FUTURE DATA ALLOCATION SHAPES REPETITION-HISTORY PERSISTENCE IN LANGUAGE MODEL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality pretraining data are finite, so repeated exposure is increasingly common. Prior work has mainly examined how repetition affects the final model. We instead ask whether predictive differences caused only by when repeated data appear persist under later training. We train matched model pairs on the same document-occurrence multiset using different repetition schedules, then continue both models on identical ordered batches. Under a fixed continuation budget, the remaining difference changes systematically with support–reuse allocation: concentrating tokens on repeated visits to narrow support can increase disagreement, whereas distributing them over broader support produces greater contraction. Token count alone does not explain this response. At 124M parameters, 15M tokens over broader support produce more contraction than 30M tokens of restricted-support reuse. A shared-first-update intervention localizes the same behavior: after an identical first update, every tested revisit increases disagreement, while matched distinct batches yield a larger one-step decrease. The allocation contrast replicates in a document-disjoint FineWeb-Edu construction and on C4, and remains positive across model scales and optimizer controls. These results show that the fate of a repetition-induced predictive difference depends on both the amount of continued pretraining and the allocation of its future data.

1 INTRODUCTION

Classical compute-optimal training assumes that additional compute is matched with additional fresh data (Qin et al., 2026), but high-quality text is finite and increasingly costly to curate (Villalobos et al., 2024; Muennighoff et al., 2023). Reusing existing documents is therefore a natural way to spend additional compute. Work in this data-constrained setting has studied how repetition affects validation loss and the value of additional processed tokens (Hernandez et al., 2022; Chudnovsky et al., 2026; Muennighoff et al., 2023), while studies of data duplication have also documented effects on memorization and privacy (Lee et al., 2022; Kandpal et al., 2022). These results describe the model produced by a repetition strategy, but leave open what happens to a predictive difference that repetition has already created when pretraining continues.

We study whether a predictive difference caused by repetition timing is reduced or amplified by subsequent pretraining. To isolate timing, we train matched model pairs from the same initialization on the same multiset of document occurrences. The two models see the same documents, occurrence counts, tokens, and optimization settings; they differ only in when repeated occurrences appear. We call the resulting predictive difference a repetition-history difference. Both models are then continued on the same ordered batches. This separates the formation of the history difference from its behavior under later training and allows us to compare different allocations of the same continuation budget.

Endpoint loss alone does not capture this pairwise behavior. Both models may improve held-out likelihood while their predictors move closer together, remain separated, or move farther apart. We therefore track the distance between their next-token distributions on fixed validation contexts, normalized by its value at continuation onset. Because the two models receive the same continuation batches, this distance measures whether later training removes or reinforces the earlier schedule difference, a response that scalar loss alone does not identify.

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.

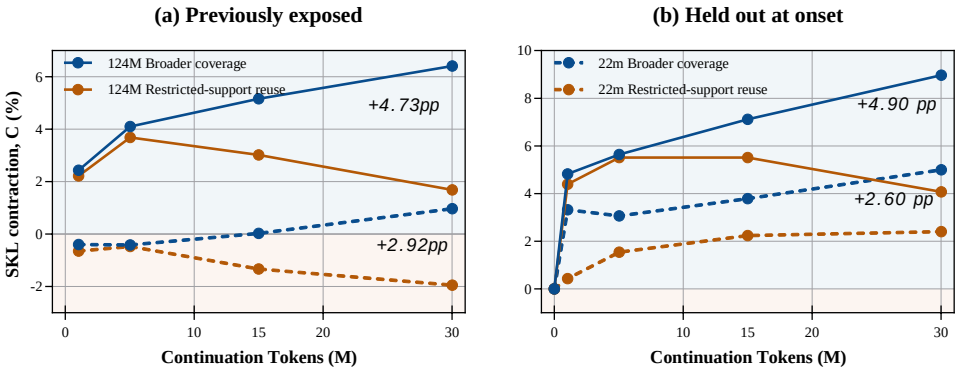

Figure 1: Broader coverage contracts repetition-history differences more strongly than restricted-support reuse. Panels show previously exposed (a) and held-out documents (b). Solid and dashed curves denote 124M and 22M models, respectively; labels give the 30M-token allocation contrast, and shading marks the sign of $C = 1 - D_t/D_0$.

Our first comparison shows that continuation length alone is not sufficient. At 124M parameters, 15M tokens spread over broader document support contract the history difference more than 30M tokens concentrated on restricted-support reuse (Figure 1). Contraction is 5.16% versus 1.68% on previously exposed documents and 7.11% versus 4.07% on documents held out at continuation onset. More processed tokens can therefore leave a larger fraction of the same predictive difference when they are concentrated on revisits.

This comparison is one endpoint of a broader pattern. Holding the continuation budget fixed at 30M tokens, we vary the fraction of a nested document support visited by the continuation from 25% to 100% (Figure 3). Under a fixed budget, broader support necessarily means fewer revisits, so the experiment measures a support–reuse allocation response rather than the isolated effect of support size. In both the primary FineWeb-Edu construction and an independently constructed document-disjoint chain, broader allocation shifts the response from divergence toward contraction. The 25%-to-100% contrast ranges from 7.65 to 10.81 percentage points across the two constructions and exposure statuses. In the document-disjoint replication, every initialization seed is strictly monotone across all five allocation levels for both symmetric KL and top-1 prediction disagreement.

We also examine the response at the level of individual updates. Both continuations begin with the same byte-identical batch. One then receives three matched, record-disjoint batches, while the other revisits the first batch. Their trajectories separate immediately afterward: every revisit increases predictive disagreement, while each matched distinct-batch update produces a larger one-step disagreement decrease. On independent validation contexts, distinct batches improve contraction by 7.25 percentage points for previously exposed documents and 6.32 points for documents held out at continuation onset. These short trajectories provide an update-level counterpart to the long-horizon allocation response.

