NEWER IS NOT ALWAYS BETTER: THE JOINT ROLE OF ADAMW MOMENT HISTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlled AdamW moment transplants across model families from 34M to 7B parameters reveal that a newer first moment paired with an older second moment can be far more disruptive than an older, coherent moment pair. We explain this asymmetry through *history coverage*: when the second moment covers the gradients represented in the first, a weighted Cauchy–Schwarz bound controls the normalized first moment. Equal-radius interventions in a 70M output head show that update magnitude alone cannot explain the damage; extended continuations reveal effects that outlast inherited moment differences. These findings motivate REBUILD_M: retain the second moment and reaccumulate the first from zero. In a common-gradient recovery model, this rule is minimax-optimal for cumulative absolute adaptive-update error relative to full-state recovery on a coverage-defined source class, given the retained second moment. For natural continuation, an exact reset representation links history coverage to a local validation-loss perturbation bound. Under uniform local regularity and a sufficiently small learning-rate scale η, complete first-moment deletion induces $O(\eta)$ loss perturbations over $O(1/\eta)$ updates, with an $O(\eta^2)$ remainder in the signed response expansion. In 1B/7B recovery with the original batches, schedules, and data cursors, the policy saves one third of checkpoint payload with validation loss close to full-state recovery, including repeated 1B recovery. Constructing a compatible second moment repairs much of the damage from zeroing it, while retaining its actual history improves recovery further at the same payload.

1 INTRODUCTION

Training recovery restores more than model weights. AdamW combines adaptive gradient updates with decoupled weight decay (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Its first moment m accumulates gradients to supply the adaptive update numerator, while its second moment v accumulates squared gradients to scale it coordinatewise. Together, these histories shape the update and occupy checkpoint storage alongside model weights.

Figure 1 illustrates the problem. At fixed current weights, optimizer clock, and future batches, a 34M model resumes with 16-step-old moments at mean excess validation negative log-likelihood (NLL; nats/token, lower is better) 0.0053. Replacing only m with its current value raises this response to 2.54 (averaged over eight stages). The newer component is not intrinsically bad: it is damaging in combination with the older denominator. This raises two linked questions: *how should moment histories be paired for continued training, and which history should a checkpoint preserve?*

Prior work establishes that optimizer history affects continuation and that resetting it can change training (Sevetlidis & Pavlidis, 2026; Asadi et al., 2023; Guo, 2026). Checkpoint systems address how training state is saved and restored (Maurya et al., 2024; Wan et al., 2025). History coverage links the two questions: it explains the recurring source-order asymmetry and motivates decay-sensitivity tests. Equal-radius interventions and extended continuations reveal how pairing-induced directions affect loss and displace training. This structure guides saving v and reaccumulating m from zero. We analyze the information lost by this choice and how its reset input propagates into NLL, then test the policy under native 1B/7B training conditions. Selecting which history to retain complements how retained states are encoded, a distinction we test using matched low-precision storage.

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.

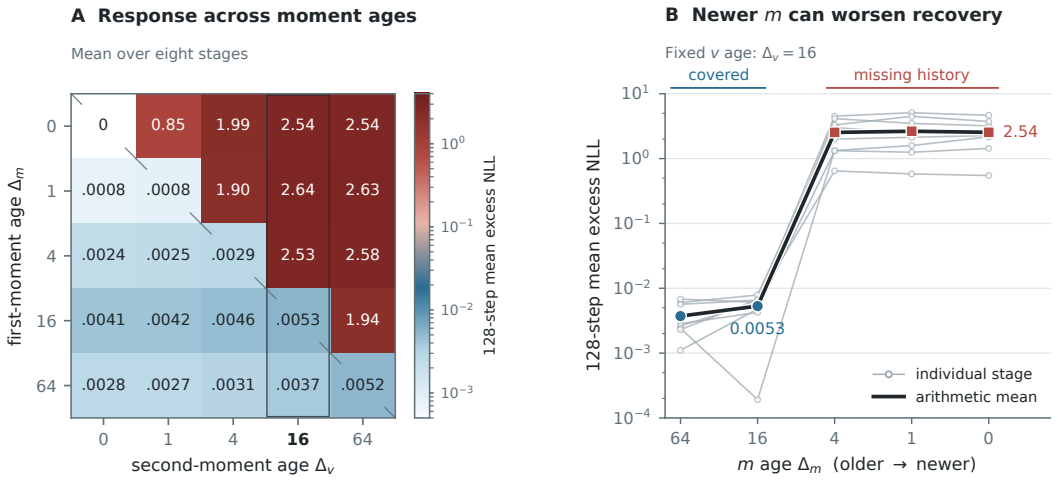

Figure 1: **Freshness alone does not determine recovery quality.** Left: the 34M atlas averages eight stages of one training trajectory; high responses concentrate where m is newer than v. Right: the highlighted $\Delta_v = 16$ column is expanded into individual stages (gray) and their arithmetic mean. Refreshing m from age 16 to 0 raises mean excess NLL from 0.0053 to 2.54. Each age is a separate intervention, not recovery time; weights, clock, schedule, randomness, future batches, and evaluator are fixed. Both panels use log displays; the zero reference cell is uncolored.

Our main contributions are:

- **A stable empirical structure.** Source pairing, rather than component freshness alone, organizes recovery responses from small models to billion-parameter models across families (Section 3).
- **An explanation with testable predictions.** History coverage explains the source-order asymmetry and motivates predictions of covered-side sensitivity to q; equal-radius interventions isolate directional effects, and dynamics explain their persistence (Sections 4–5).
- **Coverage-guided recovery.** Tight information bounds identify the value of retained v; a compensated-state theorem connects complete m deletion to a local loss-perturbation bound and signed response. Original-condition 1B/7B and repeated-recovery comparisons establish one-third payload savings and the value of retained history; matched encoding tests show composition with low-precision storage (Section 6).

2 Controlled Moment-Source Interventions

At target update t, we transplant moments from the same parent trajectory:

$$z_t(\Delta_m, \Delta_v) = (\theta_t, m_{t-\Delta_m}, v_{t-\Delta_v}, t). \tag{1}$$

Weights, optimizer clock, scheduler, RNG, future batches, and evaluation are held fixed. FULL uses $(0, 0)$. A coherent pair (d, d) restores equally old moments at *current* weights; it is distinct from replaying an old full checkpoint. The second-moment history covers the first when $\Delta_m \geq \Delta_v$.

For environment e (model, data order, and stage), the main profile statistic is

$$Y_e(a, b) = \overline{\Delta\mathrm{NLL}}_{0:128} = \frac{1}{128} \int_0^{128} [L_{(a,b)}(h) - L_{\mathrm{FULL}}(h)]\, dh, \tag{2}$$

computed by trapezoidal quadrature on a fixed evaluation grid. Positive values mean worse quality; negative values retain improvements. We evaluate a 5×5 age grid $\{0, 1, 4, 16, 64\}^2$ and summarize pairing with $B_d = Y(0, d) - Y(d, d)$ and

$$I_e(a, b) = Y(a, b) - Y(a, 0) - Y(0, b) + Y(0, 0), \tag{3}$$

$$P_d = Y(0, d) + Y(d, 0) - Y(d, d) - Y(0, 0) = -I(d, d). \tag{4}$$

B_d measures the benefit of using the old coherent numerator; P_d measures the gain beyond adding the two single-component effects.

The scale comparison uses Pythia-1B/6.9B (Biderman et al., 2023) at two matched data orders and stages 512/2048; Mistral-7B (Jiang et al., 2023) adds order 1 at step 512 and order 2 at steps 512/2048. Local moment histories are built from pretrained weights on FineWeb (Penedo et al., 2024). Pythia shares token arrays across sizes; Mistral uses the same documents with its native tokenizer. Exact recipes and metric definitions are in Appendices A.1 and B.2.

3 A Stable Source-Pairing Structure across Models

The small-model atlas suggests a directional pattern: responses are large when m contains gradient history absent from v, and much smaller in the reverse source order. Held-out 34M and 70M continuations, including an untied 70M model, reproduce this pattern. Across four additional 70M trajectories with independent initializations and data orders, both B_d and P_d remain positive at each of two checkpoints and two source ages (Appendix B.1). Figure 2 shows the same organization at larger scales and across families.

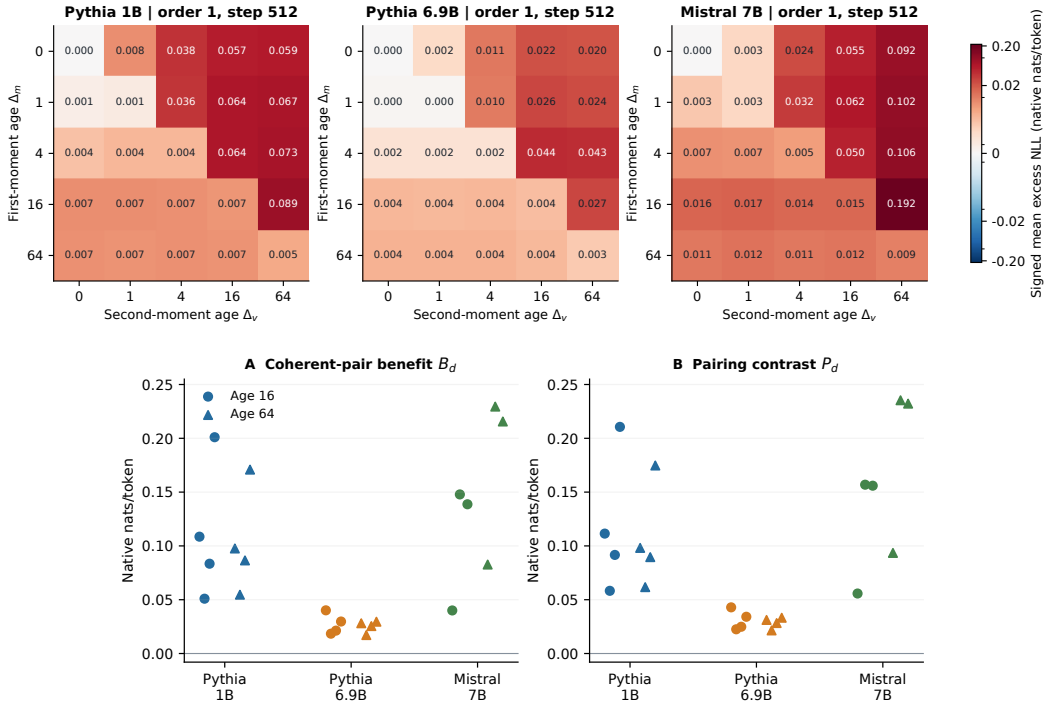

Figure 2: **The source-pairing structure recurs across models and conditions.** Top: each heatmap fixes order 1, step 512 and varies the moment-source ages. The common signed symmetric-log scale (linear threshold 0.005) preserves native excess NLL. The upper triangle has newer m than v; lower-triangle histories are covered. Bottom: coherent-pair benefit B_d (A) and pairing contrast P_d (B) across two Pythia data orders, two stages, and two ages (16 comparisons), plus six Mistral comparisons across three environments. Circles/triangles denote ages 16/64; colors denote models. Both measures remain positive across these conditions. All heatmaps and action details appear in Appendix B.2 and Figure 8.

