OPTIMIZER MEMORY: PERSISTENCE AND INTERVENTION VALUE IN LANGUAGE MODEL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimizer moments preserve gradient history, but a persistent causal effect need not identify a useful intervention. Controlled language-model adaptation separates intervention at a reached checkpoint from planning at a common warm origin. Across twelve allocations, late second-moment replacement reduces target loss by 674.54 micro-NLL while increasing reference loss by 15.60; the complete procedure nevertheless loses to higher-rate retained-state training. A separate comparison finds a small refresh benefit of 34.02 micro-NLL, but additional ordinary updates reduce target loss below an old-rich intervention when its training-time allowance is reinvested. On an unedited native trajectory, recent-history reweighting worsens both domains in all eight confirmation allocations after symmetric finite-menu suffix selection. Native graft comparisons clarify this failure: both the magnitude change and the direction change raise target loss, although their reference effects differ. Actual adaptation to WikiText-103 retains the recent-direction target penalty in one further setting. These paired interventions distinguish causal persistence, conditional marginal gain and useful continuation without establishing a general preference for older statistics or practical optimizer superiority.

1 INTRODUCTION

An optimizer checkpoint contains more than the weights used for prediction. Adam's first and second moments summarize earlier gradients and determine future updates (Kingma & Ba, 2015). A change to these tensors can leave current predictions unchanged while altering subsequent learning. This dependence is elementary. The harder question is whether a consequential component identifies an intervention worth taking, especially after a competitive continuation rule has been selected.

Language-model adaptation makes the distinction concrete. Continued pretraining can improve performance in a target domain (Gururangan et al., 2020; Ibrahim et al., 2024; Parmar et al., 2024), while changes to the training process can affect a reference domain differently (Guo et al., 2025). Optimizer comparisons are also sensitive to tuning opportunity and evaluation horizon (Choi et al., 2019; Schmidt et al., 2021; Wen et al., 2026). A favorable moment edit against one fixed learning rate therefore leaves two explanations open: the edit changes a useful aspect of gradient-history information, or an ordinary continuation control captures the relevant improvement. Neither explanation follows from a large moment-tensor difference alone.

The decision point introduces a second ambiguity. A late edit can help relative to continuing the path that produced the current checkpoint, yet the entire edited procedure can lose to a different procedure chosen before that path began. These are compatible outcomes. A comparison from the original warm state evaluates training planning; a comparison from the reached checkpoint evaluates continuation. Table 1 makes the available actions explicit; Figure 1 exposes the native-fork state intervention. Endpoint differences therefore need not define one common measure of intervention value.

The experiments use Qwen2.5-1.5B, paired future tapes and component-level interventions. Three findings form the main result. First, second-moment effects persist, but late recovery and warm-origin planning favor different actions. Second, an additional personalized readout loses to a fixed higher-rate continuation, while source refresh adds a small gain in one regime with an unresolved

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.

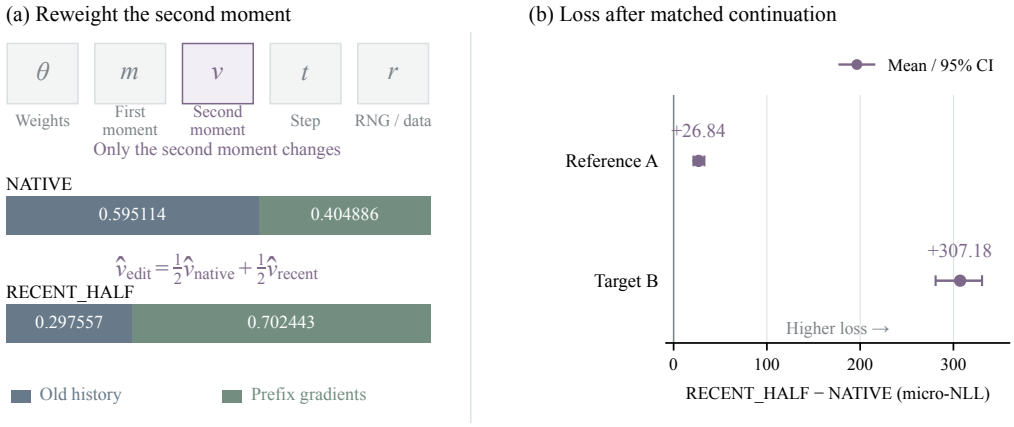


Figure 1: One state edit and its continuation effect. (a) RECENT_HALF changes only the second moment v, halving the old-history coefficient. Weights θ, first moment m, update count t and stochastic/data state r remain fixed at the fork. Bar widths encode mixture coefficients. (b) Points and lines show mean RECENT_HALF − NATIVE loss changes and crossed 95% intervals over eight confirmation allocations. Positive values indicate higher loss. Both families use identical continuation data for 1,280 updates after separate LR/STOP selection on two development allocations; both select $1.1\times$ HIGH. Figure 3 shows the individual allocation effects.

long-run target effect. Third, recent-history reweighting harms both domains on an unedited native trajectory after symmetric suffix selection. Native magnitude/direction grafts explain why a scalar step-size account is incomplete, and one additional target-domain adaptation tests the recent-direction contrast.

Both state families select the same finite-menu rate before confirmation, and every allocation remains in the analysis. Old-rich effects depend on future training, and their small target gain does not survive comparison with additional ordinary updates funded by the intervention's training-time allowance. Neither persistent state nor a non-scalar update difference prescribes a useful edit. The relevant object is a specified action, decision point and objective, compared with feasible controls.

Established grafting and optimizer recurrences separate causal persistence, conditional marginal gain and harmful continuation.

2 Related work and scope

History-dependent optimization. Adaptive gradient geometry (Duchi et al., 2011), momentum and initialization (Sutskever et al., 2013), Adam's moving averages (Kingma & Ba, 2015), and convergence analyses of their memory (Reddi et al., 2018) already establish that optimization is stateful. Sign, magnitude and variance have been studied directly (Balles & Hennig, 2018). Adafactor, Shampoo and Adam-mini alter the representation or granularity of adaptive statistics (Shazeer & Stern, 2018; Gupta et al., 2018; Zhang et al., 2025). The present experiments neither introduce these ideas nor rank alternative optimizers. The experiments intervene on specified components of an existing adaptation trajectory while separating the decision point and control family.

Nonstationarity and continuation. Resetting moments in SPAM (Huang et al., 2025) and changing clock semantics in reinforcement learning (Ellis et al., 2024) address distinct training difficulties. Those results do not imply that a native language-model checkpoint benefits from a reset. Existing continual-pretraining recipes (Ibrahim et al., 2024; Parmar et al., 2024) and optimizer-tuning studies in NLP (Gkouti et al., 2024) motivate competitive ordinary controls. The same-fork native experiment isolates a narrower question than whole-recipe optimization: which of two fixed state families performs better after equal candidate opportunity at a reached checkpoint?

Table 1: Different questions require different controls. "Same weights" is a fork property, not a claim that subsequent gradients remain equal. The HIGH retained-state control preserves native optimizer moments.

Question	Starting point and control	Evidence and boundary
Causal persistence	Same reached weights, momentum, clock and RNG; replace only v	Late cancellation helps B and harms A. No policy optimum follows.
Warm-origin planning	Common warm state; fixed higher-rate native procedure versus complete edited procedure	Higher-rate native continuation wins on A and B. Prefixes differ.
Marginal refresh gain	Common warm state; retained versus refreshed v at the same fixed higher rate	Small gain at p50/H512; B effect unresolved at p25/H1792.
Reached-state continuation	Same native fork; symmetric LR/STOP menu for each state family	Recent emphasis harms A and B in all eight confirmation allocations.

Trajectories, horizons and variability. Differentiating through training (Maclaurin et al., 2015) and short-horizon meta-optimization (Wu et al., 2018) connect local decisions with later outcomes. Concurrent finite-horizon AdamW analysis models joint state and delayed responses to minibatch perturbations (Liu & Li, 2026); the present interventions instead replace second-moment components and evaluate calibrated alternatives. Exposure and repetition also affect language-model training (Muennighoff et al., 2023; Hernandez et al., 2022). This motivates fixed nonwrapping tapes rather than a claim that one EMA coefficient dictates a universal loss half-life. Allocation and document variability are retained explicitly, following the broader concerns in Dodge et al. (2020); Bouthillier et al. (2021).

Limits. The principal evidence conditions on one pretrained ancestry and C4/FineMath adaptation; WikiText-103 supplies one further target-domain contrast, not model-family replication. Recorded exclusions do not establish unknown pretraining nonexposure. HIGH is a fixed higher-rate retained-state control, not a globally optimal rate. A second-moment (v) edit at fixed momentum (m) and clock (t) can be expressed by coordinatewise scaling, so failure of a scalar fit cannot exclude every ordinary control. Reweighting changes estimator concentration and coordinate geometry as well as temporal emphasis. Small NLL gains have no demonstrated downstream value, and unresolved intervals are not equivalence tests.

3 Decision points and controlled comparisons

Principal setting. Qwen2.5-1.5B (Qwen Team, 2025) adapts from C4 (A) (Raffel et al., 2020) toward FineMath (B), part of the data-centric training setting described by Ben Allal et al. (2025). A C4 warm phase lasts 4096 updates. Fixed nonwrapping future tapes contain either eight A/eight B windows per update (p50) or twelve A/four B windows (p25). Each window has 513 tokens and contributes 512 next-token losses. Full-token training uses FP32, eager attention, no TF32, a logical batch of 16, microbatches of four, and gradient clipping at norm one. Native AdamW uses $(\beta_1, \beta_2) = (.9, .999)$, $\epsilon = 10^{-8}$ and weight decay .01. The continuation base rate (LOW) is $\eta_0 = 10^{-6}$; the fixed higher-rate control HIGH is $1.03\eta_0$. Later development selection uses multipliers of HIGH, a separate choice. The implementation applies the rate to both the adaptive step and decoupled decay (Loshchilov & Hutter, 2019).