The pattern also holds beyond the primary construction. A document-disjoint FineWeb-Edu chain reproduces the five-level trend. On C4, increasing support from 25% to 100% changes contraction from about -40% to 11% in both exposure statuses. Top-1 disagreement changes in the same direction, while every C4 branch improves validation NLL. The allocation contrast remains positive across model scales, stable learning rates, repetition positions, and shared AdamW second-moment denominators. A fixed continuation budget specifies how many tokens are processed, but not how those tokens are distributed across document identities and revisits. That allocation can determine whether an existing history difference shrinks or grows. In summary, our contributions are threefold:

- We introduce a matched-pair protocol that isolates predictive differences caused by repetition timing and tracks them under common continuation training.

- Empirically, a five-level fixed-budget sweep reveals a graded support–reuse response, with broader allocation shifting the response from divergence toward contraction.
- A shared-first-update intervention connects the long-horizon response to repeated-visit dynamics. Document-disjoint FineWeb-Edu and C4 experiments reproduce the response beyond the primary construction.

2 Related Work

Repetition under data constraints. Data-constrained scaling work distinguishes unique corpus size from processed tokens and models the diminishing value of additional epochs or other derived data (Muennighoff et al., 2023; Qin et al., 2026). Controlled repetition studies further show that validation loss depends on both the number of copies and the size of the repeated subset (Hernandez et al., 2022; Chudnovsky et al., 2026). Deduplication studies separately show that duplicated training data affects memorization and privacy leakage (Lee et al., 2022; Kandpal et al., 2022). Other work asks how much reuse can be sustained through stronger regularization (Kim et al., 2025; Xu et al., 2026). In mixed-data settings, related studies analyze repetition of scarce target or domain data alongside broader sources (Sedova et al., 2026; Li et al., 2026; Jeha et al., 2026), as well as repetition mismatch when mixture experiments are transferred across scales (Zhou et al., 2026). These studies evaluate the model produced by a repetition strategy, usually through endpoint loss, memorization, or data efficiency. Our study begins after two repetition schedules have produced different predictors and asks how much of that difference survives subsequent pretraining.

Data allocation over training. Data-mixing methods study how a training budget should be divided across domains, including settings where mixture proportions change during training or as the available set of domains changes (Chen et al., 2024; 2026; Zhao et al., 2026). Related work on sample order and curriculum construction (Dai et al., 2026; Zhang et al., 2025; Roque & Velasco, 2025) also shows interactions with learning-rate schedules and optimizer memory (Luo et al., 2026; Sweeney, 2026). Continued-pretraining methods choose data mixtures and learning-rate schedules for adaptation (Parmar et al., 2024; Wang et al., 2025), while other work replays pretraining data during fine-tuning or schedules repeated examples using memorization dynamics (Kotha & Liang, 2026; Zuo et al., 2026). These methods generally choose a data allocation or schedule to improve a single model. Our setting instead applies the same continuation stream to both members of a matched history pair. We vary how a fixed token budget is distributed across document identities and measure how this changes the predictive difference between the two models.

Forgetting and persistence of training history. Studies of memorization dynamics track exact recall over the course of training and across model scales (Tirumala et al., 2022; Wei et al., 2025). Jagielski et al. (2023) study how later updates reduce memorization and privacy leakage from earlier examples, while Krasheninnikov et al. (2025) show that information about acquisition time can remain linearly decodable from model activations. These works focus on the retention of examples or information about when they were observed. We instead track a functional difference between two complete next-token predictors trained on the same document-occurrence multiset under different repetition schedules. Repetition-history persistence measures how much of this pairwise predictive difference remains after both models receive the same continuation training.

3 Problem Setup and Experimental Design

We hold fixed a predictive difference created by repetition timing and vary the data used in subsequent training. We measure the response of a repetition-history pair h to continuation allocation A at horizon H by $C_H(A; h)$, separating the formation of the difference from its evolution under later training.

3.1 Repetition-history pairs and common continuation

Let $\mathcal{M}$ be a multiset of document occurrences. Schedules π_E and π_L place the repeated occurrences early or late in endpoint training while preserving document identities, occurrence counts, document-internal token order, and total tokens. Starting from the same initialization θ_0, with all other training

choices fixed, the two schedules produce

$$\theta_r = \mathrm{Train}(\theta_0, \pi_r(\mathcal{M})), \qquad r \in \{E, L\}. \tag{1}$$

We call $h = (\theta_E, \theta_L)$ a *repetition-history pair*. The two models have the same total document exposure and differ only in when the repeated occurrences appear.

A *continuation allocation* A specifies the document identities, multiplicities, and stream-construction rule used to spend a fixed processed-token budget B. Let $S(A)$ denote the resulting ordered stream. Both members of h receive the same realized batches from $S(A)$ under the same optimization configuration. Unless stated otherwise, optimizer and RNG states are reset at continuation onset, so the model parameters are the only difference between the two branches at that point.

3.2 Predictive disagreement and history persistence

All allocations are evaluated on fixed contexts $\mathcal{V} = (x_i)_{i=1}^{n_V}$ whose document identities do not appear in endpoint or continuation training. Let $p_{r,t}^{h,A}(\cdot \mid x)$ denote branch r's next-token distribution after t continuation tokens, and define $\mathrm{SKL}(p, q) = \frac{1}{2}[\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)]$. Predictive disagreement is

$$D_t(A; h) = \frac{1}{n_V} \sum_{i=1}^{n_V} \mathrm{SKL}\Big(p_{E,t}^{h,A}(\cdot \mid x_i), p_{L,t}^{h,A}(\cdot \mid x_i)\Big). \tag{2}$$

For fixed h and $\mathcal{V}$, all candidate allocations share the same nonzero initial disagreement $D_0(h)$. At reporting horizon H, we define the *repetition-history persistence ratio* and its complementary contraction as

$$P_H(A; h) = \frac{D_H(A; h)}{D_0(h)}, \qquad C_H(A; h) = 1 - P_H(A; h). \tag{3}$$

Thus $P_H < 1$ means that subsequent training has reduced the incoming difference, whereas $P_H > 1$ means that the difference has grown. We report C_H because zero directly separates contraction from divergence; $C_H = 1$ denotes agreement on $\mathcal{V}$. Persistence depends on both the history pair and the continuation allocation. We omit h and H when they are fixed and replace H with t for a trajectory. Top-1 disagreement is normalized in the same way; branch-level held-out NLL is reported separately and does not define pairwise persistence.