The pattern persists when data order, stage, and source age change: B_d and P_d are positive in all 16 Pythia environment–age comparisons and all six Mistral comparisons at $d \in \{16, 64\}$ (Figure 2, bottom). Global coherent pairing also yields lower response than either output-head-only reset or rollback, which in turn improve on the raw new-m/old-v state:

$$Y_{\text{coherent}} < \{Y_{\text{head reset}}, Y_{\text{head rollback}}\} < Y_{\text{raw}}. \tag{5}$$

The full action comparison is in Figure 8.

Across the four matched Pythia scale pairs, full-profile Spearman correlations are 0.795–0.958; non-trivial 4×4 interaction correlations are 0.868–0.935. Mistral reproduces the pairing structure in a

second data order at both stages, with $B_d = 0.139$–0.229 across its four new comparisons. Its original order-1, step-512 profile correlates 0.908 with Pythia-6.9B and 0.892 with Pythia-1B. Response magnitudes differ, and lower-triangle details are less stable than the prominent asymmetry (Table 9). Joint moment sources organize the recurring structure across these environments.

4 History Coverage and Sensitivity to Moment Decay

A bound from history inclusion. For zero-initialized, uncorrected scalar moments with $0 < \beta_1, \beta_2 < 1$, $m_s = (1-\beta_1)\sum_{k=1}^{s} \beta_1^{s-k} g_k$ and $v_r = (1-\beta_2)\sum_{k=1}^{r} \beta_2^{r-k} g_k^2$. When $s \le r$, every gradient in the numerator is represented in the denominator. Weighted Cauchy–Schwarz gives the coordinatewise bound

$$m_s^2 \le C_{s,r} v_r, \qquad C_{s,r} = \frac{(1-\beta_1)^2}{1-\beta_2}\,\beta_2^{-(r-s)} \sum_{j=0}^{s-1} q^j, \qquad q = \frac{\beta_1^2}{\beta_2}. \tag{6}$$

For a fixed source gap, the history sum is bounded for $q < 1$, grows linearly at $q = 1$, and grows geometrically for $q > 1$. The source-gap factor is separate. If $s > r$, missing gradients can enlarge m_s without entering v_r, so no amplitude-independent bound of this form exists. Current-clock bias correction multiplies $C_{s,r}$ by $(1-\beta_2^t)/(1-\beta_1^t)^2$; positive ϵ reduces the update ratio. Appendix C.1 gives the proof and an exact comparison with the same-m coherent companion.

A prospective covered-side test. The bound identifies the covered triangle; its growth motivates the prediction of increased sensitivity as q crosses one *within that region*. We test this on two parameter paths: changing β_2 at fixed β_1, and changing β_1 at fixed β_2. Raising q from 0.95 to 1.05 increases covered-side damage in all four seed–stage rows on each path, with median lifts 0.00423 and 0.00595 nats/token. An external untied Pythia-1B test confirms positive lifts at both stages (Figure 3).

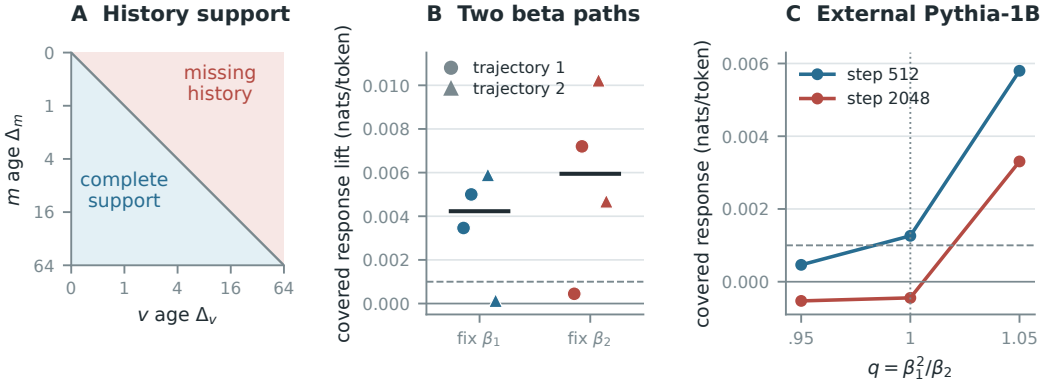

Figure 3: **A coverage-motivated prediction is supported in the covered region.** A: history inclusion maps to the lower triangle. B: the 14-cell covered-side median response increases from $q = 0.95$ to 1.05 on both parameter paths; each point is a seed–stage row. C: the four-cell covered median in an external untied Pythia-1B test shows the same direction at two stages. The q test concerns response strength inside the covered region, not movement of its source-order boundary.

The uncovered side remains substantially more damaging, but its response generally decreases along these q paths. Coverage permits amplification there; it does not order its magnitude. The realized update also depends on numerator load and its alignment with the denominator defect. Their exact factorization explains why a larger defect alone need not produce a larger update (Appendix C.2).

5 From Update Geometry to Persistent Training Response

Equal-sized displacements can have very different effects. To isolate direction from magnitude, we test output-head displacements on an untied 70M model at equal parameter-space radius, fixed

before validation. Uncovered-to-covered and uncovered-to-random absolute functional-response ratios have medians of $18.2\times$ and $160.4\times$ (20 and eight paired comparisons; Figure 4, top). The statistic averages $|\Delta L(h)|$ over time. Thus, source pairing induces directions with markedly different functional effects even at the same radius.

The update geometry shows how moment pairing shapes direction. Compare the mismatched state $F = (m_s, v_r)$ with its same-m coherent companion $C = (m_s, v_s)$, holding current weights, gradient, optimizer clock, learning rate, and ϵ fixed. Their common $\widehat{m}$ is the first moment after incorporating the current gradient and applying bias correction. The exact adaptive-update gap is

$$\|u_F - u_C\|_2^2 = \eta^2 \sum_i \widehat{m}_i^2 (d_{F,i}^{-1} - d_{C,i}^{-1})^2, \tag{7}$$

where $d = \sqrt{\widehat{v}} + \epsilon$ and decoupled weight decay cancels. Coordinatewise denominator changes reshape the update according to their alignment with the numerator. History mismatch therefore changes both update magnitude and direction.

Moment forgetting and functional recovery are different. Clamping the output-head endpoint to its coherent companion for one update leaves roughly 84% of the post-release response; clamping for 16 updates leaves 1.6%. Hidden optimizer state can therefore regenerate displacement. Under identical gradients, moment differences contract as $\delta m_K = \beta_1^K \delta m_0$ and $\delta v_K = \beta_2^K \delta v_0$. In closed-loop training, divergent gradients also create new moment differences and accumulate parameter displacement. Exact recurrences and reference-gradient replay isolate these contributions in Appendix C.5.

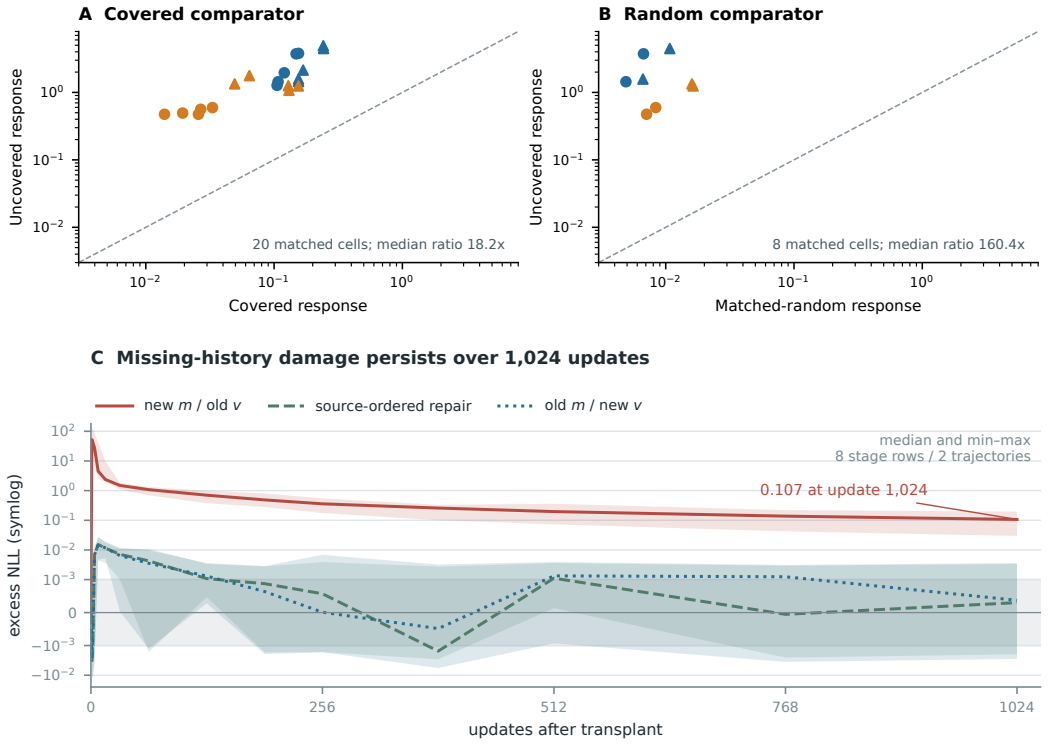

Figure 4: **The response depends on direction and persists over time.** Top: equal-radius output-head interventions compare uncovered, covered, and random directions (log axes); circles/triangles distinguish seeds and blue/orange distinguish stages 512/4096. Bottom: the age-16 mismatch is continued for 1,024 updates. Lines and bands show medians and min–max ranges over four stages on each of two 70M trajectories. The shared symmetric-log NLL axis is linear within the shaded ± 0.001 band; time is linear.

Extending the age-16 mismatch to 1,024 updates on two independent 70M trajectories, with four stages each, reveals a long-lived tail (Figure 4, bottom). Median endpoint excess NLL is 0.1068 (range 0.0297–0.1965); all 8/8 stage curves remain outside ± 0.001. Covered-order controls end near zero. Inherited history can fade while its earlier updates leave a displaced trajectory, making the initial recovery state consequential.