Interventions. A complete state is $z = (\theta, m, v, t, r)$: θ denotes weights, m and v the first and second moments, t the update count, and r the RNG, buffers, modes and data cursor. RETAIN (R) preserves that state. A source refresh V_s, $s \in \{A, B\}$, estimates mean squared clipped gradients from 16 logical probe batches at fixed weights. The refresh mixes this estimate into bias-corrected v at dose .25, then maps the result back to the unchanged clock. The probe performs no optimizer update and restores its transient state. This source-refreshed intervention differs from the native age-reweighting experiment in Section 6; the latter requires no source probe.

Table 2: Late factorial at H4096: B-derived − A-derived second moment. Means across four allocations, micro-NLL. All sixteen within-history contexts are adverse on each domain; contexts are not independent replications.

Fixed receiver weights / first moment	A mean	B mean	B range over histories
θ_A, m_A	70.79	61.66	35.43–90.74
θ_A, m_B	71.63	96.61	71.18–139.36
θ_B, m_A	87.85	85.70	72.62–98.04
θ_B, m_B	89.84	76.05	48.06–96.62

Signed effects. Let J_d be mean document-window negative log-likelihood in domain d. For two procedures beginning at state z and a common future tape U, define

$$\Delta_d(a, b \mid z, U, H) = J_d(F_H(z; a, U)) - J_d(F_H(z; b, U)). \tag{1}$$

Positive differences are adverse. Tables use micro-NLL, or 10^{-6} NLL. A HIGH-rate refresh contrast fixes HIGH and compares V_A with R. Its rate interaction subtracts the analogous LOW contrast. A favorable marginal effect does not require a favorable interaction. No arbitrary conversion combines loss, runtime and retention into one score.

Allocation and uncertainty. An adaptation allocation comprises a document-assigned warm stream, its resulting optimizer state and a fixed paired future tape, with probes when required. Repeated cells, documents and evaluations do not increase this count. Each principal bank has 1024 documents per domain; paired arms share document identities. Crossed bootstrap draws resample allocations and common document indices together. The intervals condition on one pretrained ancestry, the fixed corpora and the implemented recipe, not independent pretraining replications. The late-cancellation, HIGH-rate refresh and matched-planning studies share twelve allocations and cannot be pooled as 36 independent histories. The native confirmation has eight further allocations and two separate development parents. No equivalence or minimum useful-gain margin was prespecified. Appendix B gives the dependence and exposure boundaries.

4 Second-moment effects survive substantial continuation

A late factorial tests background dependence. Source-refreshed A and B paths first run to update 128. Crossing reached weights, first moments and second moments produces all eight $\theta/m/v$ cells per allocation, followed by the same pure-B tape through update 4096. The four-allocation experiment tests whether the second-moment ordering is confined to one receiver background. Aligning each contrast as B-derived − A-derived v gives Table 2. Every tested within-allocation contrast is positive on both domains. The ordering is consistent across the tested backgrounds, but effect magnitudes vary; four histories with four contexts each remain $n = 4$.

A late replacement asks what remains at the reached state. On a B-initialized path, a positive shadow second moment starts from the A-initialized moment but consumes the actual B-path clipped gradients:

$$\widetilde{v}_0 = v_A^0, \qquad \widetilde{v}_{k+1} = \beta_2 \widetilde{v}_k + (1 - \beta_2)(g_k^B)^{\odot 2}. \tag{2}$$

In real arithmetic, $v_k^B - \widetilde{v}_k = \beta_2^k (v_B^0 - v_A^0)$. This is Adam's existing recurrence, not a novel memory law. At $\tau = 512$, SHAM and CANCEL share weights, first moment, clock, RNG and cursor. CANCEL changes only v to the shadow. The resulting suffixes run for 1280 updates to H1792 on a fixed p25 tape.

Across twelve confirmation allocations, CANCEL − SHAM reduces B loss by 674.54 micro-NLL, crossed 95% interval $[-723.12, -627.77]$, with all twelve effects favorable. Reference A increases by 15.60, $[7.62, 23.81]$, with ten adverse allocations. Figure 2 reports the allocation-aware means and intervals. The surviving component remains consequential for the later outcome, but its effect includes feedback: gradients, momentum and weights can all change after the fork.

A different planning comparison reverses the ranking. Matched p25/H1792 ordinary controls start from the same warm parents as the cancellation procedure. CANCEL is worse than HIGH

Table 3: HIGH-rate refresh at p50/H512, twelve allocations. Arm seconds exclude shared warm, probe and preparation. The lower block uses micro-NLL and crossed 95% intervals.

Rate	State	A NLL	B NLL	Arm seconds
Warm STOP	retained	2.70858898	1.57100867	—
LOW	R	2.70341998	1.43964282	1267.45
LOW	V_A	2.70341284	1.43961045	1254.07
HIGH	R	2.70333560	1.43916048	1253.81
HIGH	V_A	2.70332844	1.43912646	1254.03
Contrast		A effect	B effect	B interval
HIGH $V_A - R$		-7.16	-34.02	$[-47.97, -22.53]$
LOW $V_A - R$		-7.14	-32.36	$[-43.47, -21.65]$
Rate $\times$ edit interaction		-0.02	-1.66	$[-7.72, 2.81]$

retained-state continuation by 1364.47 micro-NLL on B, $[1271.15, 1457.18]$, and 184.52 on A, $[167.57, 202.69]$. All twelve allocations are adverse on both domains. This is compatible with the same-fork CANCEL $-$ SHAM result: HIGH retains memory and avoids the source-refreshed path from which cancellation later recovers.

The HIGH-rate cancellation comparison also applies HIGH throughout the prefix, not only after the fork. Its HIGH $-$ LOW cancellation interaction is $+0.65$ micro-NLL on B, $[-47.38, 41.63]$, and $+0.95$ on A, $[-0.68, 2.90]$. These uncertain rate-conditioned interactions do not constitute a same-reached-state suffix-calibration test. That missing comparison motivates the native experiment below.

5 Calibration leaves a small, regime-specific marginal gain

A parent-specific readout versus fixed continuation. On six allocations, a parent-specific constant rate matches the first-step displacement norm of a source-refresh trial (Appendix A); evaluation loss selects no rate. The comparison of the personalized readout with HIGH is adverse by 299.40 micro-NLL on B, adjusted crossed 97.5% interval $[214.17, 371.76]$, and 53.36 on A, $[41.14, 64.30]$. All six history points are adverse. Mean time charged to continuation is 1602.86 seconds for the personalized readout, including 235.61 seconds of calibration, versus 1284.38 for HIGH. Thus the tested additional readout does not improve the measured quality–cost comparison. This is not evidence that native optimizer state is unnecessary: HIGH uses the moments at every step.

Direct editing remains a separate question. A four-arm LOW/HIGH by retained/refreshed comparison uses twelve allocations at p50/H512. All arms begin at one warm state; a shared A probe supplies both refresh arms. Table 3 reports absolute losses so that the marginal effect can be compared with the scale of adaptation. HIGH retained state improves B by 0.13184818 NLL relative to warm STOP. Adding A-refresh at HIGH improves it by a further 0.00003402 NLL. Its sign does not establish downstream utility.

The HIGH-rate A effect is -7.16 micro-NLL, $[-11.86, -2.64]$. One B history has a small adverse point, $+0.0468$ micro-NLL, and two A histories are adverse. The result supports a conditional mean benefit, not an allocation-wise guarantee. HIGH's ordinary rate change has a larger B effect of -482.33 micro-NLL. The rate-by-edit interaction spans zero, so the additional benefit does not establish synergy between calibration and refresh.

The benefit does not automatically transfer across recipes. The p25/H1792 LOW/HIGH by retained/refreshed comparison, using the same twelve parents, gives an additional HIGH refresh effect of -7.20 micro-NLL on B, $[-16.57, 5.34]$, and -7.98 on A, $[-12.87, -2.93]$. The B contrast is unresolved, not equivalent to zero. A direct paired raw-array comparison between the long p25 and short p50 refresh effects gives $+26.82$ micro-NLL, $[16.18, 41.04]$. Both mixture and horizon change; the contrast is a recipe difference, not a horizon-only effect.