3.3 Fixed-budget support–reuse allocations

For exposure status $z \in \{\text{exposed}, \text{held-out}\}$, let $\mathcal{U}_z$ be a fixed candidate pool. Exposed identities occur in both endpoint histories, whereas held-out identities occur in neither. These labels refer to continuation identities; evaluation identities remain disjoint from endpoint and continuation training in both strata. The support of allocation A is the set of identities it visits, and $f(A)$ denotes their share of $\mathcal{U}_z$'s tokenized mass. At a fixed processed-token budget B, the family $A_{f,z}$ defines $f \mapsto C_H(A_{f,z}; h)$. Smaller support entails more revisits, so this is a *support–reuse allocation response* rather than an isolated effect of support size.

The broader-coverage comparison is a matched-partition instance of this framework. For disjoint, equal-token partitions $\mathcal{U}_z^{(1)}$ and $\mathcal{U}_z^{(2)}$, the corresponding streams are

$$S(A_{\mathrm{cover},z}) = \mathrm{shuffle}(\mathcal{U}_z^{(1)} \uplus \mathcal{U}_z^{(2)}), \qquad S(A_{\mathrm{reuse},z}^{(j)}) = \mathrm{shuffle}(\mathcal{U}_z^{(j)} \uplus \mathcal{U}_z^{(j)}), \quad j \in \{1, 2\}. \tag{4}$$

Here $\uplus$ denotes multiset union, and both members of h receive the same realized shuffle. All streams contain the same number of tokens and updates. With h and H fixed, let $C_{\mathrm{cover}}(z) = C_H(A_{\mathrm{cover},z})$ and let $C_{\mathrm{reuse}}(z)$ be the mean contraction of the two restricted-support controls. We define the allocation contrast as

$$\Delta C_{\mathrm{alloc}}(z) = C_{\mathrm{cover}}(z) - C_{\mathrm{reuse}}(z). \tag{5}$$

Positive $\Delta C_{\mathrm{alloc}}$ means that broader allocation leaves less normalized predictive disagreement. We report $100\Delta C_{\mathrm{alloc}}$ in percentage points.

3.4 Experimental Design

Models, data, and training. We train 22M, 124M, and 355M GPT-2-style models with AdamW (Loshchilov & Hutter, 2019) on independently constructed FineWeb-Edu corpora (Penedo et al., 2024); the cross-corpus experiment uses C4 (Raffel et al., 2020). Endpoint training processes 300M, 2B, and 6B GPT-2 BPE tokens, corresponding to 13.4, 16.1, and 16.9 tokens per parameter. Repeated documents account for about 5% of these tokens. Each appears 64 times at 22M and 16 times at the larger scales, within windows of width $0.5T$ centered at $0.25T$ or $0.75T$. Appendix B gives the complete configurations.

History-pair construction. For each initialization seed, the two endpoints share the base corpus, sampled repeated subset, occurrence counts, and total processed-token budget. We move the repeated occurrences between matched windows while preserving document order within the repeated and nonrepeated subsequences. The primary comparison uses windows centered at 25% and 75% of endpoint training. The position-family experiment uses the same construction at seven centers from 25% to 75%, so its 21 pairs reuse endpoints rather than requiring independently trained models for every comparison.

Continuation protocols. We form equal-token continuation partitions by matching documents within log-length and reference-NLL strata; validation identities are disjoint from endpoint and continuation training. At 22M, a primary nested chain and an independent document-disjoint chain use $f \in \{0.25, 0.40, 0.60, 0.80, 1.00\}$. The 25% and 100% conditions visit their selected supports approximately four times and once, respectively. Matched-partition continuations use 30M tokens at 22M and 124M and 90M at 355M. C4 uses independently constructed endpoints, continuation pools, and validation data.

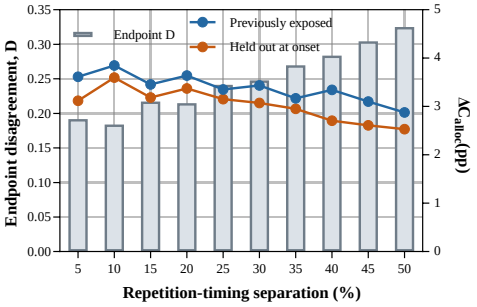

Figure 2: Allocation contrast across repetition histories. Bars show endpoint disagreement on the left axis; curves show the 30M-token allocation contrast on the right axis. Position pairs are averaged within each seed before averaging across the three initialization seeds.

To examine the response one update at a time, we start from 22M history pairs and give both arms the same byte-identical first batch. One arm then receives three matched, record-disjoint batches, while the other revisits the first batch three times. For a fixed probe Q and condition $c \in \{\text{distinct}, \text{repeated}\}$, let $D^Q_{k,c}$ denote the disagreement after update k. We define the one-step disagreement decrease as

$$\delta_{k,c} = D^Q_{k-1,c} - D^Q_{k,c}. \tag{6}$$

Positive values indicate that the update reduces predictive disagreement. The first response is identical by construction, so updates 2–4 compare extending support with revisiting the first batch. We use an 8K-target update probe and a separate set of 131K validation targets. A shared-denominator AdamW control tests whether branch-specific adaptive scaling accounts for the allocation contrast.

Evaluation. Every allocation starting from a given history pair uses the same frozen validation contexts, and D_0 and D_H are computed on the same ordered targets before normalization. Primary symmetric-KL estimates use 1,048,576 targets, while final NLL uses approximately five million. We also report top-1 disagreement, which changes the observable while preserving the same paired evaluation design. Checkpoints are evaluated at 1M, 5M, 15M, and 30M continuation tokens, with 60M and 90M added at 355M.

