤ -0.005 and upper CI < 0; neither study meets it. Budget ranges span all three replicates and five horizons, not uncertainty intervals. E counts canonical position-target exposure equivalents, D actual targets and P padded slots.

Primary contrast	Mean bpb	95% CI	Benefit rule
20%-off minus On	-0.002778	$[-0.004151, -0.001406]$	Not met
50%-off minus 20%-off	+0.012169	$[+0.011419, +0.012919]$	Not met

Policy	$\gamma = D/E$ range	$\rho = P/E$ range
Off	1.0000003–1.0000003	1.03255–1.04297
On 0.05	1.0485230–1.0486052	2.00000–2.00000
20%-off	1.0387904–1.0388844	1.80469–1.80990
50%-off	1.0242616–1.0243027	1.51432–1.52148

For the final-20%-off policy, updates $1, \ldots, \lfloor 4K/5 \rfloor$ use dropout; for final-50%-off the boundary is $K/2$. All policies share the same final $\lceil K/5 \rceil$ learning-rate decay, so the 50% boundary is not a learning-rate boundary. Under the separately frozen serial rules, all 14 original and 19 extension secondary contrasts are reported with pointwise and Bonferroni intervals but, because neither primary meets its benefit rule, are interpreted descriptively. Extending the off period to 50% changes little at low R but raises loss by $+0.0126$ and $+0.0554$ relative to final-20%-off at $R = 8$ and 16; at $R = 16$ it is also $+0.0933$ above always-on. Development measurements around the switch separate a transient response from the final outcome: at $R = 16$ the final-50%-off mean falls from 1.2317 at the midpoint to 1.2249 at about 55% of training, then rises to 1.3639 at the end. The 55% point has no old-control counterpart and is not interpolated.

HF tokenizers 0.22.2 uses the fixed 50,257 r50k IDs and 50,000 ordered BPE merges, ByteLevel with no prefix space or normalization, and the same lossless byte-aware target-boundary adapter in every policy. Structural EOT is kept separate from literal text; invalid UTF-8 boundary bytes are preserved. The p0 training path is a new common baseline, not a reused legacy tiktoken endpoint: 33/27/29 positions in the three 98,304-position prefixes have changed tokenizations. Evaluation instead replays the original canonical IDs. Natural exact-content duplicates are not filtered to make a semantic-uniqueness claim. The observed prefixes each contain 98,304 typed byte/EOT target groups; this is not document or near-duplicate independence.

The replicates use data seeds 73001/73002/73003, initialization seeds 1337/1400/1401, and occurrence-bank labels 95001/95002/95003. All four policies share the same 16-pass permutations and corresponding stored views; the labels do not seed-reconstruct native Rust BPE-dropout randomness. Views are materialized once and hash-bound. Every sampled target is retained in fragments of at most 1,024 targets; later fragments add a context-only lookbehind. Dynamic synchronized microsteps normalize summed NLL by the whole update's global valid-target count before one optimizer step. The canonical exposure batch is 131,072 and each R has $K = 768R$ updates. Actual optimized targets include training EOT; padding and context-only bytes are accounted separately. Per-run training EOT exposure was not separately logged as a scalar and is not replaced with the evaluation EOT count.

The original 45 models each have six scheduled dev evaluations and one final test; the 15-model extension adds a seventh dev time at approximately 55% of its own horizon. The 45 tests follow a freeze of all 45 final checkpoints; the extension tests similarly follow all 15 training completions. Thus there are 375 dev and 60 test evaluations. Canonical CC dev has 4,664,883 ordinary bytes and 607 EOT targets; test has 4,698,208 bytes and 615 EOT targets, each with 1,048,576 targets. Only actual shared dev times are paired; no missing 55% control value is interpolated. The extension was motivated by visible development behavior and the earlier negative screen, with rules fixed before using base-test magnitudes for extension choices. The already studied test set and shared controls preclude pristine whole-program independence.

The original primary is the equal-five-R mean of 20%-off minus constant within each of three replicates; the extension primary substitutes 50%-off minus 20%-off. Both use $t_{0.975,2} = 4.303$. Each separately requires mean ≤ -0.005 and upper CI < 0 before its secondary family can support confirmatory claims. The original 14 and extension 19 simultaneous contrasts use critical values 16.688 and 19.455. All intervals remain visible when a gate fails, but are then descriptive; shared controls do not increase n. Normality/conditional replicate independence are assumptions, and three replicates give a minimum two-sided exact sign-flip $p = 0.25$.

A separate six-model partial-replacement development screen stopped without confirmatory models or tests; it is not part of the 60-model tokenization study. Two execution incidents in the full-corpus study are documented in the supplementary records and are not counted as additional models: a storage interruption before one model's first update, fixed by a migration that preserved payload hashes, and a resumed checkpoint that required an additional directory sync before publication. Model, optimizer, RNG and loader payloads are hash-bound; the analysis verified raw metrics and metadata but did not rescan every checkpoint. Actual token and padding budgets, resource costs and all contrasts are in the supplementary material.