6 Coverage-Guided Recovery

6.1 From coverage to a recovery policy

Coverage links the source-pairing results to recovery design. Its same-source specialization supplies a constraint that new moment histories can preserve. For $q < 1$, define

$$C = \frac{(1-\beta_1)^2}{(1-\beta_2)(1-q)}, \qquad \mathcal{E}_C = \{(m, v) : v_i \geq 0,\ m_i^2 \leq C v_i\ \forall i\}. \tag{8}$$

This envelope is forward invariant under Adam's moment recurrence, including a recovered branch's own gradients (Appendix C.6). Setting $m = 0$ enters it for any valid v: new first-moment history is covered by the second moment, while retained v supplies additional support. This gives **REBUILD_M**: save parameters, v, and training metadata; zero m once on recovery, preserve the clock, and reaccumulate m.

We next ask which retained history limits information loss under common future gradients, then how deleting m affects NLL when each branch follows its own gradients.

Which history preserves recovery information? First, consider scalar Adam driven by common future gradients, without past-history replay. Let $\mathfrak{R}_H$ be the smallest worst-case cumulative absolute adaptive-update error relative to FULL among causal recovery rules. **Proposition 1 (state choice).** For a fixed clock $t \geq 2$, $q < 1$, constant $\eta > 0$, zero weight decay and $\epsilon = 0$, consider the same reachable source class $\mathcal{S}_{l,u} = \{(m, v) : l \leq v \leq u,\ m^2 \leq C_t v\}$, where $C_t = C(1 - q^t)$ and $0 < l < u < \infty$. For fixed $R = u/l > 1$,

$$\mathfrak{R}_H^{\text{keep } v} = \Theta(1), \qquad \mathfrak{R}_H^{\text{keep } m} = \Theta(\sqrt{H}). \tag{9}$$

REBUILD_M attains the first minimax risk exactly: retaining v makes the omitted numerator history geometrically summable, whereas missing v permits persistent denominator ambiguity. Appendix C.7 proves the tight joint H, R law. For $\epsilon > 0$, its lower bounds persist when $\epsilon \leq \sqrt{l\beta_2^H}$; fixed-scale growth is eventually cut off.

From omitted history to loss response. Now let each branch generate gradients from its own evolving parameters. Coverage bounds the omitted-history input; its propagation through training and alignment with the validation gradient shape the signed loss response (Section 5). Define $\widetilde{m}_k = m_k^R + \beta_1^k m_0$. This auxiliary moment starts at m_0 and follows the ordinary recurrence, expressing full m deletion as the input $A_k \beta_1^k m_0 / (\sqrt{v_k^R} + e_k)$ to an otherwise ordinary update. Its cumulative coordinatewise magnitude is at most $\overline{A} \kappa_0 \rho / (1 - \rho)$, with $\rho = \sqrt{q}$ and $\kappa_0 = \max_{i : v_{0,i} > 0} |m_{0,i}| / \sqrt{v_{0,i}} \leq \sqrt{C}$. Here A_k, e_k include learning rate and clock corrections (Appendix C.8). **Proposition 2 (local loss perturbation and response).** Fix the source, betas, and $T > 0$. Under the uniform local smoothness and fast–slow block bounds in Appendix C.8, let $\eta_k = \eta \ell_k$ with bounded nonnegative ℓ_k. The compensated map's first-order response $z_0 = 0$, $z_k = J_k z_{k-1} + b_k$, uses its Jacobian J_k and reset input b_k along FULL. With uniformly bounded validation gradient and curvature,

$$\sup_{h \leq T/\eta} |\Delta L_h| \leq C_1 \eta + C_2 \eta^2, \qquad \Delta L_h = \nabla L(\theta_h^F)^\top z_{h,\theta} + O(\eta^2), \tag{10}$$

for sufficiently small η, with constants independent of η, h. The first bound follows from the $O(\eta)$ parameter displacement induced by complete m deletion. The second characterizes the leading signed response through propagated reset inputs and validation sensitivity (Appendix C.8). Coverage thus bounds the reset input, while closed-loop dynamics characterize its local effect on validation loss relative to FULL.

The envelope also provides construction controls. If only m is saved, **V_FROM_M** constructs $\widetilde{v}_i = m_i^2/C$ on its boundary. Resetting both moments also enters the envelope. Comparing these states tests the value of retained history beyond satisfying the ratio constraint. Each construction acts once; subsequent updates incorporate the current gradient into both moments.

Repairing a mismatched pair. Clipping $|m_i|$ to $\sqrt{Cv_i}$ while retaining v reduces two Pythia-1B new-m/old-v 64-update mean excess NLL values from $19.41/24.32$ to $0.00744/0.00962$, leaving FULL and the old coherent pair unchanged (Appendix C.6).

6.2 Retained history, recovery quality, and storage

Direct recovery under native training conditions. We compare five policies at OLMo-2-0425-1B ingredient-1 stage-2 step 5k, and FULL/REBUILD_M at OLMo-2-7B stage-1 step 500k (Team OLMo et al., 2025; Allen Institute for AI, 2025). Both retain the original batch ($512/1024 \times 4096$ tokens for 1B/7B), learning-rate schedule, and training-data cursor, without an adaptation prefix or added warmup. Each branch undergoes a real process kill and disk-only recovery for 128 updates. FULL saves (θ, m, v); REBUILD_M saves (θ, v); REBUILD_V and V_FROM_M save (θ, m); REBUILD_MV saves θ. Omitted moments are zeroed once, except for the envelope construction in V_FROM_M.

Table 1: Native-condition recovery at update 128: primary NLL differences from FULL (10^{-3} nats/token; paired 95% intervals) and payloads (GB).

Policy	Payload	ΔNLL	95% interval
A: OLMo-2 1B, five policies / 5k source			
FULL	17.819	0	—
REBUILD_M	11.879	-0.300	$[-0.479, -0.098]$
REBUILD_MV	5.940	$+1.202$	$[+0.611, +2.206]$
REBUILD_V	11.879	$+5.679$	$[+5.172, +6.303]$
V_FROM_M	11.879	$+1.149$	$[+0.940, +1.382]$
B: OLMo-2 7B, FULL / REBUILD_M, 500k			
FULL	87.584	0	—
REBUILD_M	58.389	-0.369	$[-0.611, -0.138]$

REBUILD_M saves one third of the payload at both scales, with endpoint differences from FULL of -0.000300 (1B) and -0.000369 (7B) nats/token (Table 1). In the five-policy 1B comparison, constructing v from m removes 79.8% of the zero-v endpoint excess; preserving real v improves further at the same payload. These results distinguish constructing a compatible state from retaining useful history (Figure 5A; evaluation details in Appendix D.1).

Retaining history through repeated recovery. Continuing the native 1B FULL, REBUILD_M, and REBUILD_MV branches to 384 updates with faults at 0/128/256 isolates history choice at identical recovery positions. REBUILD_M stays close to FULL, with endpoint difference -0.000427 nats/token, versus $+0.001176$ when both moments are omitted. Across the three faults, its first-16-update mean excess NLL ranges from -0.001951 to -0.001574, versus $+0.002269$ to $+0.002414$ for REBUILD_MV (Figure 5B; paired intervals in Appendix D.2). Omitting both moments repeatedly reintroduces a larger transient, even when endpoint differences are small.

Recovery across source states and scales. A complementary 512-update study tests all five policies at six source states: 1B ingredients 1/2 at stage-2 steps 5k/10k and 7B at stage-1 steps 500k/800k. The 1B states receive a 256-update common FineWeb prefix; 7B starts directly from the sources. FULL-only diagnostics fix effective batches of 4/16 and a 0.1 learning-rate multiplier for 1B/7B. Recovery uses real process kills, disk-only loading, and no added warmup; evaluation uses FineWeb for 1B and eleven published OLMo held-out sources for 7B (Appendix D.3).

Here the primary late statistic weights excess NLL at updates 384/448/512 by $0.25/0.50/0.25$. FULL ends below its initial NLL at all six states. REBUILD_M stays within 10^{-4} nats/token of FULL on every late point estimate (Table 2), with payload reduced from 17.819 to 11.879 GB at 1B and from 87.584 to 58.389 GB at 7B.

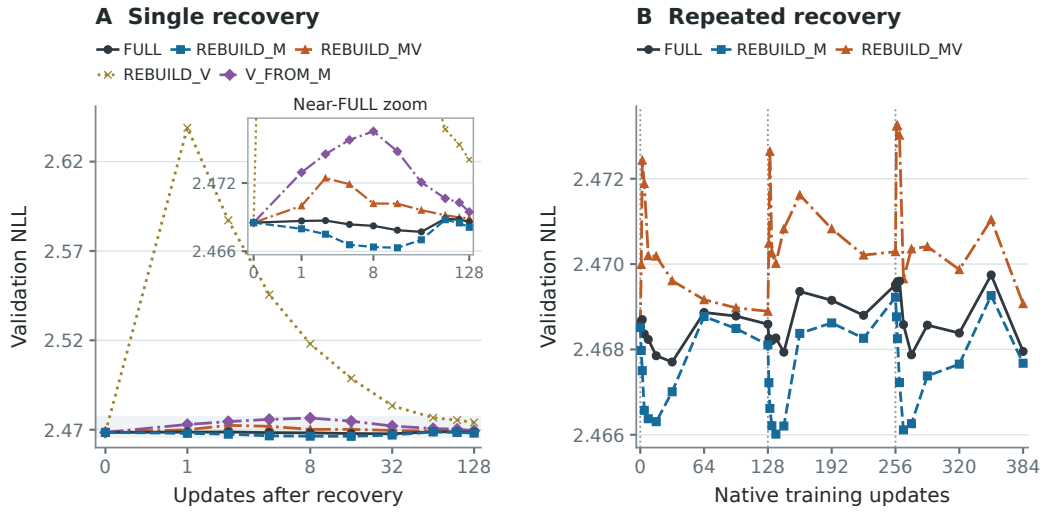

Figure 5: **Retaining v keeps native recovery close to FULL after single and repeated faults in OLMo-2 1B.** (A) Five-policy single recovery, with a near-FULL inset. (B) Three continuing branches; dotted lines mark real process kills and disk-only recoveries at updates 0/128/256. Both panels show absolute NLL on the same fixed 44 sequences, with separate linear vertical ranges. The horizontal axis is linear through update 1 and logarithmic thereafter in (A), and linear in (B). The first 128 updates in (B) reuse the corresponding trajectories in (A), not independent training runs. Evaluation details and 352-sequence primary estimates with paired intervals appear in Appendices D.1 and D.2.

Table 2: Late excess NLL relative to FULL under matched reduced-batch continuation settings (10^{-3} nats/token; lower is better). Each row is a source state. Retaining either one moment saves one third of the payload; removing both saves two thirds.