Initialization seed is the replication unit. Matched-partition comparisons use five seeds at 22M, three at 124M, and two at 355M; the graded chains, local intervention, and C4 experiment use three seeds each. At 124M, one initialization seed selects the continuation learning rate, while the other two are reserved for confirmation. Effects are computed within seed and summarized by mean and sample SD.

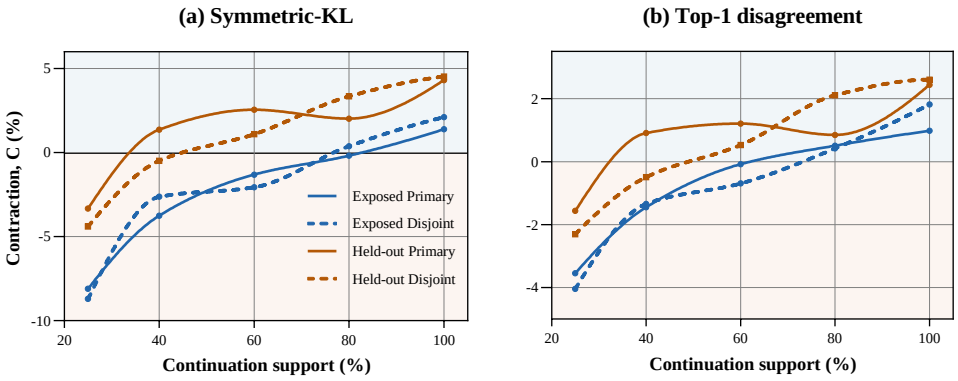

Figure 3: Fixed-budget support–reuse response. With a 30M-token continuation budget, support varies from 25% to 100%. Panels show contraction in symmetric KL (a) and top-1 disagreement (b). Blue and orange denote previously exposed and held-out documents, respectively, while solid and dashed curves denote the primary and document-disjoint constructions. Curves are means across three initialization seeds; the zero line separates divergence from contraction.

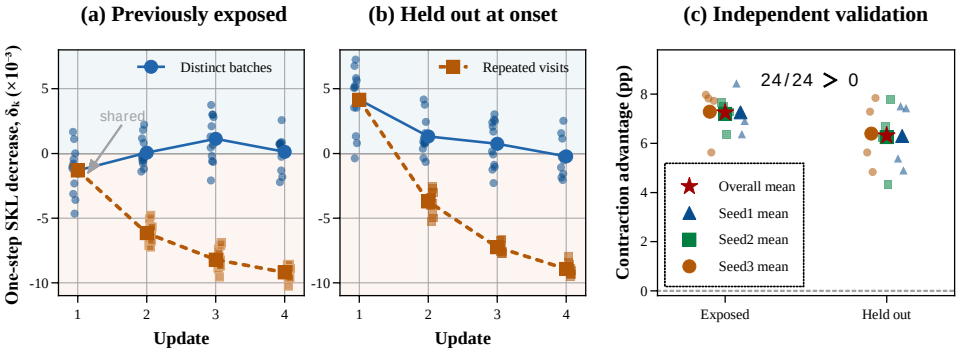

Figure 4: Local response to distinct batches and repeated visits. Panels (a) and (b) show $\delta_{k,c} = D^Q_{k-1,c} - D^Q_{k,c}$ for previously exposed and held-out documents. After a shared first batch, the distinct arm receives three matched record-disjoint batches, while the repeated arm revisits the first batch. Panel (c) shows the distinct-minus-repeated contraction difference on independent validation contexts. In (a) and (b), pale points are nested stimulus observations and curves show their means. In (c), shapes denote seed means and red stars denote overall means; initialization seed is the replication unit.

Nested stimuli and validation targets are treated as measurements rather than independent replicates. Appendix B reports support diagnostics, run counts, and evaluation horizons.

4 Results

4.1 The allocation contrast holds across repetition histories

The allocation contrast extends beyond the primary 25%–75% repetition-history pair. We train seven matched endpoints with window centers from 25% to 75% of endpoint training and compare all 21 position pairs (Figure 2).

Endpoint disagreement generally increases with the separation between repetition timings. The allocation contrast, however, does not depend on choosing a widely separated pair. At 30M tokens, it is positive for every evaluated position pair, initialization seed, and exposure status, including adjacent timing pairs, and ranges from 2.04 to 4.73 percentage points. Mean D_0 rises from approximately 0.18–0.19 at separations of 5–10 percentage points to 0.325 at a 50-point separation. Even at the adjacent 5-point separation, the mean allocation contrast is 3.62 and 3.12 percentage points in the two strata. The magnitude of the history difference at formation therefore does not by itself predict the strength of the later allocation response.

Figure 1 provides a direct counterexample to an explanation based only on continuation length within the primary 25%–75% pair. For the two 124M seeds excluded from learning-rate selection, 15M tokens drawn once from both matched partitions yield 5.16% contraction for previously exposed documents and 7.11% for documents held out at continuation onset. Processing 30M tokens by revisiting one partition twice yields only 1.68% and 4.07%, respectively. The 15M broader-coverage run exceeds the 30M reuse mean for both seeds and both exposure statuses. The allocation contrast is also larger at 30M than at 15M in both exposure statuses. It remains positive for previously exposed documents, showing that the contrast does not require continuation identities that are new to the endpoint models.

4.2 History response varies systematically with allocation

We next vary support over five levels while fixing the continuation budget at 30M tokens (Figure 3). In the primary chain, increasing support from 25% to 100% changes mean symmetric-KL contraction from -8.11% to 1.39% for previously exposed documents and from -3.33% to 4.32% for documents held out at continuation onset. The corresponding endpoint contrasts are 9.50 and 7.65 percentage points. Top-1 disagreement changes in the same direction, by 4.53 and 4.00 percentage points.

An independently constructed document-disjoint chain shows the same pattern. Its symmetric-KL contraction changes from -8.71% to 2.11% for previously exposed documents and from -4.39% to 4.53% for held-out documents, giving endpoint contrasts of 10.81 and 8.92 percentage points. All 12 seed-by-stratum-by-metric trajectories in this construction are strictly increasing (Spearman $\rho = 1$). Every primary-chain seed also has a positive endpoint contrast. The exposed trajectories are strictly increasing, while the held-out trajectories remain strongly monotone ($\rho \in \{0.7, 0.9\}$).