A four-allocation p50 extension holds the first 512 batches fixed before continuing to H1792. Complete neural-prefix replays recover every H512 document loss exactly. The B refresh effect

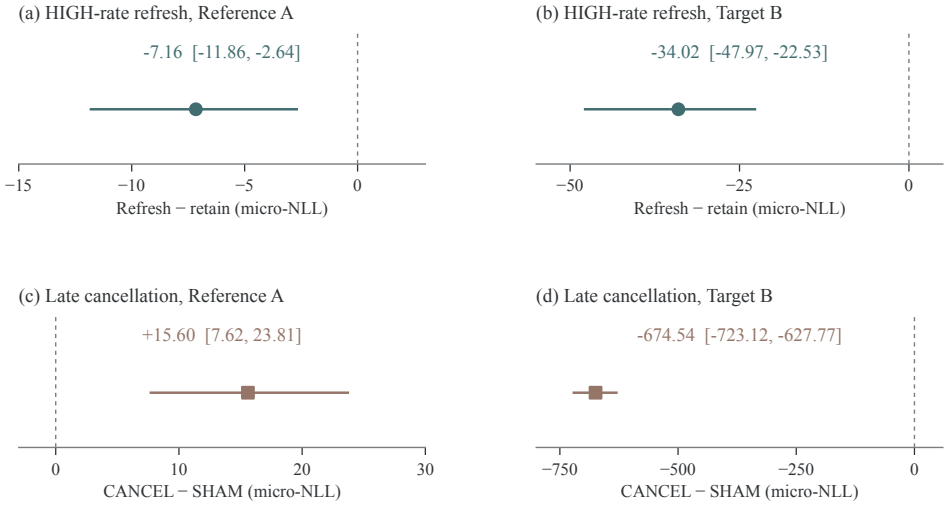


Figure 2: Two fixed questions shown separately for reference A and target B. (a,b) HIGH-rate refresh at p50/H512; (c,d) late cancellation at p25/H1792. Points are means over the same twelve allocations, and horizontal lines are marginal crossed 95% intervals. Negative effects favor the intervention. Refresh improves both means; late cancellation trades target improvement for reference harm. Panel scales and recipes differ, so the contrasts do not define a same-regime ranking.

changes from -27.80 to $+8.26$ micro-NLL, but the paired horizon contrast, $+36.06$ with interval $[-.19, 93.00]$, remains unresolved. All four histories, including the adverse outcome, enter the analysis. Appendix C gives the full table and separate selection-bank evaluation. The result limits a durable-benefit claim without establishing a population sign reversal or creating new independent allocations.

6 Native history reweighting at the same checkpoint

The earlier source refresh intentionally changes the warm second moment. A separate experiment asks whether an ordinary, unedited path contains old statistics that can profitably be downweighted. This test combines a native-history intervention with a competitive control selected at the same reached state, rather than comparing only warm-origin procedures.

Construction without a source probe. After the W4096 C4 warm phase, HIGH runs an ordinary p25 prefix for 512 updates. A positive accumulator records the accepted clipped gradients:

$$r_0 = 0, \qquad r_{k+1} = \beta_2 r_k + (1-\beta_2) g_k^{\odot 2}, \tag{3}$$

$$\widehat{v}_{\mathrm{native}} = a_\tau \widehat{v}_{\mathrm{old}} + (1-a_\tau)\widehat{v}_{\mathrm{recent}}, \qquad a_\tau = \frac{\beta_2^\tau (1-\beta_2^W)}{1-\beta_2^{W+\tau}}. \tag{4}$$

Here $\widehat{v}_{\mathrm{old}} = v_0/(1-\beta_2^W)$ and $\widehat{v}_{\mathrm{recent}} = r_\tau/(1-\beta_2^\tau)$. The real-arithmetic coefficient is $a_{512} \simeq .595114$. NATIVE retains the reached v. RECENT_HALF replaces the bias-corrected moment with $\frac{1}{2}\widehat{v}_{\mathrm{native}} + \frac{1}{2}\widehat{v}_{\mathrm{recent}}$ and converts it back to the same raw-moment clock. The relative old coefficient becomes .297557. This is a mixture coefficient, not a fraction of semantic knowledge. The intervention changes scale, variance and coordinate weighting as well as age emphasis.

Symmetric selection before confirmation. Two development parents evaluate multipliers $\{.9, 1, 1.1\}$ of the prefix HIGH rate for each state family, plus the actual fork STOP. Each family selects one common recipe minimizing mean B among candidates whose mean A does not exceed fork STOP. Ties favor STOP and then the lower rate. The zero-increase development rule is not a population safety guarantee. Both families select continuation at the upper menu endpoint, 1.1 (Appendix B). The eight confirmation allocations then evaluate only those two fixed procedures, each

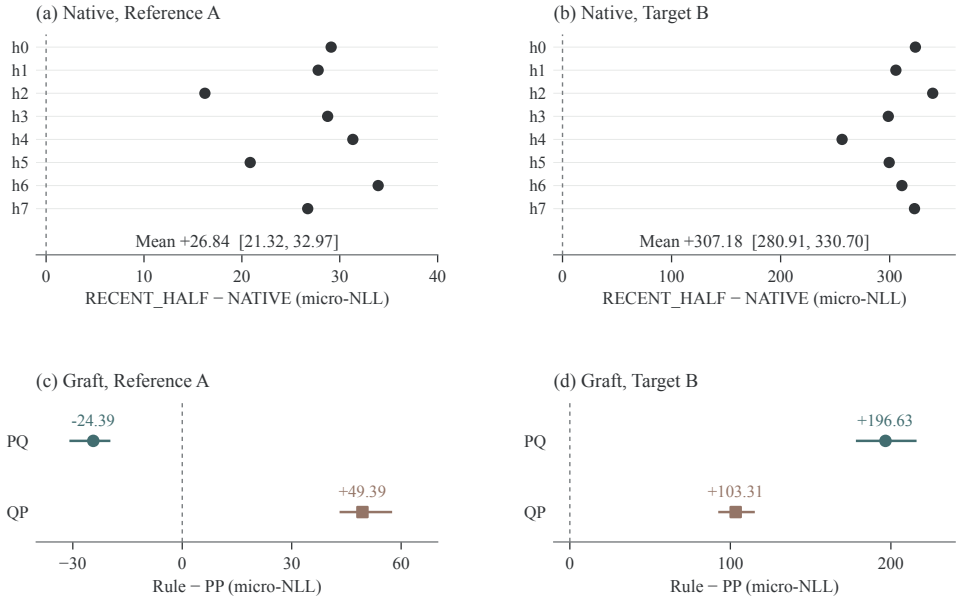


Figure 3: Domain-specific effects expose the native boundary. Effects are micro-NLL; positive is adverse. (a,b) Each h0–h7 row is one of eight allocations, with RECENT_HALF − NATIVE effects on A and B; every allocation is adverse. Printed means retain crossed 95% intervals. (c,d) The separate twelve-allocation graft changes direction (PQ, circle) or magnitude (QP, square) relative to PP. P is native and Q recent-rich. Horizontal lines are marginal 99.5% intervals from the ten-contrast family. The cohorts share an evaluation bank and are not pooled. Figure 4 preserves the acute diagnostic.

for the same 1280-update suffix. Weights, first moment, clock, RNG and future cursor are identical at each fork. No confirmation outcome changes the selected rate, horizon or reweighting dose.

Downweighting old statistics is harmful here. RECENT_HALF raises B loss by 307.18 micro-NLL, crossed 95% interval $[280.91, 330.70]$, and A by 26.84, $[21.32, 32.97]$. All eight allocations are adverse on both domains (Figure 3). Absolute mean A/B losses are $2.67758259/1.42178865$ for NATIVE and $2.67760943/1.42209583$ for RECENT_HALF. The penalty persists on new documents and at a higher selected rate, without establishing an optimal schedule (Appendix B.1).

The update changes beyond a scalar, without improving the outcome. The first suffix gradient is common because weights and stochastic state are still identical. Applying the actual moment/clock update supplies diagnostic vectors u_N and u_E, with weight decay recorded separately. For nonzero u_N, the best nonnegative scalar fit and its residual are

$$c^* = \max\left(0, \frac{\langle u_E, u_N \rangle}{\|u_N\|^2}\right), \qquad r_{\text{fit}} = u_E - c^* u_N. \tag{5}$$

Across the eight allocations, relative update change $\|u_E - u_N\|/\|u_N\|$ is 5.59–7.72%, cosine similarity is .99783–.99875, and $\|r_{\text{fit}}\|/\|u_E\|$ is 5.01–6.59%. The scalar fits are positive, .9555–.9726. These bias-corrected diagnostics use the common updated moments but are not claimed as a bitwise copy of the fused numerical update. Their nonzero residual shows that a scalar cannot exactly match this particular adaptive vector. The adverse final result shows why this acute vector difference does not prescribe a useful edit.

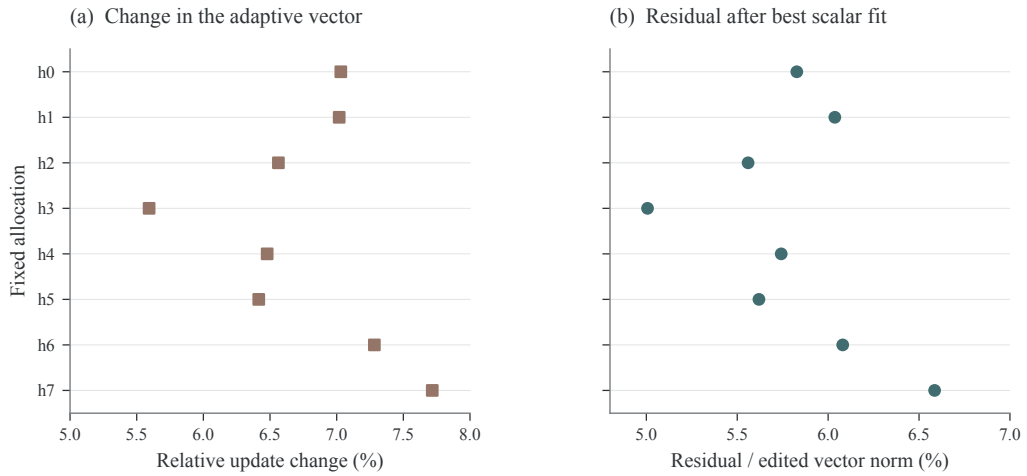


Figure 4: Common-gradient diagnostic at eight native forks. Aligned rows retain allocation identity. Squares show $\|u_E - u_N\|/\|u_N\|$; circles show the residual after the best nonnegative scalar fit, divided by the edited-vector norm. Points are individual diagnostics, with no uncertainty intervals. The different normalizations do not form an additive decomposition. The diagnostic uses common updated moments and is not a bitwise reconstruction of the numerical optimizer update.