Source	REBUILD_M	REBUILD_MV	REBUILD_V	V_FROM_M
1B / I1 / 5k	+0.036	+0.246	+22.106	+0.180
1B / I2 / 5k	-0.013	+0.256	+37.503	+0.273
1B / I1 / 10k	-0.005	+0.174	+13.059	+0.111
1B / I2 / 10k	+0.095	+0.232	+14.927	+0.241
7B / 500k	+0.044	+0.133	+0.827	+0.101
7B / 800k	-0.087	+0.190	+0.478	+0.032

At 1B, zeroing v causes late excess NLL 0.01306–0.03750; envelope reconstruction reduces it to the 10^{-4} scale. Keeping actual v further improves late quality at equal payload (Figure 17; paired intervals in Appendix D.3.1). At 7B, envelope reconstruction removes $88\%/93\%$ of the zero-v excess. Removing both moments saves two thirds of the payload, with larger late deviations than REBUILD_M. Together these comparisons connect the state constraint to the practical choice of which history to retain.

Composition with storage precision. The recovery bound extends to approximate v, separating omission and relative-precision error (Appendix C.7.1). Across the four 1B states, omitting m saves a further 33.3% with BF16 storage and 25.2% with block-wise 8-bit moment quantization. Matched late NLL differences from full-state recovery at the same encoding lie within 6.0×10^{-5} nats/token, with training computation unchanged (Appendix D.4).

History versus checkpoint frequency. The Pythia-1B state comparison holds current parameters fixed: zeroing both moments gives seed-wise mean excess NLL $0.03044/0.02903$, versus $-0.00430/+0.00078$ when retaining v (Table 24). At matched planned write budgets, parameter-only saving (S5) checkpoints every 16 updates, compared with 32 for REBUILD_M (S3) and 48 for FULL. Despite this frequency advantage, S5 has larger recovery excursions (Figure 22). Reevaluating the repeated-fault terminals on 2,048 documents gives S5 excess NLL $0.06281/0.05040$, versus $0.00061/-0.00150$ for S3 (Appendix D.6). More frequent parameter snapshots do not, in these runs, replace retained second-moment history.

7 Related Work

Optimizer history and adaptive dynamics. Optimizer history affects continuation (Sevetlidis & Pavlidis, 2026; Asadi et al., 2023; Guo, 2026). SPAM combines momentum resets with spike-aware clipping (Huang et al., 2025). Adam analyses also connect stability to its decay parameters (Reddi et al., 2018; Nguyen, 2026). Bai et al. (2026) connect loss spikes to decoupling between v_t and g_t^2. Joint-state response analysis tracks minibatch-gradient shocks (Liu & Li, 2026); our exact compensation treats complete moment deletion and connects its input budget to history coverage. We vary moment sources to identify coverage and derive recovery rules. Worst-case optimal recovery provides a related information-based perspective (Foucart & Liao, 2024). Our analysis concerns information lost at restore; quantized-optimizer convergence concerns ongoing updates (Tang et al., 2026).

Checkpoint systems. Asynchronous persistence, distributed formats, and selective saving reduce checkpoint costs (Maurya et al., 2024; Wan et al., 2025; Lian et al., 2025; Sun et al., 2025). Concurrent LayerCheck restores layers from different timestamps while keeping each layer's weights and optimizer states together (Sun et al., 2026); we vary moment histories independently at fixed current weights. Lossy checkpointing studies the compression–recovery trade-off (Chen et al., 2020); ExCP combines checkpoint residuals, weight–momentum joint shrinking, and quantization (Li et al., 2024). Memory-efficient optimizers quantize states (Dettmers et al., 2022), factor second moments (Adafactor; Shazeer & Stern, 2018), or project gradients (GaLore; Zhao et al., 2024). Adafactor also links oversized updates to slow second-moment adaptation and proposes update clipping and a changing decay rate. We study moment-source pairing at recovery rather than ongoing optimizer design: our checkpoint policy resumes ordinary AdamW and composes with BF16 and 8-bit encoding.

8 Conclusion

AdamW recovery depends on how two moments jointly encode history. Coverage explains the source-order asymmetry; directional and dynamical experiments show how the resulting updates affect loss and persist in training. These findings guide retaining v and rebuilding m. Complementary information and response analyses quantify the history lost at recovery and connect its coverage-controlled input to local loss perturbations and signed response. Original-condition 1B/7B recovery, repeated 1B recovery, and matched continuations support this history choice, with near-FULL quality while saving one third of the payload. Constructing missing v through the envelope repairs much of the zero-v damage; retaining its history improves recovery further. Together with low-precision composition, these results make joint history a basis for selecting and reconstructing checkpoint state.

AI USE STATEMENT

OpenAI Codex and ChatGPT assisted with hypothesis and experimental-design discussions, mathematical formulation and proof drafting, implementation, result interpretation, literature search, figure preparation, writing, and translation. Numerical results come from executed experiments and retained evaluation records. The empirical findings are supported by matched comparisons, state-consistency checks, and recomputation of reported statistics from saved measurements, as documented in the paper. The authors take responsibility for the mathematical arguments, empirical claims, and final manuscript.

REPRODUCIBILITY STATEMENT

Appendix A details experimental configurations, paired evaluation, and outcome definitions. Appendices B, C, and D document source-state interventions, derivations, and recovery protocols, including checkpoint contents, fault schedules, and paired estimates.

REFERENCES

Allen Institute for AI. OLMo-2-1124-7B: Model card and intermediate checkpoints. Hugging Face model repository, 2024. URL https://huggingface.co/allenai/OLMo-2-1124-7B.

Allen Institute for AI. OLMo-2-0425-1B: Model card and intermediate checkpoints. Hugging Face model repository, 2025. URL https://huggingface.co/allenai/OLMo-2-0425-1B.

Kavosh Asadi, Rasool Fakoor, and Shoham Sabach. Resetting the optimizer in deep RL: An empirical study. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3164. URL https://papers.nips.cc/paper_files/paper/2023/hash/e4bf5c3245fd92a4554a16af9803b757-Abstract-Conference.html.

Zhiwei Bai, Zhangchen Zhou, Jiajie Zhao, Xiaolong Li, Zhiyu Li, Feiyu Xiong, Hongkang Yang, Yaoyu Zhang, and Zhi-Qin John Xu. Adaptive preconditioners trigger loss spikes in Adam. *arXiv preprint arXiv:2506.04805*, 2026. URL https://arxiv.org/abs/2506.04805. Accepted to ICML 2026.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Yu Chen, Zhenming Liu, Bin Ren, and Xin Jin. On efficient constructions of checkpoints. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1627–1636. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20m.html.

Liliaokeawawa Cothren, Francesco Bullo, and Emiliano Dall'Anese. Online feedback optimization and singular perturbation via contraction theory. *SIAM Journal on Control and Optimization*, 64(3):1482–1506, 2026. doi: 10.1137/24M1684736. URL https://doi.org/10.1137/24M1684736.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.02861.

Simon Foucart and Chunyang Liao. Radius of information for two intersected centered hyperellipsoids and implications in optimal recovery from inaccurate data. *Journal of Complexity*, 83: 101841, 2024. doi: 10.1016/j.jco.2024.101841. URL https://doi.org/10.1016/j.jco.2024.101841.

Jinhui Guo. Delayed optimizer-state transport shapes short-horizon training decisions. *arXiv preprint arXiv:2608.24593*, 2026. URL https://arxiv.org/abs/2608.24593.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7a70ad3d9c704fb9b81b5c69eda722dc-Abstract-Conference.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Wenshuo Li, Xinghao Chen, Han Shu, Yehui Tang, and Yunhe Wang. ExCP: Extreme LLM checkpoint compression via weight-momentum joint shrinking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27575–27588. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24m.html.

Xinyu Lian, Sam Ade Jacobs, Lev Kurilenko, Masahiro Tanaka, Stas Bekman, Olatunji Ruwase, and Minjia Zhang. Universal checkpointing: A flexible and efficient distributed checkpointing system for large-scale DNN training with reconfigurable parallelism. In *2025 USENIX Annual Technical Conference (USENIX ATC 25)*, pp. 1519–1534. USENIX Association, 2025. URL https://www.usenix.org/conference/atc25/presentation/lian.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Avinash Maurya, Robert Underwood, M. Mustafa Rafique, Franck Cappello, and Bogdan Nicolae. DataStates-LLM: Lazy asynchronous checkpointing for large language models. In *Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing*, 2024. doi: 10.1145/3625549.3658685. URL https://arxiv.org/abs/2406.10707.

Mistral AI. Mistral-7B-v0.3: Model card. Hugging Face model repository, 2024. URL https://huggingface.co/mistralai/Mistral-7B-v0.3.

Quan M. Nguyen. How to set β_1, β_2 in Adam: An online learning perspective. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 2026. URL https://proceedings.mlr.press/v313/nguyen26b.html.

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 Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Vasileios Sevetlidis and George Pavlidis. Training memory in deep neural networks: Mechanisms, evidence, and measurement gaps. *arXiv preprint arXiv:2601.21624*, 2026. URL https://arxiv.org/abs/2601.21624.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604. PMLR, 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

Minqiu Sun, Xin Huang, Luanzheng Guo, Nathan R. Tallent, Kento Sato, and Dong Dai. LLMTailor: A layer-wise tailoring tool for efficient checkpointing of large language models. In *Proceedings of the SC '25 Workshops of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pp. 1366–1374. ACM, 2025. doi: 10.1145/3731599.3767515. URL https://doi.org/10.1145/3731599.3767515.

Minqiu Sun, Xin Huang, Luanzheng Guo, Nathan R. Tallent, Kento Sato, and Dong Dai. LayerCheck: Adaptive layer-wise checkpointing for large language model post-training. *arXiv preprint arXiv:2609.27193*, 2026. URL https://arxiv.org/abs/2609.27193.

Xuan Tang, Jichu Li, and Difan Zou. A convergence analysis of adaptive optimizers under floating-point quantization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/07ea874e9e4f71ec6680a3574a485a36-Abstract-Conference.html.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025. URL https://arxiv.org/abs/2501.00656.

Joris van der Hoeven. Certifying trajectories of dynamical systems, 2018. URL https://www.texmacs.org/joris/lagrange/lagrange-v2018.pdf. Revised manuscript.

Borui Wan, Mingji Han, Yiyao Sheng, Yanghua Peng, Haibin Lin, Mofan Zhang, Zhichao Lai, Menghan Yu, Junda Zhang, Zuquan Song, Xin Liu, and Chuan Wu. ByteCheckpoint: A unified checkpointing system for large foundation model development. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, pp. 559–578. USENIX Association, 2025. URL https://www.usenix.org/conference/nsdi25/presentation/wan-borui.

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3a6935d11910d6f9142b0a1e36fc6753-Abstract-Conference.html.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

A Experiments and Measurement

A.1 Experimental Overview and Shared Controls

Experimental overview. The experiments connect source-pairing structure, its mechanism, and recovery design (Table 3). We evaluate each training trajectory at multiple checkpoints, applying the same set of moment-source interventions within each checkpoint. Profile correlations summarize each response matrix; recovery confidence intervals resample paired evaluation documents or sequences, as specified with the corresponding results.