Both constructions cross $C = 0$ within the measured support range. The same token budget can therefore either enlarge or reduce the incoming history difference, depending on how it is allocated. Broader coverage and restricted-support reuse are better viewed as endpoints of a graded allocation response than as an isolated binary comparison.

4.3 Repeated visits increase disagreement after the shared update

We examine the allocation response at the level of individual updates. In Figure 4, the distinct-batch and repeated-batch arms receive the same first update and therefore have identical first-step responses for every seed, exposure status, and stimulus construction. Their trajectories separate as soon as the training data differ. Across the three initialization seeds, every post-shared revisit increases probe disagreement, and the corresponding distinct batch has a larger $\delta_{k,c}$ for every nested stimulus comparison. By update 4, mean $\delta_{4,c}$ under repeated visits is -9.20×10^{-3} and -8.93×10^{-3} in the two strata, compared with 0.13×10^{-3} and -0.23×10^{-3} under distinct batches.

The separation also appears on validation contexts disjoint from the update probe. Relative to repeated visits, distinct batches improve contraction by 7.25 percentage points for previously exposed documents and 6.32 points for held-out documents. The contrast is positive for all three initialization seeds, with every nested stimulus comparison agreeing in direction. The long-horizon allocation response therefore has a corresponding local pattern: after the shared first update, revisiting the same batch increases the history difference, whereas matched distinct batches produce larger one-step disagreement decreases.

The intervention isolates the point at which the two allocations begin to differ. Before update 2, the two arms share the same incoming endpoints, probe, optimizer configuration, and byte-identical first batch. Subsequent batches are matched in size and construction; the relevant difference is whether an update uses a fresh matched batch or revisits the first batch. The immediate separation at update

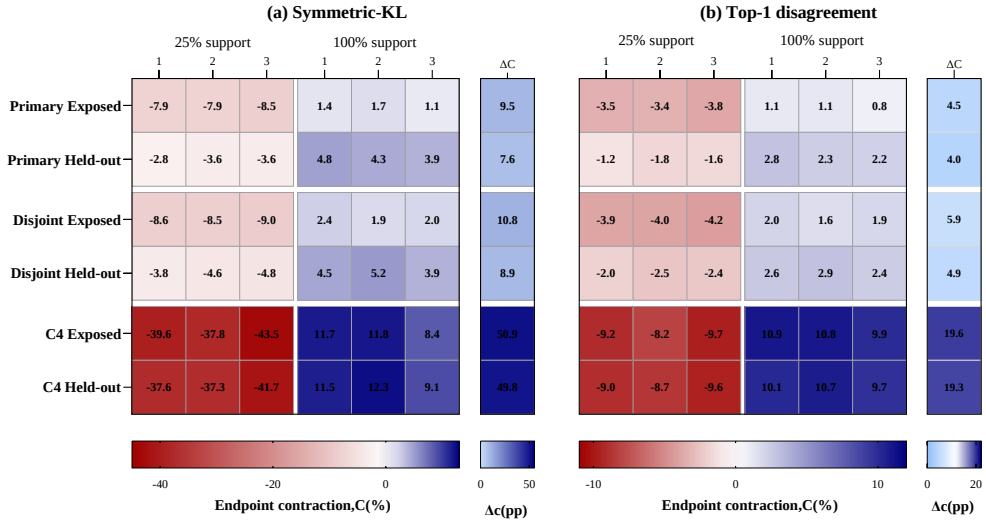

Figure 5: Endpoint allocation contrasts across support constructions and corpora. Rows show the primary FineWeb-Edu, document-disjoint FineWeb-Edu, and C4 constructions for each exposure status. Within each row, cells show three seeds at 25% and 100% support under the same 30M-token budget; the adjacent strip shows $\Delta C_{25\to100} = C_{100\%} - C_{25\%}$. Panels report symmetric-KL contraction (a) and top-1 disagreement contraction (b). Each panel uses separate color scales for endpoint contraction and the adjacent contrast.

2 therefore connects the long-horizon allocation response to the sequence of document identities supplying the updates. The same separation on an independent validation set also shows that the effect is not confined to the probe used for the one-step measurement.

4.4 The allocation response replicates on C4

Figure 5 compares the two FineWeb-Edu constructions with a separately constructed C4 experiment. On C4, increasing support from 25% to 100% changes mean symmetric-KL contraction from -40.31% to 10.63% for previously exposed documents and from -38.84% to 10.96% for documents held out at continuation onset. The corresponding contrasts are 50.94 and 49.81 percentage points, with the same sign for all six seed-by-stratum comparisons. Top-1 disagreement gives parallel contrasts of 19.56 and 19.26 percentage points.

The larger C4 percentages are not caused by a small normalization denominator. Endpoint disagreement D_0 ranges from 0.279 to 0.289 across seeds. At 30M tokens, mean absolute disagreement is 0.400 and 0.396 under 25% support in the two strata, compared with 0.255 and 0.254 under 100% support. All 24 continuation branches also improve validation NLL relative to their endpoints, by 0.282–0.420 nats per target. Concentrated reuse can therefore increase pairwise disagreement even while both branches improve as language models.

The C4 contrast is roughly five times the FineWeb-Edu endpoint contrasts despite a similar scale of D_0. The response magnitude is therefore regime-dependent, while its direction replicates across the two corpora.

4.5 The contrast persists across scales and controls

Table 1 summarizes the matched-allocation contrast across model scales and stable continuation learning rates. All 24 model–seed–stratum comparisons are positive at their stated horizons. Mean effects are 2.60–2.92 percentage points at 22M and 4.71–4.91 points at 124M. At 355M, the magnitude varies with the continuation learning rate, but the sign is unchanged across both rates, strata, and seeds.