Table 4: Native magnitude/direction controls and one target-domain transport, micro-NLL. Negative values favor the first rule. P is native and Q recent-rich; the first letter fixes the magnitude rule. Native rows use the ten-contrast 99.5% adjustment; WT103 uses a separate two-contrast 97.5% adjustment. Cohorts and interval families are not pooled.

Fixed comparison	A effect [interval]	B effect [interval]
Native PQ − PP, $n = 12$	-24.39 [$-30.93, -19.70$]	196.63 [178.21,216.00]
Native QP − PP, $n = 12$	49.39 [43.12,57.53]	103.31 [92.45,115.23]
Native interaction, $n = 12$	$-.04$ [$-.55, .53$]	-1.03 [$-7.48, 1.78$]
WT103 PQ − PP, $n = 4$	-3.18 [$-9.01, 2.51$]	296.04 [278.36,313.80]

7 Mechanism, reproducibility and cost

Magnitude and direction are established controls. Following Agarwal et al. (2020), two shadows within each live branch yield $u_P = a_P d_P$ and $u_Q = a_Q d_Q$. PP, QP, PQ and QQ select magnitude then direction. Both shadows consume the branch's actual clipped gradient; different branches subsequently have different gradients. Matching the native-shadow magnitude rule therefore does not match another branch's norm sequence. Table 4 reports the native controls and one transport; Appendix D retains the source-refresh factorial and additional direction/future contrasts.

Native recency changes magnitude and direction differently. At the original eight forks, PQ raises B by 203.30 micro-NLL and lowers A by 22.85; QP raises B by 103.48 and A by 49.69. A separate twelve-allocation graft reproduces these signs (Table 4); the cohorts share the evaluation bank and are not pooled as $n = 20$. The recent-direction target penalty survives under the native-magnitude rule, so the magnitude change alone cannot account for the native result. The interaction $J(QQ) - J(QP) - J(PQ) + J(PP)$ is unresolved on both domains. Neither exact additivity nor mediation percentages follow.

At four further fixed forks, old-rich reweighting combines native-shadow magnitude with an old-rich direction. PO − PP improves B by 318.94 micro-NLL but worsens A by 31.78 under p25 training. All four forks show this trade-off. This finite contrast supplies a conditional gain without identifying optimal age weights; the effect also depends on the subsequent training allocation.

Future allocation changes the effect, not its target sign. Replacing p25 with pure B at the same four forks holds 1280 updates, rate and total tokens fixed, while jointly quadrupling B exposure and removing A replay. PQ − PP changes from $+201.32$ to $+45.43$ micro-NLL on B and from -25.38 to $+46.20$ on A: the target penalty persists and the reference benefit becomes harm. The PO comparison reduces the old-rich B gain to 90.30 micro-NLL, while A changes from a cost to an 81.32 gain. Table 5 reports all simple effects and paired interactions. A symmetric pure-B rate menu preserves a smaller PO advantage; matched coefficient profiles also differ on B (Appendix D, Tables 16–18). Neither control isolates a pure age effect.

Table 5: Direction and future-allocation controls at the same four forks, micro-NLL. Interactions subtract the p25 effect from the pure-B effect. Superscripts retain distinct analyses: [a]original p25 six-contrast adjustment, 99.17%; [b]descriptive crossed 95%; [c]each follow-up's own two-contrast adjustment, 97.5%. No combined interval family or enlarged sample is implied.

Fixed comparison	A effect [interval]	B effect [interval]
p25 PQ − PP[a]	-25.38 [$-30.90, -20.25$]	201.32 [176.81,224.54]
p25 PO − PP[a]	31.78 [26.21,37.60]	-318.94 [$-365.51, -257.22$]
p25 PO − PQ[a]	57.16 [46.82,67.88]	-520.26 [$-586.40, -448.06$]
Pure-B PQ − PP[b]	46.20 [39.87,52.69]	45.43 [14.23,69.75]
Pure-B PO − PP[b]	-81.32 [$-89.55, -72.94$]	-90.30 [$-114.55, -72.87$]
PQ future interaction[c]	71.59 [62.82,80.77]	-155.89 [$-186.63, -134.91$]
PO future interaction[c]	-113.10 [$-123.41, -102.51$]	228.63 [174.73,273.64]

One contrast transports to a second target. Qwen/C4–WikiText-103 adaptation (Merity et al., 2017) repeats PP/PQ through actual training, not evaluation alone. PP-only selection on two development allocations chooses the upper endpoint 2 from $\{.5, 1, 2\}\times$HIGH. Four fixed confirmation allocations retain the B penalty; the A effect is unresolved (Table 4). The changed selection rule prevents a controlled FineMath–WT interaction claim. This confirms neither model-family generalization nor transport of old-rich gains.

Small loss differences require numerical anchors. The exact replay underlying Table 12 includes a predefined anchor with both endpoint weights retained. This is a complete neural-prefix replay, not merely a CPU recalculation of a table. The result supports reproducibility under the validated FP32/microbatch implementation, not cross-hardware bitwise portability. A full-batch/microbatch comparison exceeded its maximum-coordinate threshold; same-path replay does not remove that failure (Appendix D). Deleted checkpoints remain unrecoverable.

A shared experiment is not a deployment timing claim. The twelve shared warm states cost 27.261 GPU-hours, charged once; late cancellation adds 21.232 and HIGH-rate refresh 18.269. Refresh preparation includes a mean 32.44-second A-probe per quartet, excluded from arm-phase differences. Native recent accumulation adds storage and memory traffic without an extra backward pass. The native study costs 42.620 incremental GPU-hours, or 43.343 including two checkpoint-writing failures. Restoration and evaluation are included. A pure-B follow-up reinvests the development-measured training-time allowance: ordinary continuation receives 1,421 rather than 1,280 updates. OLD_PLUS at 1,280 has 986.03 micro-NLL higher B loss, adjusted 97.5% interval $[611.34, 1431.70]$; A remains unresolved (Appendix F). This is not an equal-token or exact-runtime comparison.

8 Conclusion

Second-moment history remains consequential without prescribing a useful edit. Late recovery improves a reached path yet loses to higher-rate warm-origin planning. Source refresh adds a small benefit with unresolved durability. Native recent emphasis harms both domains after finite-menu selection; grafts and one further target test constrain the explanation. Additional ordinary updates outperform old-rich reweighting on B under a development-time budget; the A contrast remains unresolved. Claims of optimizer-memory utility require a specified decision point, intervention, objective and feasible comparator.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, mathematical formulation and checking, the design, implementation and analysis of optimizer-state experiments, literature search and reference verification, figure preparation, and manuscript drafting and editing. The authors are responsible for the final content.

ETHICS STATEMENT

The study uses existing text corpora and pretrained language models, with the recorded licensing and data-exposure limitations retained. The measured changes in log-likelihood do not establish downstream semantic benefit or population safety.

REPRODUCIBILITY STATEMENT

Appendices A and B specify checkpoint restoration, paired training branches, selection rules and uncertainty analysis. Appendix E describes checkpoint availability, replay evidence and technical failures; Appendix F specifies the additional ordinary-training control. Exact replay is limited to the tested FP32/microbatch implementation; deleted earlier checkpoints cannot be reconstructed from hashes.

REFERENCES

Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling adaptive gradient methods from learning rates. *arXiv:2002.11803*, 2020. URL https://arxiv.org/abs/2002.11803.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *ICML*, pp. 404–413, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, et al. SmolLM2: When smol goes big—data-centric training of a small language model. *arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. In *MLSys*, 2021. URL https://arxiv.org/abs/2103.03098.

Dami Choi, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. On empirical comparisons of optimizers for deep learning. *arXiv:1910.05446*, 2019. URL https://arxiv.org/abs/1910.05446.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv:2002.06305*, 2020. URL https://arxiv.org/abs/2002.06305.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011. URL https://jmlr.org/papers/v12/duchi11a.html.

Benjamin Ellis, Matthew T. Jackson, Andrei Lupu, Alexander D. Goldie, Mattie Fellows, Shimon Whiteson, and Jakob N. Foerster. Adam on local time: Addressing nonstationarity in RL with relative Adam timesteps. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2412.17113.

Nefeli Gkouti, Prodromos Malakasiotis, Stavros Toumpis, and Ion Androutsopoulos. Should I try multiple optimizers when fine-tuning a pre-trained transformer for NLP tasks? should I tune their hyperparameters? In *EACL*, pp. 2555–2574, 2024. URL https://aclanthology.org/2024.eacl-long.157/.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. Efficient domain continual pretraining by mitigating the stability gap. In *ACL*, 2025. URL https://aclanthology.org/2025.acl-long.1578/.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *ICML*, pp. 1842–1850, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, pp. 8342–8360, 2020. URL https://aclanthology.org/2020.acl-main.740/.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, et al. Scaling laws and interpretability of learning from repeated data. *arXiv:2205.10487*, 2022. URL https://arxiv.org/abs/2205.10487.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *ICLR*, 2025. URL https://arxiv.org/abs/2501.06842.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv:2403.08763*, 2024. URL https://arxiv.org/abs/2403.08763.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL https://arxiv.org/abs/1412.6980.