Table 3: Experimental overview: each group addresses a part of the source-pairing, mechanism, and recovery argument.

Question	Models and experimental settings
Source-pairing structure	GPT-2-style 34M discovery: 8 stages; 34M / 70M / untied 70M: 1 / 2 / 2 further trajectories, 40 checkpoint evaluations
Pairing across training histories	Tied GPT-2-style 70M: 4 additional initialization/data-order trajectories, checkpoints 512/2048, source ages 16/64; four-state pairing comparisons
Stability across scale and family	Pythia 1B / 6.9B: 2 data orders and 2 stages per size; Mistral-7B: order 1 at 512 and order 2 at 512/2048
Coverage predictions	Two 70M parameter paths varying q; untied Pythia-1B comparison at two stages
Direction and persistence	Untied NeoX-style 70M: 2 trajectories, 4 checkpoints at equal radius; 70M long horizon: 2 trajectories, 8 checkpoints
Recovery from durable files	Two write schedules, 128 paired crash phases each; 72 disk-recovery branches at 8 checkpoints from 2 trajectories; three phase priors
Selective saving and repair	Randomly initialized Pythia-1B: 2 trajectories, 48 state-matrix / 36 end-to-end / 24 envelope records; 8 terminals reevaluated on 2,048 documents
Direct native recovery	OLMo-2 1B: five policies; 7B: FULL/REBUILD_M; original batches, schedules, and data cursors, 128 updates
Repeated native recovery	OLMo-2 1B: FULL/REBUILD_M/REBUILD_MV, 384 updates, faults at 0/128/256
Recovery across source states	OLMo-2 1B/7B: 6 source states, five policies, 512 recovery updates
Composition with precision	Four OLMo-2 1B sources: matched FULL/REBUILD_M at BF16 and block-wise 8-bit storage, 512 updates

Sampling grids and finite contrasts. Scale and Mistral profiles use lags $0, 1, 2, 4, 8, 16, 24, 32, 48, 64, 96, 128$; the small-model, direction, and system experiments use $0, 1, 2, 4, 8, 16, 32, 64, 128$. The first row and column of I are zero by definition, yielding nine structural zeros in its 25-cell profile. We also examine the remaining 16 cells to resolve the nontrivial interaction structure. Rank-two reconstruction summarizes within-profile matrix structure.

Pairing and evaluation. Each original transplant branch is paired with a FULL continuation sharing the same current parameters, optimizer clock, scheduler, RNG, future batch IDs, validation examples, and arithmetic path. The signed primary statistic in Equation 2 uses the stated lag grid and trapezoidal integration. Negative values indicate lower NLL than FULL. We use ± 0.001 nats/token as a reference scale for interpreting small differences.

Architectures and restoration controls. The small-model experiments use GPT-2-style decoders (Radford et al., 2019) and GPT-NeoX-style decoders (Black et al., 2022), with tied or untied input/output weights as indicated. Serialization round trips, repeated FULL branches, and copy controls check that restored parameters and optimizer states reproduce the intended continuation.

Recorded execution environments. The following summarizes retained runtime records for the recovery experiments; batch sizes, precision, and evaluation protocols are specified with each experiment.

Experiment	Compute and software	Storage context
70M durable recovery	RTX PRO 6000 Blackwell; PyTorch 2.7.1, CUDA 12.8; FP32	Shared host, local block-device filesystem
1B native recovery	Single RTX PRO 6000 Blackwell; PyTorch 2.9.1, CUDA 12.8	Host filesystem; complete serialized recovery payloads
7B native recovery	Two-GPU full-shard FSDP; PyTorch 2.9.1, CUDA 12.8	Local hot state; separate cold archive

Timing measurements are observed end-to-end costs under the recorded shared workloads, not isolated storage-throughput benchmarks. Timing conditions and the operations included in each measurement are described in Appendices D.5.3 and D.7.

A.2 Reading NLL and Recovery Responses

For N evaluated next-token targets, negative log-likelihood is

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x_i \mid x_{<i}). \tag{11}$$

We use natural logarithms, so the unit is nats/token. Lower NLL means that the model assigns higher average log-probability to the observed tokens. Perplexity is $\exp(L)$.

Absolute loss versus recovery difference. $L_b(h)$ is a branch's absolute validation loss. Its matched recovery difference is $\Delta L_b(h) = L_b(h) - L_{\mathrm{FULL}}(h)$ at the same logical update and on the same targets. Positive differences mean worse performance; negative differences mean better performance. A curve near zero follows FULL closely even when FULL itself changes over training. We therefore report absolute FULL losses for the reduced-batch continuation experiments (Appendix D.3) in Table 18.

Time-integrated response versus late quality. Profile response Y is the signed trapezoidal time average over 128 updates (Equation 2); the application matrix uses 512 updates (Equation 73). Early and late effects can cancel in these signed summaries, so trajectories are shown as well. The reduced-batch continuation's fixed late statistic instead uses weights $0.25, 0.50, 0.25$ at updates 384,448,512: the normalized trapezoidal average over the final 128-update interval. Its endpoint statistic uses update 512 alone. The direct native-condition experiment instead reports the update-128 endpoint (Table 1). Equal-radius direction tests use the time average of $|\Delta L|$, a different quantity.

Interpreting a small difference. On matched targets, an instantaneous NLL difference Δ corresponds exactly to a perplexity ratio e^{Δ}. Thus 0.001 nats/token is about a 0.1% relative perplexity increase. This interpretation does not turn a time-averaged statistic into a single endpoint perplexity. Comparisons across tokenizers retain each model's native units rather than treating tokens as interchangeable.

B History-Pairing Structure

B.1 Pairing across Independent Training Histories

The source-pairing asymmetry recurs across independently trained 70M models. We vary both initialization and data order across four trajectories and compare the same moment pairings at two checkpoints in each trajectory (Table 4).

Training histories and interventions. Each model is a GPT-2-style decoder with tied input/output weights, 69,997,056 parameters, 14 layers, eight attention heads, hidden width 512, and feed-forward width 2,048. Seeds 20260921–20260924 determine independent initializations and hash-ordered FineWeb training streams, using the same document-role and token-window rules and 48 fixed, disjoint validation sequences. Parents train for 2,048 updates with batch eight and sequence length 256. AdamW uses learning rate 3×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$ and weight decay 0.1; training is FP32 with zero dropout.

At $t \in \{512, 2048\}$ and $d \in \{16, 64\}$, we compare FULL, (m_t, v_{t-d}), (m_{t-d}, v_{t-d}), and (m_{t-d}, v_t) at the same current weights. Each checkpoint shares its FULL branch across the two ages, giving seven branches per checkpoint and 56 in total. We hold the optimizer clock, RNG, future batches, and validation sequences fixed within each comparison. All branches continue for 128 updates and are evaluated at $h \in \{0, 1, 2, 4, 8, 16, 32, 64, 128\}$. The response Y and contrasts B_d, P_d use Equations 2–4; each trajectory summary equally averages its two checkpoints and two ages.

Table 4: Pairing across four independent 70M training histories. Contrasts use the time-averaged excess NLL over 128 updates (nats/token). Positive B_d indicates a coherent-pair benefit; positive P_d indicates a gain beyond additive single-component effects. Each trajectory mean equally averages its two checkpoints and two ages.

Trajectory (seed)	Checkpoint	Age 16		Age 64	
		B_{16}	P_{16}	B_{64}	P_{64}
20260921	512	5.7687	5.7728	7.5715	7.5748
	2048	2.5297	2.5325	2.7320	2.7370
	Mean	$\bar{B} = 4.6505$,		$\bar{P} = 4.6543$	
20260922	512	7.9470	7.9517	6.7278	6.7315
	2048	2.5283	2.5313	2.6699	2.6703
	Mean	$\bar{B} = 4.9683$,		$\bar{P} = 4.9712$	
20260923	512	6.8768	6.8808	7.3192	7.3213
	2048	2.2182	2.2212	1.9805	1.9872
	Mean	$\bar{B} = 4.5987$,		$\bar{P} = 4.6026$	
20260924	512	7.8619	7.8660	6.0264	6.0299
	2048	2.6836	2.6887	2.6374	2.6402
	Mean	$\bar{B} = 4.8024$,		$\bar{P} = 4.8062$	

Both contrasts are positive in every checkpoint–age comparison and in each trajectory summary. The benefit of a coherent older pair over newer-m/older-v therefore persists across the four initialization and data-order choices, alongside the cross-model pattern in Section 3.

B.2 Complete Local-History Profiles and Sensitivity Checks

Exact construction and labels. We use Pythia-1B and Pythia-6.9B weights (Biderman et al., 2023); optimizer histories start locally from zero moments. The model repository revisions used for these local-history profiles are listed below.

Source	Model / release	Revision (commit)
P1	Pythia-1B	`f73d7dcc545c8bd326d8559c8ef84ffe92fea6b2`
P7	Pythia-6.9B	`c0e3eee36dc47af0c49f361c74cfe459c09f7f23`
M7	Mistral-7B-v0.3	`c882233d224d27b727b3d9299b12a9aab9dda6f7`

Orders 1 and 2 are the data permutations 2026090701 and 2026090702, shared across sizes. Pair labels in the appendix are: pair 1 = order 1 / step 2048; pair 2 = order 1 / step 512; pair 3 = order 2 / step 2048; pair 4 = order 2 / step 512. Both sizes use sequence length 256, global batch eight, FP32 parameters and moments, eager attention, zero dropout, deterministic math, AdamW $(.9, .95)$, $\epsilon = 10^{-8}$, and weight decay 0.1, with FSDP execution. Historical sources cover ages $\{0, 1, 4, 16, 64\}$ at each target; all 31 branches continue for 128 updates.

Cross-family matching. Mistral uses the base `Mistral-7B-v0.3` release (source M7 above) (Mistral AI, 2024), also with locally constructed moment histories. The 18,000 background and 48 validation FineWeb documents (Penedo et al., 2024) match the Pythia selection. Document roles, order keys, and window rule are shared, while native Mistral tokenization determines its token IDs and boundaries. The learning rate is 1.2×10^{-5}, shared with Pythia-6.9B. Profiles cover order 1 / step 512 and order 2 / steps 512, 2048 (Figures 2 and 9). The original order-1 profile's rank-two relative reconstruction errors are 0.0872 for Y and 0.1220 for I. The 16-cell interaction correlations are 0.824 against Pythia-6.9B and 0.915 against Pythia-1B, respectively.