The 30M horizon at 124M and the 90M horizon at 355M correspond to similar relative continuation budgets of 0.241 and 0.254 tokens per parameter.

Table 1: Allocation contrasts across scales and continuation learning rates. Contraction is reported in percent and ΔC_{alloc} in percentage points (mean $\pm$ sample SD). Horizons are 30M tokens for 22M and 124M models and 90M tokens for 355M models.

Model	LR	Status	C_{cover}	C_{reuse}	ΔC_{alloc}	$+/n$
22M	2×10^{-4}	Exposed	0.97	-1.94	2.92 ± 0.20	5/5
		Held-out	5.00	2.40	2.60 ± 0.31	5/5
124M	2×10^{-5}	Exposed	6.68	1.97	4.71 ± 0.03	3/3
		Held-out	9.21	4.30	4.91 ± 0.12	3/3
355M	1×10^{-5}	Exposed	8.15	-0.44	8.59 ± 0.19	2/2
		Held-out	10.63	3.09	7.54 ± 0.13	2/2
355M	5×10^{-6}	Exposed	7.93	3.91	4.01 ± 0.17	2/2
		Held-out	10.41	6.90	3.51 ± 0.02	2/2

The 355M trajectories show how the contrast develops over training (Figure 6). The four learning-rate–stratum means are close to zero at 1M tokens (-0.31 to -0.05 percentage points), become positive by 5M, reach 3.31–7.72 points at 30M, and remain at 3.51–8.59 points through 90M. Appendix C reports the complete horizon-level aggregates.

Partition- and optimizer-matched controls preserve the contrast. The allocation contrast against each matched reuse partition is positive for every 22M and 124M initialization seed (Appendix Table 4). Sharing AdamW's second-moment denominator across the two branches also preserves the contrast: mean effects are 2.98 and 2.83 percentage points in the exposed and held-out strata, with every seed positive. Branch-specific adaptive scaling is therefore not required for the contrast in this intervention. Retaining the endpoint optimizer and RNG states produces a mean exposed contrast of 2.82 points, compared with 2.88 under matched resets. At the second stable 124M learning rate, the two confirmation seeds yield exposed contrasts of 2.89 and 2.65 points and held-out contrasts of 2.86 and 2.63 points. These controls change the optimizer state or step size without changing the direction of the allocation response.

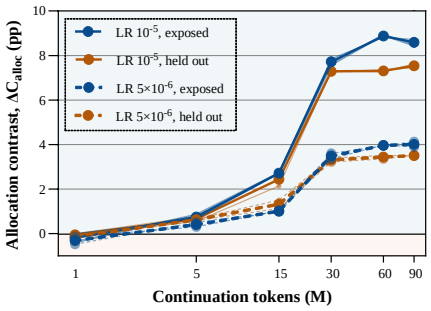

Figure 6: 355M allocation contrast through 90M tokens. Curves are two-seed means; color denotes exposure status and line style denotes the learning rate.

Pairwise divergence coexists with improving next-token prediction. Every reported allocation improves validation NLL, including all C4 and 355M branches. For the two 124M seeds excluded from learning-rate selection, broader coverage and reuse differ in mean branch NLL by less than 10^{-4} after 1M tokens, while their allocation contrasts are already 0.21 and 0.42 percentage points. Appendix Table 5 gives the corresponding loss changes.

5 Conclusion

Repetition timing can produce different predictors from the same document occurrences. Under a fixed continuation budget, concentrating updates on repeated visits can preserve or amplify this difference, whereas distributing them over broader support produces greater contraction. We observe this pattern at the level of individual updates, across repetition-history pairs, and on independently constructed corpora. These results show that repetition-history persistence depends on both the earlier training schedule and the allocation of later data. Comparisons of data-reuse schedules should therefore report not only processed-token count but also support coverage and revisit multiplicity, since matching compute alone does not imply the same history response.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing, and to brainstorm ways to present the work. The authors reviewed all AI-assisted text and take responsibility for the final manuscript.

ETHICS STATEMENT

The experiments use established public text corpora and do not involve human participants or the collection of new personal data. The analysis concerns aggregate training dynamics rather than individual documents or people.

REPRODUCIBILITY STATEMENT

Section 3 specifies the history-pair construction, continuation allocations, metrics, and replication units. The appendix reports model and optimizer configurations, data construction and matching diagnostics, run counts, evaluation horizons, and complete control results.

REFERENCES

Mayee F. Chen, Michael Y Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. *ArXiv*, abs/2411.05735, 2024. URL https://api.semanticscholar.org/CorpusID:273950317.

Mayee F. Chen, Tyler Murray, David Heineman, Matt Jordan, Hannaneh Hajishirzi, Christopher Ré, Luca Soldaini, and Kyle Lo. Olmix: A framework for data mixing throughout LM development, 2026. URL https://arxiv.org/abs/2602.12237.

Jessica Chudnovsky, Joshua Kazdan, Noam Levi, Rylan Schaeffer, Yegor Denisov-Blanch, Bo-Yang He, Mehmet Donmez, Oluwasanmi Koyejo, and David L. Donoho. Internal data repetition destroys language models. *ArXiv*, abs/2606.24998, 2026. URL https://api.semanticscholar.org/CorpusID:289629874.

Yalun Dai, Yangyu Huang, Tongshen Yang, Yonghan Wang, Xin Zhang, Wenshan Wu, Qihao Zhao, Hao Li, Yuanyuan Gao, Kim-Hui Yap, and Scarlett Li. Demystifying data organization for enhanced llm training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2605.30334. Main Conference.

Danny Hernandez, Tom B. Brown, Tom Conerly, Nova Dassarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Thomas Henighan, Tristan Hume, Scott Johnston, Benjamin Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data. *ArXiv*, abs/2205.10487, 2022. URL https://api.semanticscholar.org/CorpusID:248986979.

Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=7bJizxLKrR.