Kang Liu and Suyan Li. Finite-horizon input-output dynamics of minibatch perturbations in AdamW. *arXiv:2608.19762*, 2026. URL https://arxiv.org/abs/2608.19762.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://arxiv.org/abs/1711.05101.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, pp. 2113–2122, 2015. URL https://proceedings.mlr.press/v37/maclaurin15.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

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:2305.16264*, 2023. URL https://arxiv.org/abs/2305.16264.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Reuse, don't retrain: A recipe for continued pretraining of language models. *arXiv:2407.07263*, 2024. URL https://arxiv.org/abs/2407.07263.

Qwen Team. Qwen2.5 technical report. *arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

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. *JMLR*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *ICLR*, 2018. URL https://arxiv.org/abs/1904.09237.

Robin M. Schmidt, Frank Schneider, and Philipp Hennig. Descending through a crowded valley—benchmarking deep learning optimizers. In *ICML*, pp. 9367–9376, 2021. URL https://proceedings.mlr.press/v139/schmidt21a.html.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, pp. 4596–4604, 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pp. 1139–1147, 2013. URL https://proceedings.mlr.press/v28/sutskever13.html.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pre-training optimizers and where to find them. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ea4f0a6ad7088c9b93f22d74c0c9e8c2-Abstract-Conference.html.

Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger Grosse. Understanding short-horizon bias in stochastic meta-optimization. In *ICLR*, 2018. URL https://arxiv.org/abs/1803.02021.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *ICLR*, 2025. URL https://arxiv.org/abs/2406.16793.

A IMPLEMENTED UPDATE AND FORK SEMANTICS

The executable state, rather than the weight tensor alone, defines a reproducible training fork. The implemented training path preserves optimizer parameter order, step counters, first and second moments, gradient buffers, module modes, RNG streams, model weights and the next logical-batch cursor. Paired arms consume the same future examples, but gradient equality is asserted only before their first divergent update. Later gradients are endogenous to each branch. Evaluation is read-only and restores training mode and stochastic state.

For accepted logical gradient g_t after clipping, the adaptive part follows

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^{\odot 2}, \tag{6}$$

$$\widehat{m}_t = m_t/(1-\beta_1^t), \qquad \widehat{v}_t = v_t/(1-\beta_2^t), \tag{7}$$

$$u_t = \widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon). \tag{8}$$

The optimizer applies decoupled decay and the adaptive step in the native implementation order. Parameters without a gradient follow the native skipped-update and clock semantics. The diagnostic expression above is a mathematical readout of updated moments, not a replacement kernel. Training updates use FP32; aggregate diagnostic calculations use FP64.

Source refresh computes a positive mean-square estimate from sixteen clipped logical probe gradients without an optimizer step. The blend occurs in bias-corrected coordinates at dose .25 and is converted to raw v using the preserved clock. Probe RNG and transient buffers are restored. A-refresh is not synonymous with a target-domain probe: A is the reference domain throughout the manuscript. The observed source-by-future behavior is not a monotone rule that a source's own statistics must improve that source.

Late cancellation maintains an A-initialized positive shadow along the actual B-initialized path. The shadow is not the second moment obtained on a separately evolving A trajectory. Its positivity avoids constructing invalid coordinates by subtracting two nearly equal tensors. At the fork, exact one-step controls check unchanged weights, momentum and stochastic state. After replacing the moment, native AdamW resumes; the experiment measures the total post-fork response.

Native recent accumulation is different. No refresh constructs the prefix. The accumulator starts at zero and receives ordinary prefix gradients. At W4096 and $\tau = 512$, the old mixture coefficient is .5951141661 before reweighting and .2975570831 after the half blend. The accumulator requires persistent memory and bandwidth. Avoiding an additional source backward pass does not make the operation cost-free. Table 6 summarizes the fixed recipe.

Table 6: Principal fixed recipe. The p25 and p50 horizons are distinct comparisons.

Model / numerical path	Qwen2.5-1.5B; FP32, eager attention, no TF32
Logical batch / microbatch	16 / 4; 513 input tokens, 512 shifted losses
Warm / prefix / late endpoint	4096 / 512 / 1792 continuation updates
AdamW / clipping	$(.9, .999)$, $\epsilon = 10^{-8}$, decay .01; norm one
LOW / HIGH	10^{-6} / 1.03×10^{-6}
Native suffix candidates	$.9, 1, 1.1$ times HIGH, plus reached-fork STOP
Principal evaluation	1024 document windows per domain

Parent-specific first-step norm matching. The readout restores the full parent after 4,096 updates and forms an A-refresh counterfactual from sixteen logical C4 probe batches. At base rate $\eta_0 = 10^{-6}$, separate trial updates on the first p50 continuation batch measure native and refreshed FP32 parameter displacements. Their ratio $r = \|\Delta\theta_{V_A}\|_2/\|\Delta\theta_R\|_2$ sets one constant rate $r\eta_0$. A third trial from the native parent checks the resulting displacement norm against the refreshed norm. Both norms must be positive, $.25 \le r \le 4$, and relative norm error must be at most 1%; inadmissible ratios are not clipped into range. FP64 accumulation covers all unique parameter coordinates. Each trial and the refresh are discarded by exact restoration before actual continuation.

The resulting rate remains fixed for all 512 updates, with native moments and rate-dependent weight decay. Norm matching asserts neither vector nor trajectory equivalence. The six multipliers are

1.010852, 1.012638, 1.008339, 1.003556, 1.016300 and 1.012406. Fixed comparison rates are $1.0135638767449517\eta_0$ and HIGH, $1.03\eta_0$; evaluation losses do not choose among these arms. The rule and bounds precede the six outcomes. Calibration includes probes, three trial updates, snapshots and restoration. Continuation time also includes setup and evaluation, while excluding shared warm construction and waiting time. The comparison does not match wall-clock budgets.

B DEPENDENCE, SELECTION AND UNCERTAINTY

Let $D_{hj} = L^a_{hj} - L^b_{hj}$ be a paired document loss difference for allocation h and document j. The displayed mean is $\widehat{\Delta} = (nM)^{-1} \sum_{h,j} D_{hj}$. A crossed bootstrap draw samples allocations with replacement and a common set of document indices with replacement. The same sampled document indices are used for every paired arm and allocation within a domain. This respects the shared bank instead of treating every allocation–document pair as independent. The intervals use the original bootstrap seeds and resampling settings; the same estimates appear in the figures.

Construction and reuse of adaptation allocations. Fixed hashes of domain and document-group identity assign source documents to histories and roles. Each original allocation uses 65,536 C4 warm windows for 4,096 updates, 256 probe windows per domain, and future banks of 21,504 C4 and 7,168 FineMath windows. A window contains 513 tokens and supplies 512 shifted losses, without concatenating documents. Domain cursors do not wrap or repeat. Within-update slot hashes place exactly twelve C4 and four FineMath windows in each p25 batch. The p50 comparison instead uses the first 4,096 windows from each existing domain bank, eight per domain per update. The twelve-allocation cancellation, refresh and rate controls therefore share parents and document banks; later matched controls are follow-ups to previously observed results.

The native eight-allocation confirmation uses separately assigned source documents and new warm states, excluding recorded prior document and content identities. No reference probe is needed. Each parent receives its own 512-update native prefix and paired 1,280-update suffix. The two development parents reuse histories 8 and 9 from the original cohort; native confirmation history 0 is a different parent from original-cohort history 0. Both cohorts share pretrained ancestry and a recipe. Fixed seeds alone do not define an allocation, and unrecorded exposure and near duplicates remain unknown.

Primary B and mandatory A are not interchangeable. A target gain with reference harm remains a trade-off. The six-allocation readout comparison retains adjusted 97.5% intervals; late-cancellation, HIGH-rate refresh and original native contrasts retain crossed 95% intervals. Native grafts use a fixed ten-contrast family with 99.5% intervals; p25 old-rich reweighting uses a six-contrast family. The PQ future interaction, PO future interaction and WT103 each use a separate two-contrast family. Descriptive simple effects do not inherit those adjustments, and the two future follow-ups do not form one simultaneous family. A narrower document-only interval cannot replace an allocation-aware interval. Table 7 records reuse.

The native selection rule evaluates the same three rate multipliers and STOP for each state family on two development histories. Eligibility requires mean A no worse than the actual fork STOP. Among eligible procedures, the minimum mean B is selected once for the entire family. Both selected multipliers are 1.1. Confirmation is therefore a paired test of two fixed procedures, not a per-history oracle or the minimum of all observed final losses. Development eligibility is finite-sample preference information, not a certified retention condition.

The retained native development results include unrounded losses, means, eligibility and time charged to each candidate. STOP reuses the fork evaluation; its zero incremental suffix time does not remove the cost of reaching or evaluating the fork.

Local exclusions use whole-document/group and normalized-content identities. Unknown pretraining exposure, near duplicates and incomplete older exposure records remain limitations. The separate twelve-allocation native comparison gives RECENT_HALF − NATIVE B $+298.91$ micro-NLL, crossed 95% $[282.01, 316.53]$, and A $+24.96$, $[20.60, 30.06]$, with all twelve adverse on both domains. Its complete graft is included in Table 4, without pooling the cohorts. Failed computations provide no effect estimate and are not assigned zero effects; their costs are reported explicitly.