Definitions of diagnostics. The full-profile statistic is Spearman correlation over all 25 cells, including the zero FULL cell. Interaction correlation uses 25 cells, nine of which are identically zero. The strict-upper and strict-lower comparisons each use ten cells; the additional interaction summary uses $I[1:, 1:]$ (16 cells). All-zero ties use standard midranks. For $x = \mathrm{vec}(Y_{1B})$ and $y = \mathrm{vec}(Y_{6.9B})$, $\alpha = x^\top y/(x^\top x)$ and $r = \|y - \alpha x\|_2/\|y\|_2$. These give the least-squares amplitude and relative residual. Singular-value truncation is applied separately to each Y or I; a low rank-two residual can coexist with weak transport between two different matrices. These diagnostics use native signed values without per-model normalization.

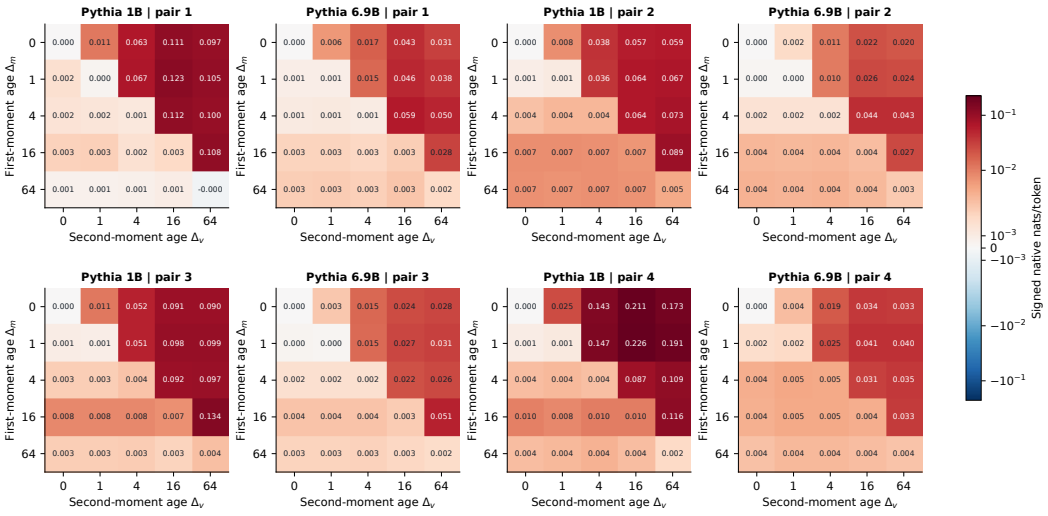

Figure 6: **All eight Pythia response profiles.** Pair identities are defined above. One common signed symmetric-log color scale covers every panel. Labels are untransformed native nats/token, rounded only for display.

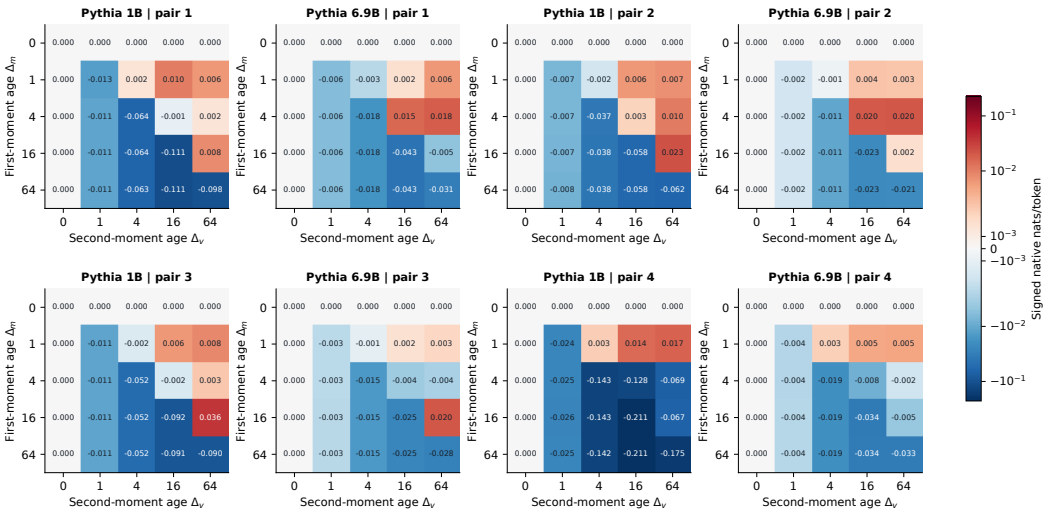

Figure 7: **All eight anchored Pythia interaction profiles.** The entire first row and column vanish under the anchored contrast in Equation 3. The 16-cell correlations summarize the remaining interaction entries.

Table 5: Per-environment pairing contrasts and rank-two relative Frobenius reconstruction error.

Model / pair	B_{16}	B_{64}	P_{16}	P_{64}	$r_2(Y)$	$r_2(I)$
Pythia 1B / 1	0.1086	0.0976	0.1114	0.0982	0.1550	0.2482
Pythia 6.9B / 1	0.0401	0.0281	0.0429	0.0312	0.1392	0.1801
Pythia 1B / 2	0.0509	0.0546	0.0583	0.0617	0.1495	0.2566
Pythia 6.9B / 2	0.0184	0.0172	0.0225	0.0215	0.1540	0.2188
Pythia 1B / 3	0.0835	0.0865	0.0916	0.0896	0.1347	0.2429
Pythia 6.9B / 3	0.0213	0.0255	0.0249	0.0283	0.1161	0.1969
Pythia 1B / 4	0.2010	0.1709	0.2106	0.1747	0.0904	0.1034
Pythia 6.9B / 4	0.0297	0.0296	0.0342	0.0333	0.1194	0.1779

Table 6: All four Pythia scale pairs. Rank correlations use the full response (ρ_{25}) and anchored interaction on all cells (ρ_{I25}) or its 16-cell nonzero support (ρ_{I16}); α and r give the fitted amplitude and relative residual.

Family / pair	ρ_{25}	ρ_{I25}	ρ_{I16}	α	r
Pythia / 1	0.795	0.806	0.903	0.377	0.247
Pythia / 2	0.921	0.982	0.935	0.418	0.304
Pythia / 3	0.958	0.891	0.868	0.304	0.163
Pythia / 4	0.869	0.981	0.932	0.198	0.277

Table 7: Pythia and Mistral action outcomes, as signed time-averaged excess NLL relative to FULL. Coherent means global first-moment rollback to the age of the second moment; current weights and clock remain fixed.

Model / row	d	Coherent	Head reset	Head rollback	Raw
Pythia 1B O1/2048	16	0.00251	0.01572	0.01729	0.11109
Pythia 1B O1/2048	64	-0.00045	0.02295	0.02384	0.09717
Pythia 1B O1/512	16	0.00652	0.01079	0.01113	0.05746
Pythia 1B O1/512	64	0.00470	0.01502	0.01642	0.05935
Pythia 1B O2/2048	16	0.00718	0.01551	0.01722	0.09063
Pythia 1B O2/2048	64	0.00373	0.02032	0.02246	0.09026
Pythia 1B O2/512	16	0.01003	0.01792	0.02022	0.21104
Pythia 1B O2/512	64	0.00250	0.02288	0.02402	0.17342
Pythia 6.9B O1/2048	16	0.00306	0.00993	0.01132	0.04320
Pythia 6.9B O1/2048	64	0.00249	0.01097	0.01019	0.03060
Pythia 6.9B O1/512	16	0.00385	0.00430	0.00502	0.02227
Pythia 6.9B O1/512	64	0.00324	0.00717	0.00800	0.02039
Pythia 6.9B O2/2048	16	0.00320	0.00589	0.00650	0.02448
Pythia 6.9B O2/2048	64	0.00230	0.00825	0.00890	0.02781
Pythia 6.9B O2/512	16	0.00448	0.00748	0.00850	0.03421
Pythia 6.9B O2/512	64	0.00359	0.00914	0.00995	0.03317
Mistral 7B O1/512	16	0.01468	0.02909	0.03005	0.05470
Mistral 7B O1/512	64	0.00928	0.05474	0.06513	0.09197
Mistral 7B O2/512	16	0.01697	0.05443	0.05402	0.15568
Mistral 7B O2/512	64	0.01177	0.06634	0.06757	0.22741
Mistral 7B O2/2048	16	0.00864	0.01770	0.01697	0.15649
Mistral 7B O2/2048	64	0.00368	0.02825	0.02910	0.23317

B.3 PROFILE CONSTRUCTION AND FULL PAIRING COMPARISONS

Table 8: Profile experiments. Each environment contains 25 age cells, two repeat/copy controls, and four head-local actions, continued for 128 updates.

	Pythia 1B / 6.9B	Mistral 7B extension
Orders × target steps	$2 \times \{512, 2048\}$ per size	O1: 512; O2: 512, 2048
Environments / branches	8 / 248 (4 scale pairs)	3 / 93
Stream matching	Identical token arrays	Same documents; native tokenizer
AdamW learning rate	3×10^{-5} / 1.2×10^{-5}	1.2×10^{-5}
Shared construction	FP32; $(\beta_1, \beta_2) = (.9, .95)$; sequence 256; batch 8	