Paul Jeha, Anastasiia Sedova, Louis Béthune, Skyler Seto, Jes Frellsen, Pierre Ablin, and Natalie Schluter. Mix, don't tune: Bilingual pre-training outperforms hyperparameter search in data-constrained settings. *ArXiv*, abs/2605.13225, 2026. URL https://api.semanticscholar.org/CorpusID:288257205.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10697–10707. PMLR, 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. *CoRR*, abs/2509.14786, 2025. doi: 10.48550/ARXIV.2509.14786. URL https://doi.org/10.48550/arXiv.2509.14786.

Suhas Kotha and Percy Liang. Replaying pre-training data improves fine-tuning. *ArXiv*, abs/2603.04964, 2026. URL https://arxiv.org/abs/2603.04964.

Dmitrii Krasheninnikov, Richard E. Turner, and David Krueger. Fresh in memory: Training-order recency is linearly encoded in language model activations. *ArXiv*, abs/2509.14223, 2025. URL https://api.semanticscholar.org/CorpusID:281332495.

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. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

Jingwei Li, Xinran Gu, Rui Dai, Xintong Hao, Chengyin Xu, Yan Wu, Shuran Zheng, and Jingzhao Zhang. Scaling domain data repetition in LLM pretraining, 2026. URL https://arxiv.org/abs/2608.14071.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kairong Luo, Zhenbo Sun, Haodong Wen, Xinyu Shi, Jiarui Cui, Chenyi Dang, Kaifeng Lyu, and Wenguang Chen. How learning rate decay wastes your best data in curriculum-based llm pretraining. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.18903. Oral.

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. *ArXiv*, abs/2305.16264, 2023. URL https://api.semanticscholar.org/CorpusID:258888192.

Jupinder Parmar, Sanjeev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Reuse, don't retrain: A recipe for continued pretraining of language models. *ArXiv*, abs/2407.07263, 2024. URL https://api.semanticscholar.org/CorpusID:271089092.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. Datasets and Benchmarks Track.

Tian Qin, Kimia Hamidieh, and David Alvarez-Melis. Bridging compute- and data-optimal pretraining. *ArXiv*, abs/2607.25271, 2026. URL https://api.semanticscholar.org/CorpusID:290635341.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Matthew Theodore Roque and Dan John Velasco. Beyond repetition: Text simplification and curriculum learning for data-constrained pretraining. *ArXiv*, abs/2509.24356, 2025. URL https://api.semanticscholar.org/CorpusID:281675296.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints. *ArXiv*, abs/2605.12715, 2026. URL https://api.semanticscholar.org/CorpusID:288257987.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise, 2026. URL https://arxiv.org/abs/2606.29554.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fa0509f4dab6807e2cb465715bf2d249-Abstract-Conference.html.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49523–49544. PMLR, 2024. URL https://proceedings.mlr.press/v235/villalobos24a.html.

Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. Learning dynamics in continual pre-training for large language models, 2025. URL https://arxiv.org/abs/2505.07796.

Johnny Tian-Zheng Wei, Ameya Godbole, Mohammad Aflah Khan, Ryan Wang, Xiaoyuan Zhu, James Flemings, Nitya Kashyap, Krishna P. Gummadi, Willie Neiswanger, and Robin Jia. Hubble: a model suite to advance the study of llm memorization. *ArXiv*, abs/2510.19811, 2025. URL https://api.semanticscholar.org/CorpusID:282272225.

Zhiwei Xu, Shihao Wu, Hanseul Cho, Wei Hu, and Yixin Wang. Data-constrained language model pretraining: Improved regularization and scaling laws. *ArXiv*, abs/2606.06888, 2026. URL https://api.semanticscholar.org/CorpusID:288984802.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. Beyond random sampling: Efficient language model pretraining via curriculum learning. *ArXiv*, abs/2506.11300, 2025. URL https://api.semanticscholar.org/CorpusID:279392162.

Kai-Yan Zhao, Zhong-Tao Miao, Akiko Aizawa, and Yoshimasa Tsuruoka. Regmix-d: Dynamic data mixing via proxy training trajectories. *ArXiv*, abs/2606.18663, 2026. URL https://api.semanticscholar.org/CorpusID:289623468.

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

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*, abs/2607.04969, 2026. URL https://api.semanticscholar.org/CorpusID:289933606.

A Appendix

B Experimental Details

B.1 Models and training

Table 2 gives the three model regimes. FineWeb-Edu documents are tokenized with GPT-2 BPE, truncated at 2,048 tokens, and terminated with an end-of-text token. Endpoint training uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, gradient clipping at 1.0, and a cosine schedule with 20% warmup. Continuations use a constant learning rate. Unless a control states otherwise, AdamW moments and RNG state are reset at continuation onset.

Table 2: Model and training configurations.

Configuration	22M	124M	355M
Parameters	22.4M	124.4M	354.6M
Blocks / width / heads	12 / 256 / 8	12 / 768 / 12	24 / 1,024 / 16
Context length	256	1,024	1,024
Batch size	256	32	32
Endpoint tokens	300M	2B	6B
Repeat count R	64	16	16
Repeated-token share	5%	5%	5%
Peak LR	1.2×10^{-2}	6×10^{-4}	3×10^{-4}
Floor LR	0	6×10^{-5}	3×10^{-5}
Continuation LR	2×10^{-4}	$\{1, 2\} \times 10^{-5}$	$\{0.5, 1\} \times 10^{-5}$
Initialization seeds	5	3	2

B.2 Continuation construction

Documents are paired within the cross-product of log-length and reference-NLL deciles and then assigned to opposing partitions. Table 3 shows that the paired supports have nearly identical token mass, document length, and reference NLL. The document-disjoint 22M construction excludes all document identities used by the primary continuation before forming a new support chain. Validation documents are disjoint from endpoint and continuation training.

For the graded study, one frozen permutation of the matched $\mathcal{U}_1/\mathcal{U}_2$ pairs defines nested supports with unique-token mass $f \in \{0.25, 0.40, 0.60, 0.80, 1.00\}$ of a 30,003,200-target maximum. Each stream cycles through its selected support for as many complete visits as the budget permits, then assigns the residual targets by a seeded permutation; document visit counts differ by at most one. The document-disjoint chain repeats this construction after excluding all primary-chain identities. C4 uses the same exact-budget construction at $f \in \{0.25, 1.00\}$, with exposed candidates drawn from endpoint identities and held-out candidates from disjoint frozen C4 shards.