Table 7: Independent units and reuse. Study labels identify comparison families, not separate populations of pretrained models.

Study	Allocations	Dependency and analysis role
Late factorial	4	Four contexts per allocation; post-hoc common-sign display
Readout and source graft	6	Same six parents across readout and graft comparisons
Refresh, cancellation, planning and rate controls	12 shared	Different estimands, same warm ancestry and banks; not $n = 48$
Native development	2	Rate/STOP selection only; excluded from confirmation n
Native confirmation	8	Fixed selected pair; common bank; no new pretraining roots
Larger native graft	12	Separate native cohort, shared bank; no pooling with the original eight
old-rich reweighting / future allocation	4 shared	Same four forks and previously examined final-evaluation bank; new arms do not increase allocation n
WT103 transport	4	Two separate development allocations; common Qwen ancestry, 512 target articles
Mixture–horizon extension	4 reused	Prefix-preserving descriptive extension; no new allocations
Functional evaluation	2 reused	Document-only uncertainty; inference-only endpoint comparison

Table 8: Complete native development menu: absolute A/B losses for the two development histories. Multipliers apply to HIGH; both families reuse the actual fork STOP. All candidates satisfy the mean-A eligibility rule. The selected 1.1 rate is the upper menu boundary for both families.

Family	Choice	A, h8	B, h8	A, h9	B, h9
Both	STOP	2.67848200	1.42111188	2.67726094	1.42175019
NATIVE	.9	2.67195076	1.40187499	2.67114222	1.40219929
NATIVE	1	2.67158464	1.40111584	2.67081377	1.40146632
NATIVE	1.1	2.67125365	1.40045644	2.67051973	1.40076821
RECENT	.9	2.67197189	1.40218543	2.67115694	1.40251364
RECENT	1	2.67160512	1.40141504	2.67082872	1.40173064
RECENT	1.1	2.67127713	1.40076277	2.67053471	1.40106268

Old-rich reweighting uses four shared native forks, with B PO − PP fixed as primary and A retained. The later future challenge reuses exactly these forks, rate 1.133×10^{-6}, p25 controls and previously examined 1024-window final-evaluation banks. In 8192 crossed draws, allocation and domain-specific document indices are shared across all four cells; B and A interactions form the two-contrast family. Pure-B replaces the 15,360 A suffix windows per allocation with additional B windows while preserving the original 5,120 B windows, yielding 20,480 B windows without wrapping. Known-role/content exclusions do not imply global nonexposure. The intervention jointly changes B exposure and A replay.

The pure-B PO comparison adds four PO suffixes at the same four forks and reuses p25 PO/PP and pure-B PP. Its primary contrast is the B difference of PO − PP effects between pure B and p25; A is mandatory. The prespecified 8192 draws and two-contrast 97.5% adjustment remain separate from the earlier PQ analysis. No new adaptation allocations, pretrained ancestries, rate selection or evaluation bank are introduced.

WT103 constructs actual target histories using C4-only warm checkpoints. Its two development allocations evaluate both PP/PQ at multipliers $.5, 1, 2$, but only PP determines the common rate; no confirmation outcome selects it. Four fixed confirmation allocations use 1024 shared A documents and 512 B articles. The selected upper boundary 2 is reported, not presented as optimal tuning. The transport establishes neither a new pretrained ancestry nor transport of old-rich reweighting or the future interaction.

Table 9: Original native endpoints on two document banks, eight fixed allocations. Absolute losses are NLL; RECENT_HALF − NATIVE effects and intervals are micro-NLL. New-bank intervals use a separate two-contrast 97.5% family.

Bank	Domain	Native	Recent-half	Effect	Interval
Original	A	2.67758	2.67761	26.84	[21.32,32.97]
New	A	2.68220	2.68222	22.49	[18.11,27.46]
Original	B	1.42179	1.42210	307.18	[280.91,330.70]
New	B	1.40556	1.40584	285.94	[257.95,310.98]

Table 10: Added development cells at $1.2\times$HIGH. Absolute NLL and incremental GPU-seconds; original candidates remain in Table 8.

Family	Parent	A	B	Seconds
Native	h8	2.67095573	1.39990118	2649.80
Native	h9	2.67025602	1.40019109	2647.65
Recent-half	h8	2.67098024	1.40011519	2648.03
Recent-half	h9	2.67027140	1.40050237	2664.90

B.1 Sensitivity to evaluation documents and finite rate menus

All sixteen original native endpoints were evaluated without training on 1024 additional document groups per domain. The new groups were fixed before endpoint readout and excluded against recorded adaptation, development, probe and evaluation identities. The first eligible groups from fixed C4 and FineMath shards each contribute one 513-token window under the original tokenizer. Exact and normalized content and URL-group checks do not establish absence of near duplicates, unrecorded use or pretraining exposure.

The paired evaluation uses the same FP32/eager microbatch-four kernel and 10,000 crossed allocation/document draws, seed 26092602. New A/B contrasts form a separate two-contrast family with 97.5% intervals; the original crossed 95% intervals remain unchanged. All eight new-bank allocation effects are adverse on both domains. No rate, model or continuation rule is selected using the new bank.

A later development extension adds only $1.2\times$HIGH to both original families at the two original development forks. The old candidates, STOP, 1,280-update suffix and mean-A eligibility/minimum-mean-B selection rule remain fixed; ties favor STOP, then the lower rate. Both families select the new upper boundary, leaving optimal rates unresolved. Table 10 reports all four new development cells. The four-allocation sensitivity cohort was fixed before the new development outcomes and reuses previously evaluated forks and banks. Both selected 1.2 procedures improve on their own 1.1 results; recent-half remains adverse in every allocation.

The B penalty is smaller at the extended rate; its A change remains unresolved (Table 11). These previously exposed allocations provide finite-rate sensitivity, not independent confirmation or additional sample size. New-document evaluation costs .4911 GPU-hours (1768.10 seconds), excluding training, and source preparation costs 27.64 CPU seconds. The four development suffixes cost 2.9473 GPU-hours and the eight sensitivity suffixes 5.8964, totaling 8.8437.

C Control results and reproducibility check

Table 3 reports the main higher-rate control; Table 12 preserves the matched-horizon extension. The earlier six-allocation HIGH-refresh comparison has B -24.95 micro-NLL, crossed 95% $[-52.41, 4.25]$, and A -6.39, $[-10.96, -2.33]$. This reused group is not pooled with the twelve-allocation confirmation. The latter's recipe contrast retains raw-array precision $+26.8222$, $[16.1810, 41.0433]$; rounding does not alter the analysis.

The prefix-preserving extension costs 9.093 incremental GPU-hours, excluding reused warm states. A separate selection-bank evaluation gives p25 late − early B $+27.36$ micro-NLL, $[15.21, 44.21]$,

Table 11: Rate sensitivity at four reused, exposed forks, micro-NLL. Extended denotes $1.2\times$HIGH; old denotes $1.1\times$HIGH. All ten coordinates use the separate crossed 99.5% family, with 8192 draws and seed 2026092601.

Contrast	A effect [interval]	B effect [interval]
Extended recent-half − native	25.01 [15.90,34.86]	281.10 [257.35,306.51]
Old recent-half − native	24.62 [15.72,34.19]	290.87 [266.82,316.37]
Native extended − old	-283.02 [-339.13, -241.80]	-580.55 [-681.57, -485.74]
Recent-half extended − old	-282.64 [-338.57, -241.56]	-590.31 [-691.62, -494.97]
Rate interaction	.39 [$-.60$, 1.45]	-9.77 [-13.75, -5.31]

Table 12: Same-mixture horizon control: HIGH $V_A - R$ on four reused allocations, micro-NLL and descriptive crossed 95% intervals. The final row is a paired effect difference.

Fixed p50 tape	A effect [interval]	B effect [interval]
H512	-5.12 [$-14.50, 2.59$]	-27.80 [-40.10, -13.77]
H1792	-4.35 [$-12.59, 3.41$]	8.26 [$-19.68, 64.72$]
H1792 − H512	.77 [$-2.27, 4.00$]	36.06 [$-.19, 93.00$]

and H512 p25 − p50 B -2.42, $[-11.13, 11.31]$. Those descriptive role-separated contrasts cannot be combined with final-evaluation values to construct an unexecuted factorial. Neither the unresolved horizon contrast nor the smaller long-run gain supports a universal persistence-of-benefit claim.

The replay anchor preserves paired endpoint weights and the complete per-document H512 equality checks, not merely aggregate means. The observed numerical equality is limited to the tested implementation and does not establish portability to other kernels. Deleted earlier checkpoints remain unavailable. The four-history late endpoint is a descriptive extension, not an independent transport claim.

A separate four-allocation source-refresh comparison increased B's share from p25 to p75. The B-domain $V_B - V_A$ penalty narrowed at H2048, contrary to the predicted positive mixture interaction; the interaction magnitude also decreased from H512 to H2048. These contrasts used different future tapes from Table 12 and reused previously examined evaluation banks. Neither attenuation nor an unresolved interval establishes equivalence. No result alters the fixed β_2 recurrence; its relationship to loss involves gradients, preconditioning and path-dependent weights.