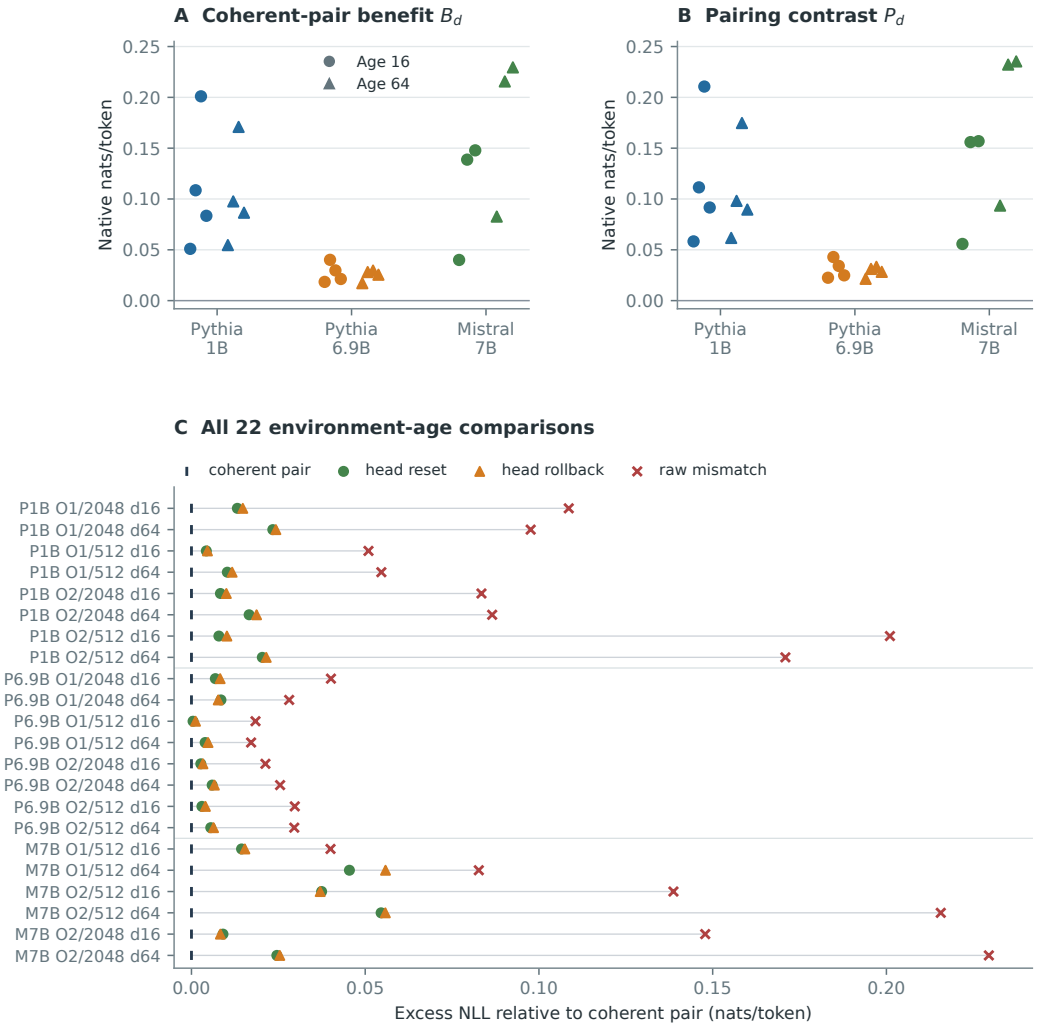

Figure 8: **Pairing benefits are consistent across the tested environments.** (A–B) Rollback benefit B_d and advantage over additivity P_d, for age 16 (circles) and 64 (triangles). (C) Each action row is centered on its coherent outcome; horizontal distances are signed loss differences. Rows identify environment–age comparisons: 16 for Pythia and six for Mistral.

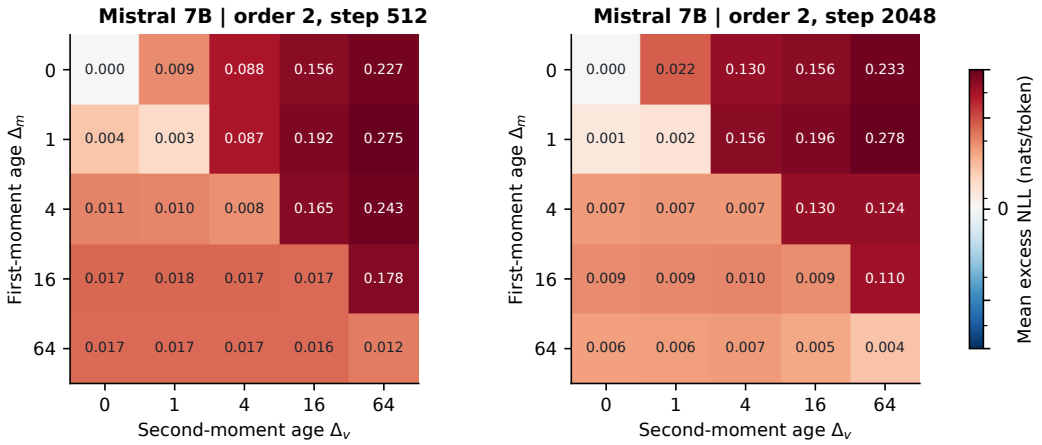

Figure 9: **The Mistral source-pairing structure repeats across stages in a second data order.** All 25 age cells at each stage use the same response definition and evaluation grid as the original profile. Both panels share a signed symmetric-log color scale (linear threshold 0.005).

Table 9: All four Pythia 1B-to-6.9B pairs. Spearman correlations ρ_{25}, ρ_{U10}, and ρ_{L10} compare all 25 response cells and the ten cells in each strict triangle; ρ_{I16} compares the 16-cell interaction support. α fits $Y_{6.9B} \approx \alpha Y_{1B}$ without an intercept; r is its relative Frobenius residual.

Order / stage	ρ_{25}	ρ_{U10}	ρ_{L10}	ρ_{I16}	α	r
O1 / 2048	0.795	0.794	-0.588	0.903	0.377	0.247
O1 / 512	0.921	0.842	0.491	0.935	0.418	0.304
O2 / 2048	0.958	0.867	0.903	0.868	0.304	0.163
O2 / 512	0.869	0.673	0.661	0.932	0.198	0.277

B.4 Untied Localization and Selective Reset

Untying input embedding and output head makes role attribution possible. Across eight seed-by-stage clusters, the output head carries median 99.22% of absolute group-local response, compared with 0.70% for the input embedding and below 0.1% for each remaining group (Figure 10A).

The localization motivates a selective-reset comparison for the age-16 fault (m_t, v_{t-16}). It zeros only output-head m while retaining historical v and all non-head state. After h gradients on the reset branch,

$$m^R_{t+h} = (1-\beta_1)\sum_{j=0}^{h-1}\beta_1^j g_{t+h-j}, \tag{12}$$

$$v^R_{t+h} = \beta_2^h v_t^{\text{old}} + (1-\beta_2)\sum_{j=0}^{h-1}\beta_2^j g^2_{t+h-j}. \tag{13}$$

Weighted Cauchy–Schwarz therefore gives

$$|m^R_{t+h}|^2 \le \frac{(1-\beta_1)^2}{1-\beta_2}\left(\sum_{j=0}^{h-1} q^j\right) v^R_{t+h}. \tag{14}$$

The new numerator is covered by denominator mass from the same gradients; retained historical v only adds nonnegative support. The statement is valid on the reset branch's own closed-loop gradient sequence. The following experiments measure the resulting recovery quality.

On one untied 70M trajectory, head-m reset lowers mean risk from 2.3362 to 0.01051 and removes median 99.52% of measured fault response. Across two Pythia-1B trajectories it lowers mean risk from 0.1297 to 0.01082 and removes median 92.45%. Input-embedding, non-head, equal-size non-head, and reset-both controls do not match it (Figure 10B–C). Several clusters remain above 0.01, and global rollback gives a smaller residual response.

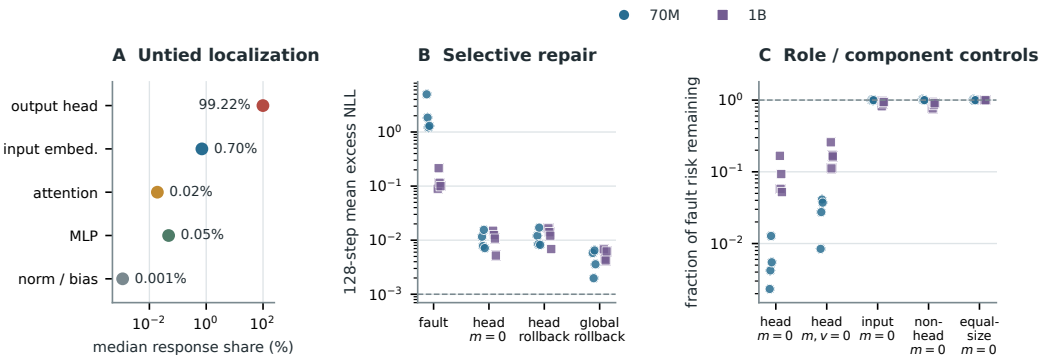

Figure 10: **Output-head localization guides selective reset.** (A) Median absolute group-local response share in the untied 70M battery. (B) Absolute 128-update risk for the fault and repair choices at 70M and 1B; points show the checkpoint stages. (C) Remaining-fault fractions for role-, component-, and size-matched controls.

C Mechanisms and State Construction

C.1 Coverage Bound and Moment-Coherent Companion

For one coordinate, define

$$a_k = (1-\beta_1)\beta_1^{s-k}, \qquad b_k = (1-\beta_2)\beta_2^{r-k}.$$

When $s \le r$, v_r includes every squared gradient indexed by $k \le s$. Weighted Cauchy–Schwarz gives

$$|m_s|^2 = \left|\sum_{k=1}^{s} a_k g_k\right|^2 \tag{15}$$

$$\le \left(\sum_{k=1}^{s} \frac{a_k^2}{b_k}\right)\left(\sum_{k=1}^{s} b_k g_k^2\right) \le C_{s,r} v_r, \tag{16}$$

where

$$\sum_{k=1}^{s} \frac{a_k^2}{b_k} = \frac{(1-\beta_1)^2}{1-\beta_2} \sum_{k=1}^{s} \frac{\beta_1^{2(s-k)}}{\beta_2^{r-k}} \tag{17}$$

$$= \frac{(1-\beta_1)^2}{1-\beta_2} \beta_2^{-(r-s)} \sum_{j=0}^{s-1} \left(\frac{\beta_1^2}{\beta_2}\right)^j = C_{s,r}. \tag{18}$$

The actual transplants retain the current clock t. With current-clock bias correction,

$$\frac{|\widehat{m}_s|^2}{\widehat{v}_r} \le C_{s,r} \frac{1-\beta_2^t}{(1-\beta_1^t)^2}. \tag{19}$$

Adding $\epsilon > 0$ to AdamW's denominator can only reduce the corresponding coordinatewise preconditioned magnitude. For fixed or bounded $r - s$, the shared-history sum is bounded by $(1-q)^{-1}$ when $q < 1$. The prefactor $\beta_2^{-(r-s)}$ prevents an unqualified bound for a source gap that itself grows without limit.

For $s > r$, choose the gradients up to r and hence v_r as fixed, then scale one gradient in $(r, s]$ arbitrarily. It enters m_s but not v_r, so no amplitude-independent bound can depend only on the four scalar indices and decays. A data-dependent comparison remains available. Define

$$\bar{g}_{r+1:s}^2 = \frac{1-\beta_2}{1-\beta_2^{s-r}} \sum_{k=r+1}^{s} \beta_2^{s-k} g_k^2. \tag{20}$$

The EMA recurrence yields

$$v_s = \beta_2^{s-r} v_r + (1-\beta_2^{s-r})\bar{g}_{r+1:s}^2, \tag{21}$$

$$\frac{v_s}{v_r} = 1 + (1-\beta_2^{s-r})\left(\frac{\bar{g}_{r+1:s}^2}{v_r} - 1\right) = \Lambda_{r:s}. \tag{22}$$

For positive denominators, the fault (m_s, v_r) and companion (m_s, v_s) have the same numerator, giving

$$\frac{m_s^2}{v_r} = \frac{m_s^2}{v_s}\Lambda_{r:s}. \tag{23}$$

$\Lambda_{r:s}$ exceeds one precisely when the missing interval has larger weighted squared-gradient load than the older v_r. This identity explains why source order changes the available guarantee, while leaving functional magnitude dependent on the realized history and current training state.