Table 3: Matching diagnostics for continuation supports.

Construction	Status	Documents $\mathcal{U}_1/\mathcal{U}_2$	Tokens per partition	$\lvert\Delta\rvert$ mean length	$\lvert\Delta\rvert$ mean ref. NLL
22M primary	Exposed	18,624 / 18,624	15.002M	0.000	0.00182
	Held-out	18,624 / 18,623	15.002M	0.043	0.00081
22M disjoint	Exposed	18,693 / 18,693	14.980M	0.000	0.00039
	Held-out	18,806 / 18,806	14.980M	0.000	0.00151
124M primary	Exposed	18,673 / 18,673	15.002M	0.000	0.00001
	Held-out	18,760 / 18,760	15.002M	0.000	0.00009
355M blocks 1–3	Exposed	18,744–18,968 / same	15.000M	0.000	0.00084
	Held-out	18,676–18,727 / 18,677–18,727	15.000M	0.043	0.00188

At 22M and 124M, each matched partition contains approximately 15M prediction targets, and each matched-partition continuation contains approximately 30M. The independent 22M support construction supplies 29.96M targets. The 355M continuations extend to 90M targets. The graded allocation study fixes the budget at 30M while varying nested support fractions. The local repeated-visit study uses four stimulus constructions per initialization seed; these are nested measurements, and initialization seed remains the replication unit. Each local update contains 256 fixed 256-target records. Within each stimulus construction, 1,024 records are selected without replacement and stratified by reference NLL into four balanced, record-disjoint batches. The distinct arm receives these four batches in order, whereas the repeated arm receives the first batch four times. AdamW state is reset at the start of each arm and then evolves normally.

B.3 Evaluation and replication

Symmetric KL is accumulated from FP32 log-softmax values without materializing full-vocabulary probability tensors. Primary estimates use exactly 1,048,576 validation targets; final NLL uses approximately five million. For every seed and treatment, D_0 and D_H are evaluated on the same ordered targets before forming C_H. The 22M and 124M regimes are evaluated at 0, 1M, 5M, 15M, and 30M continuation tokens; the 355M regime additionally uses 60M and 90M. Reported token budgets count next-token prediction targets.

All summaries vary initialization seed. Matched-partition comparisons use five seeds at 22M, three at 124M, and two at 355M; graded allocation, local dynamics, and C4 use three. At 124M, one seed sets the continuation learning rate and the other two provide held-out confirmation. Corpus, occurrence multiset, schedule, and continuation random variables are fixed within each paired construction.

C Additional Results and Controls

C.1 Both restricted-support controls

The main text averages the two symmetric restricted-support continuations. Table 4 instead compares broader coverage with each one separately. Every contrast is positive, ruling out a result driven by the choice of restricted partition.

Table 4: Seed ranges of the allocation contrast against each reuse control.

Model	Status	vs. reuse 1 (pp)	vs. reuse 2 (pp)	$+/n$
22M	Exposed	1.74–2.31	3.67–4.07	10/10
	Held-out	2.43–3.53	1.69–2.30	10/10
124M	Exposed	4.28–4.53	4.95–5.08	6/6
	Held-out	4.84–5.04	4.70–5.03	6/6

C.2 Validation NLL

Table 5 reports branch-relative NLL changes after 30M tokens. Both allocations improve mean NLL in every row. Their early/late branch gaps remain similar even when their pairwise-disagreement contraction differs. At the 355M final horizon, all 48 branches improve NLL, with changes from -0.0254 to -0.0113 nats per target; all 24 C4 branches also improve.

Table 5: Validation-NLL changes and early/late branch gaps at 30M tokens.

Model	Status	Mean change (nats)		Branch gap (nats)	
		Cover	Reuse	Cover	Reuse
22M	Exposed	-0.0099	-0.0038	0.0466	0.0466
	Held-out	-0.0115	-0.0059	0.0458	0.0461
124M	Exposed	-0.0191	-0.0122	0.0135	0.0137
	Held-out	-0.0196	-0.0129	0.0132	0.0136

C.3 Optimizer-state controls

Sharing AdamW's second-moment denominator across the two members of each history pair leaves mean allocation contrasts of 2.98 ± 0.24 and 2.83 ± 0.70 percentage points in the exposed and held-out strata; all three seeds are positive in both strata. Retaining endpoint optimizer and RNG state gives a mean exposed contrast of 2.82 points, compared with 2.88 under matched resets. At the second stable 124M learning rate, both confirmation seeds remain positive: 2.65–2.89 points in the exposed stratum and 2.63–2.86 points in the held-out stratum.

C.4 355M HORIZONS

Table 6 provides the numerical values underlying Figure 6. The 124M 30M horizon and 355M 90M horizon correspond to similar relative continuation budgets of 0.241 and 0.254 tokens per parameter.

Table 6: 355M contractions and allocation contrasts at 30M and 90M tokens.

LR	Status	C_{cover} (%)		C_{reuse} (%)		ΔC_{alloc} (pp)	
		30M	90M	30M	90M	30M	90M
10^{-5}	Exposed	7.03 ± 0.04	8.15 ± 0.02	-0.69 ± 0.23	-0.44 ± 0.20	7.72 ± 0.18	8.59 ± 0.19
	Held-out	9.29 ± 0.15	10.63 ± 0.07	2.01 ± 0.19	3.09 ± 0.19	7.29 ± 0.04	7.54 ± 0.13
5×10^{-6}	Exposed	6.40 ± 0.02	7.93 ± 0.06	2.92 ± 0.15	3.91 ± 0.22	3.48 ± 0.17	4.01 ± 0.17
	Held-out	8.64 ± 0.21	10.41 ± 0.13	5.33 ± 0.09	6.90 ± 0.16	3.31 ± 0.12	3.51 ± 0.02