D MECHANISTIC DIAGNOSTICS AND FUNCTIONAL READOUT

Table 13 preserves the source-refresh factorial. Magnitude-only QP improves B relative to PP but loses to the fixed HIGH control by $+412.31$ micro-NLL, $[351.07, 470.00]$. Explaining an improvement against PP therefore does not establish a competitive policy. The separate native-age graft in Table 14 uses unedited p25/H1792 forks and the selected $1.1\times$HIGH suffix rate. Both shadows receive the live branch's clipped gradient and updated momentum; decay remains separate. Eight histories and 1024 shared windows/domain are jointly resampled in 8192 draws with a ten-contrast 99.5% family.

Table 13: Source-refresh graft at p50/H512: six reused allocations, micro-NLL and descriptive crossed 95% intervals. Complete evolving rules are compared, not pathwise mediation components.

Rule contrast	A effect [interval]	B effect [interval]
QP − PP	-10.43 [-17.36, -2.83]	-66.39 [-103.93, -21.30]
PQ − PP	2.74 [$-2.08, 7.02$]	28.35 [$-2.82, 55.62$]
QQ − PQ	-10.94 [-17.54, -3.34]	-61.33 [-96.90, -19.53]
QQ − QP	2.24 [$-3.10, 6.97$]	33.41 [8.98,56.01]
QQ − QP − PQ + PP	$-.51$ [-2.30, .32]	5.06 [$-.31, 17.53$]

Table 14: Native-age graft: eight reused native forks, micro-NLL and family-adjusted 99.5% intervals. Negative values favor the first rule. P is native; Q emphasizes recent statistics.

Rule contrast	A effect [interval]	B effect [interval]
PQ − PP	-22.85 [$-28.82, -17.04$]	203.30 [182.13,225.40]
QP − PP	49.69 [42.84,58.70]	103.48 [88.14,120.04]
QQ − QP	-22.85 [$-28.97, -16.91$]	203.70 [171.20,229.25]
QQ − PQ	49.70 [42.82,58.79]	103.88 [73.26,121.93]
QQ − QP − PQ + PP	.001 [$-2.00, 2.25$]	.40 [$-26.57, 15.53$]

Each of the sixteen off-diagonal continuation runs verified that one full-model PP update matched the native update, then restored the fork and RNG. PP and QQ endpoints are reused from their native trajectories; the original checks included small CPU PP/QQ comparisons. A later check at one native fork exactly matches a direct recent-half step with a full-model QQ step in weights, buffers, moments, clocks, options, modes and RNG. Both steps use the first suffix batch at $1.1\times$HIGH; restoring the fork discards both trial updates. This costs 187.03 GPU-seconds. Full 1,280-update QQ replay and cross-hardware agreement remain untested. Subsequent norm traces differ across branches and do not record angles. The off-diagonal continuation runs add 13.215 GPU-hours, excluding already-paid prefixes and diagonal endpoints. The unresolved interaction does not establish exact additivity.

The separate twelve-allocation graft retains the remaining simple effects as well: QQ − QP gives A -24.43, adjusted 99.5% $[-30.84, -19.78]$, and B $+195.59$, $[176.61, 215.27]$; QQ − PQ gives A $+49.35$, $[43.12, 57.47]$, and B $+102.28$, $[91.71, 113.40]$. Together with Table 4, these complete the prespecified ten A/B contrasts. The 24 off-diagonal continuations cost 19.767 GPU-hours, including a repeated computation for one interrupted allocation; interrupted attempts cost an additional .168 GPU-hours.

For the native first-suffix readout in Figure 4, the common gradient is accepted after clipping and moment/clock updates. The reported best scalar is nonnegative for all eight allocations, so no negative learning-rate implementation is implied. Zero-norm cases are explicitly handled in the implementation. The residual measures a vector property for that one gradient. Subsequent branches no longer share gradients, and no long-horizon approximation guarantee is asserted. A larger residual can coexist with a worse endpoint, as the native confirmation shows.

Fixed direction and future-allocation boundaries. Old-rich reweighting moves the initial old coefficient from .5951141661 to .8926712492, opposite to RECENT_HALF's .2975570831. Positive old/recent accumulators construct the moment directly; subtracting nearly equal tensors is unnecessary. PO uses the native-shadow magnitude and old-rich direction within each live branch. The equal coefficient displacement does not imply equal angles, equal gradients across branches, or a pure intervention on age rather than estimator concentration. The original p25 PO − PP effect favors B and harms A at all four forks.

Table 5 preserves both successful and adverse outcomes. Recent direction remains adverse on B under pure-B training; old-rich direction retains a smaller B gain, and both reference effects reverse relative to p25. The pure-B PO − PP effects are favorable on both domains at all four forks. This result supports a conditional fixed-rule benefit, without establishing broad learning-rate optimality, equal-time policy superiority or model-family transport. The PQ and PO follow-ups add 6.261 and 3.556 GPU-hours respectively, excluding reused controls and prefixes. Interpretation is limited to the fixed paired contrasts and their uncertainty intervals.

The full-vocabulary evaluation compares two prespecified late-cancellation endpoint pairs on common document prefixes. Stable log probabilities and FP64 accumulation cover the complete vocabulary; discarded top-k mass is not treated as zero. The evaluator is checked by comparing each checkpoint with itself before the cross-model comparison. Table 15 reports Jensen–Shannon divergence (JS, in nats), argmax disagreement and gold-token NLL change. Document-only uncertainty is retained in the machine-readable report, but no two-history population conclusion is drawn.

This inference-only experiment costs .1553 aggregate GPU-hours and creates no new adaptation or pretraining roots. Gold-token loss differences, full predictive divergence and downstream behavior

Table 15: Fixed endpoint evaluation. Gold loss is CANCEL − SHAM. Positive JS measures a distributional difference, not semantic harm or benefit.

Domain	History	Mean JS (nats)	Argmax disagreement	Gold change (micro-NLL)
A	0	9.52×10^{-7}	.125%	24.11
A	1	5.06×10^{-7}	.071%	4.85
B	0	1.34×10^{-5}	.185%	−727.22
B	1	9.19×10^{-6}	.194%	−731.17

Table 16: Complete pure-B development menu. Absolute mean NLL over the two fixed development parents; both families select $2.2\times$HIGH by B. A imposes no eligibility constraint.

Family	Choice	A NLL	B NLL
Both	STOP	2.67787147	1.42143103
PP	.55	2.68015881	1.39387730
PP	1.1	2.68099571	1.38866028
PP	2.2	2.68233423	1.38650263
PO	.55	2.68009859	1.39375856
PO	1.1	2.68090809	1.38852879
PO	2.2	2.68221448	1.38641098

remain distinct measurements. The HIGH-rate refresh document–allocation B-loss RMS of 162.448 micro-NLL and its 34.68% adverse pair fraction describe heterogeneous log-likelihood changes; those values are not substitutes for the full-distribution evaluation.

Symmetric selection under pure-target continuation. The later pure-B comparison gives PP and PO the same STOP/$.55, 1.1, 2.2$ times HIGH menu on the original two development parents. Each family minimizes mean B; ties favor STOP and then the lower rate. A remains reported without an eligibility constraint, unlike the earlier p25 selector. The original prefix and 5,120 B suffix windows are retained while 15,360 further FineMath windows replace the A slots. Recorded-role and content exclusions and nonwrapping cursors are preserved. Both families select the upper boundary 2.2, not a demonstrated optimum (Table 16).

The two fixed procedures then use the same four exposed forks and evaluation bank as the fixed-rate pure-B study. Both rates change, so all eight selected continuations are trained; four shared STOP evaluations provide absolute references. A/B paired contrasts use 8192 crossed draws with seeds 26092661/26092662 and a separate two-contrast 97.5% family. No new allocation is added and the original fixed-rate interactions retain their own estimates and families.

All four selected PO − PP effects favor PO: B effects are $-74.63, -38.83, -56.66, -67.04$ micro-NLL and A effects $-125.64, -108.75, -126.10, -108.18$. Table 17 reports the pooled effects and absolute losses. The target increment is .0000592915 NLL against PP's .03821442 reduction from STOP; the descriptive .15515% ratio has no ratio inference. Both selected procedures increase A loss relative to STOP. Relative gains therefore do not establish retention safety or equal-time policy superiority. Training and evaluation take 16.3590 GPU-hours; including full allocated-GPU elapsed time gives 16.4558 hours, neither a CUDA-active measure. Shared warm/prefix costs are excluded. The selected endpoint weights are unavailable for reanalysis.

Matching coefficient concentration while changing temporal emphasis. At the same four old-rich-reweighting forks, exact 512-update prefix replay reconstructs the fork state. Let O be the bias-corrected warm second moment and E, L the mean clipped squared-gradient vectors over the first and last 256 prefix updates. With $a = .5951141661$, replacements are $\widehat{v}_E = aO + (1-a)(.75E + .25L)$ and $\widehat{v}_L = aO + (1-a)(.25E + .75L)$. Multiplication by $1 - .999^{4608}$ restores the raw-moment clock. Positive construction changes only the second moment. Old mass and the coefficient multiset, sum and squared sum match exactly. Each prefix half contains 3072 A and 1024 B sequences; equal weighted input exposure does not separate domain contributions to squared mixed gradients.

Table 17: Selected pure-B procedures on four reused, exposed forks. Upper rows are absolute NLL; the final row gives PO − PP effects and adjusted 97.5% intervals in micro-NLL.

Procedure	A NLL	B NLL
Shared fork STOP	2.68494948	1.44346393
PP, 2.2×HIGH	2.68960909	1.40524952
PO, 2.2×HIGH	2.68949192	1.40519022
Contrast	A effect [interval]	B effect [interval]
Selected PO − PP	-117.17 [$-131.18, -103.20$]	-59.29 [$-86.41, -35.83$]

Table 18: Matched coefficient concentration at four reused forks, micro-NLL; positive values are adverse. EARLY − LATE uses the primary two-contrast 97.5% family. EARLY/LATE − PP use a separate four-contrast 98.75% family.

Contrast	A effect [interval]	B effect [interval]
EARLY − LATE	3.98 [$-1.29, 8.86$]	155.86 [66.22,329.07]
EARLY − PP	3.36 [$-.66, 6.74$]	124.18 [59.03,323.47]
LATE − PP	$-.62$ [$-2.80, 2.82$]	-31.68 [$-46.51, 10.69$]

Both replacements use the native-shadow magnitude on their own live branch and the same 1,280-update p25 tape at $1.1\times$HIGH. The branches do not share subsequent gradients or a common cross-branch norm; uniform within-half weights also differ from the original exponential weighting. Exact one-step diagonal checks and fork restoration precede each continuation. Previously exposed PP controls and 1024 shared documents per domain are reused. Crossed bootstrap uses 8192 draws, seeds 26092631/26092632 for A/B, with 97.5% primary and 98.75% secondary intervals (Table 18).

The four EARLY − LATE B effects are 71.16, 328.96, 117.26 and 106.04 micro-NLL, all adverse. Coefficient summaries alone therefore do not determine this loss ordering. Actual correlated-gradient variance is not matched and content remains associated with temporal position; neither a pure age effect nor the fraction of the original recent-half penalty is identified. LATE − PP is unresolved under its specified adjustment despite a negative descriptive 95% interval. The A contrasts do not establish equivalence. Numerical reconstruction of paired arrays and intervals is separate from independent neural replication; technical interruptions and their repairs add no scientific observations.

Cross-batch numerical check. A training-only check compared physical batch 16 with accumulation over four microbatches of four on the same first sixteen WikiText-103 windows, initial Qwen weights, empty AdamW state and RNG. Both paths used FP32/eager attention, no TF32 or autocast, 8,192 shifted targets, one clip and one AdamW update at 10^{-5}. Table 19 preserves the original criteria and observations. A separate trial of the unchanged microbatch-four kernel exactly reproduced the instrumented microbatch-four weights, moments and clock. Current paired experiments condition on that fixed implementation; the failed cross-batch check is not reversed by same-path reproduction.

The largest updated-parameter discrepancy occurs in the last layer's key-projection tensor; 225 of 338 tensors exceed the threshold at some coordinate, not 225 failing coordinates. Available summaries do not identify the responsible operation or establish a gradient sign reversal. The check neither supplies a universal error bound for small effects nor automatically invalidates every paired comparison on the fixed microbatch-four path.

E Artifacts, costs and bounded claims

The analyses include the twelve-allocation native graft, four-allocation old-rich intervention, WikiText-103 transport and both pure-B direction comparisons, alongside the original cancellation and refresh studies. All four fixed forks in the old-rich pure-B comparison enter Table 5. Appendices A and B specify the training and resampling procedures. Reported effect estimates exclude incomplete computations; failed computations remain relevant to measured cost.

Table 19: Original numerical check on training data. Relative L_2 checks compare global vectors; the maximum-coordinate check concerns updated parameters. Thresholds are implementation criteria, not margins for scientific effects.

Comparison	Discrepancy	Threshold	Outcome
Native labels versus shifted NLL	2.3842×10^{-7}	10^{-4}	Within
Full batch versus microbatch NLL	2.9802×10^{-7}	10^{-4}	Within
Raw gradient, relative L_2	4.7790×10^{-4}	5×10^{-4}	Within
Clipped gradient, relative L_2	4.7222×10^{-4}	5×10^{-4}	Within
Actual update, relative L_2	.00403483	.05	Within
Updated parameter, L_∞	1.9830×10^{-5}	2×10^{-6}	Exceeds
Same-path weights / Adam state	$0/0$	Exact	Matches

The native experiment includes eight confirmation prefixes, twelve development suffixes, sixteen selected confirmation suffixes and their own checks. Two checkpoint-writing failures used 2603.00 GPU-seconds, included in the total below. The native study's incremental total is 153430.62 seconds; including failed attempts gives 156033.62 seconds. These are measured experiment costs, not optimized deployment costs. Table 20 separates the retained reproducibility capabilities.

Table 20: Different retained artifacts support different reproducibility claims. A digest identifies bytes but cannot reconstruct a deleted tensor.

Evidence group	Retained capability and limitation
Late cancellation	Twelve complete warm states and 24 final weight endpoints; exact first-step fork checks and later inference evaluation
HIGH-rate refresh	Frozen recipe and paired loss arrays; original comparison did not save new endpoints; a bounded complete replay now supplies an anchor
Native confirmation	Retained native-prefix checkpoints, selected endpoint pairs, selection procedure, accepted-gradient diagnostic and final paired arrays
Earlier checkpoints	The 1508 earlier checkpoints were deleted; remaining records cannot reconstruct their model states
Larger native/graft studies	Twelve-allocation primary and graft comparisons; interrupted attempts are documented separately. No pooled $n = 20$ claim
Direction and future contrasts	Four shared forks, p25 and pure-B PP/PQ/PO; common previously examined final-evaluation bank and positive old/recent construction
WT103 transport	Actual target prefix/suffix, PP-only finite development selection, four paired confirmation allocations; no model-family or old-rich reweighting transport

An ordinary policy may have access to different information than a shadow-based edit. The late-cancellation shadow must have been maintained or reconstructed; a generic checkpoint need not contain it. A suffix comparison may treat an already-paid prefix as sunk, but an end-to-end implementation must include the prefix, shadow maintenance, restoration, observation and evaluation costs. Conversely, charging a common warm state twice exaggerates a shared experiment's cost. Cost attribution therefore depends on the decision point and which work is shared.

The evaluation does not assess practical speedup, a downstream utility threshold, privacy leakage, semantic history decoding, population safety or a minimum checkpoint payload. Positive higher-rate refresh effects and adverse native effects both remain part of the same explanation. No favorable subset replaces a prespecified paired comparison. A technical failure is not a null effect; an interval containing zero is not proof of equivalence; a small effect with a stable sign is not automatically scientifically important.

F ORDINARY CONTINUATION WITH REINVESTED TRAINING TIME

This follow-up asks whether the small pure-B OLD_PLUS benefit survives a stronger ordinary-continuation control. PP retains the native state; PO uses the existing OLD_PLUS intervention. Both families selected 2.2×HIGH on the same two development allocations. Their accepted-update training times determine the ordinary budget before the additional continuations:

$$H_{\mathrm{PP}} = \left\lfloor 1280 \frac{\overline{T}_{\mathrm{PO,dev}}}{\overline{T}_{\mathrm{PP,dev}}} \right\rfloor = 1421.$$

The learning rate is $(10^{-6} \times 1.03) \times 2.2$, with weight decay .01. The four already evaluated allocations reproduce the original 1,280-step PP traces and readouts before continuing for 141 updates. Each receives 2,256 further FineMath windows, disjoint from the available training, development and evaluation records. This does not establish non-exposure during pretraining. The comparison uses the previous PO endpoints and no further rate or stopping-point selection.

Table 21: Pure-B continuation on four exposed allocations. A is C4 and B is FineMath; lower NLL is better. STOP is the reached starting checkpoint.

Method	Updates	A NLL	B NLL
STOP	0	2.684949484	1.443463933
PP	1280	2.689609086	1.405249516
PO	1280	2.689491920	1.405190224
PP, reinvested	1421	2.689805465	1.404204199

Table 22: $\mathrm{PO}_{1280}-\mathrm{PP}_{1421}$ in micro-NLL. The two-domain family uses 97.5% intervals from 8,192 crossed allocation/document resamples, with 1,024 paired evaluation documents per domain. Positive values favor additional ordinary continuation.

Domain	Mean difference	Adjusted interval
A	-313.544602	$[-572.658523, +30.857017]$
B	$+986.025358$	$[+611.342509, +1431.696286]$

The B differences in the four allocations are 888.61, 865.59, 1365.28 and 824.63 micro-NLL, all favoring PP_{1421}. The A differences are -444.53, 119.49, -428.77 and -500.37. At the original equal 1,280 updates, PO's B difference was only -59.29 micro-NLL. Additional ordinary updates therefore remove that target advantage. The A interval includes zero: neither Pareto dominance nor a resolved two-domain trade-off follows. All three continuation arms have higher A loss than STOP.

$\mathrm{PP}_{1421}-\mathrm{PP}_{1280}$ is -1045.32 micro-NLL on B, with a separate descriptive 95% interval $[-1396.06, -712.66]$, and 196.38 on A, $[-94.35, 429.48]$. These descriptive intervals do not replace the two primary 97.5% intervals.

Table 23: Measured seconds per allocation. The step allowance uses prior development training time; these observed times are not an exact runtime match or a controlled speed benchmark.

Timing scope	PP_{1421}	PO_{1280}	PP / PO
Accepted-update training	2712.620	2714.468	.999319
Worker execution	3015.774	3167.840	.951997
GPU allocation elapsed time	3035.995	3182.664	.953916

The ordinary arm consumes more target examples and tokens. Source preparation, evaluation and process costs differ, and runs occurred at different times. Consequently this design does not establish equal FLOPs, equal end-to-end cost or a serving speedup. The four allocations are reused experimental histories, not independent pretrained models or an untouched confirmation cohort. This negative practical comparison separates the causal effect of optimizer memory from a reason to spend additional computation on state intervention.