C.2 Prospective q Tests and Realized Forcing

The first-update output-head forcing statistic in the gap experiment is

$$F(q,d) = \|u_{\text{fault}}^{\text{head}}(q,d) - u_{\text{FULL}}^{\text{head}}(q)\|_2. \tag{24}$$

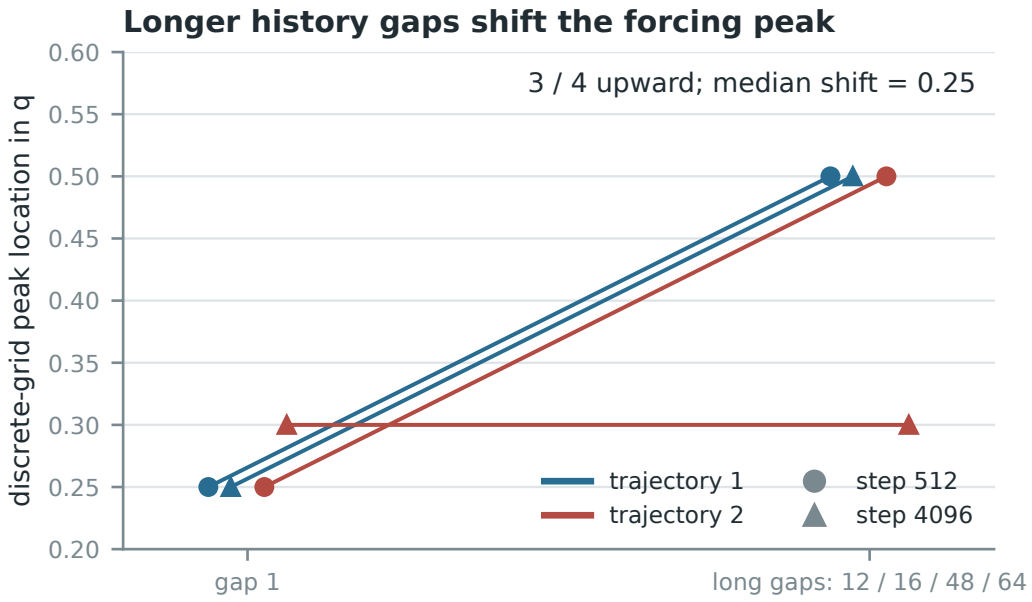

Figure 11: **History gap changes the location of the forcing peak.** Each line connects the discrete-grid peak at gap one to the median peak across gaps 12/16/48/64 in the same seed–stage row. Three of four rows shift upward. This first-update diagnostic complements the covered-side prediction in Figure 3.

Two independent crossings. The prospective 70M experiment varies β_2 at fixed β_1 and varies β_1 at fixed β_2. Both paths cross $q = \beta_1^2/\beta_2$ from .95 to 1.05. Covered-side lifts are positive along both paths; different beta pairs matched on q retain median surface-profile Spearman 0.986. The uncovered triangle remains far larger in absolute magnitude but generally moves in the opposite direction: all four seed-by-stage region medians decline on each path, while the endpoint cell direction is 40/40 and 34/40. An untied Pythia-1B anchor declines at 2/2 measured stages. Increasing q therefore redistributes response between the covered and uncovered regions.

History-gap peak shift. The two-seed gap experiment uses two stages per trajectory and tests a q grid concentrated around a predicted forcing maximum. At each (q, d), forcing is the first-update output-head update gap $F(q,d) = \|u^{\text{head}}_{\text{fault}}(q,d) - u^{\text{head}}_{\text{FULL}}(q)\|_2$. Long gaps move the discrete-grid peak upward relative to gap one in 3/4 environments, by median 0.25 in q. Across all cells, the exclusive-history kernel has Spearman 0.744 with forcing, but it does not beat a simpler common-hump baseline in both cross-seed folds. The peak shift and its history dependence are the robust findings of this comparison.

For a gap $d = s - r > 0$, the exclusive-history kernel score is

$$K(\beta_1, d) = \left[\frac{\beta_1^2(1-\beta_1)}{1+\beta_1}(1-\beta_1^{2d})\right]^{1/2}.$$

It is the coefficient norm of the missing numerator history after one further recurrence, $\beta_1(1-\beta_1)\sum_{j=0}^{d-1}\beta_1^j g_{s-j}$. The common-hump score is $H(\beta_1) = \beta_1(1-\beta_1)$. Each cross-seed fit predicts $\log_{10} F$ by least squares using an intercept, stage and source-cell indicators, and either $\log_{10} K$ or $\log_{10} H$; the other seed supplies the test fold, scored by log-scale RMSE.

Exact first-update factorization. Compare a fault $F = (m_s, v_r)$ and moment-coherent companion $C = (m_s, v_s)$ under the same current gradient, parameters, clock, and learning rate. Let $\widehat{m}_i$ be their common post-gradient bias-corrected numerator and $d_{F,i}, d_{C,i}$ the corresponding denominators. Weight decay cancels, and the squared output-head update gap is exactly

$$E = \|u_F - u_C\|_2^2 = \eta^2 L_m \bar{D} P^\star, \qquad L_m = \sum_i \widehat{m}_i^2, \quad \bar{D} = \frac{1}{n}\sum_i (d_{F,i}^{-1} - d_{C,i}^{-1})^2, \tag{25}$$

with normalized coordinate-overlap factor

$$P^\star = \frac{n \sum_i \widehat{m}_i^2 (d_{F,i}^{-1} - d_{C,i}^{-1})^2}{(\sum_i \widehat{m}_i^2)(\sum_i (d_{F,i}^{-1} - d_{C,i}^{-1})^2)}. \tag{26}$$

$P^\star = 1$ is the fixed-marginal random-pairing reference.

Across the fixed-β_1 endpoints, increasing q raises numerator load and denominator defect in 38/40 and 38/40 cells, respectively, but lowers overlap in 40/40 and net squared gap in 34/40. Median log changes are 0.079, 0.603, -0.998, and -0.190. The exact decomposition residual is numerical zero. The first-step gap has pooled Spearman 0.641 with later natural response (Figure 12A). The matched-impulse experiment in Section C.3 then isolates directional gain by holding the perturbation radius fixed.

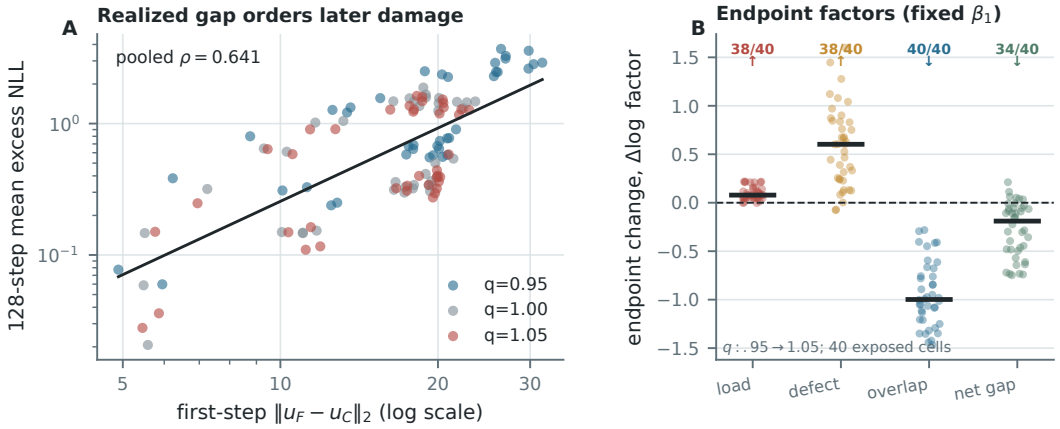

Figure 12: **First-update geometry connects history to later response.** (A) The realized output-head update gap orders later functional response. (B) Along the fixed-β_1 endpoints, numerator load and denominator defect usually increase with q, coordinate overlap always falls, and the net gap falls in 34/40 cells. Each point is a grid cell in a seed–stage environment.

C.3 Equal-Radius Directions and the Local Response Expansion

Direction construction. Within each untied NeoX-style 70M seed–stage environment, we compute the output-head first-update delta of the $q = .4$ upper source cell $(\Delta_m, \Delta_v) = (0, 16)$ relative to FULL. This defines the upper unit direction and its raw norm. The lower direction uses the corresponding $(16, 0)$ delta, normalized to the same radius. A deterministic Gaussian direction is also normalized to that radius. None uses validation outcomes. Within that environment, we inject the same fixed vector into every tested q runtime and continue for 128 updates. The $\alpha = 0$ pair, $\alpha \in \{.5, 1, 2\}$ upper scaling, and matched lower/random branches check endpoint semantics and radius dependence.

For $\alpha = 1$, the median absolute functional responses are 1.3887 (interquartile range 0.955–2.001), 0.1246 (IQR 0.045–0.155), and 0.00773 (IQR 0.0067–0.0121). All four seed–stage q profiles vary by more than the 0.001 resolution, but one is non-monotone. Median parameter-distance retention ρ_{128} (Equation 29) is 2.318. The equal-radius contrasts and retained displacement establish directional sensitivity and persistent trajectory effects.

Local response expansion. The directional model contains 70,378,496 parameters. For augmented state $x = (\theta, m, v, c)$ and fixed exogenous batches and randomness ω_k, write

$$x_{k+1} = F_k(x_k; \omega_k), \qquad \delta x_0 = E_s \delta s. \tag{27}$$

Along the reference trajectory, with Jacobian J_k and parameter projection P_θ, the local readout is

$$\delta\theta_h \approx P_\theta J_{h-1} \cdots J_0 E_s \delta s, \qquad \Delta L(h) \approx g_h^\top \delta\theta_h + \tfrac{1}{2} \delta\theta_h^\top \mathcal{H}_h \delta\theta_h. \tag{28}$$

This local approximation separates injection, propagation, and readout; the experiments measure their combined effect along the actual trajectory. Here $g_h = \nabla_\theta L_{\text{val}}(\theta_h)$ and $\mathcal{H}_h = \nabla_\theta^2 L_{\text{val}}(\theta_h)$ are the validation gradient and Hessian at the reference endpoint, rather than the training-batch gradient.

The linear readout term is necessary because the outcome is a signed loss difference that need not be nonnegative; a positive-semidefinite quadratic form alone cannot represent both signs. The quadratic term captures curvature-mediated directional gain. For the specific operation of completely resetting m, Appendix C.8 controls the response-expansion residual through an exact compensated-state representation.

The equal-radius experiment retains the following parameter-distance ratio at update 128:

$$\rho_{128} = \frac{\|\theta_{128}^{\text{impulse}} - \theta_{128}^{\text{reference}}\|_2}{\|\theta_0^{\text{impulse}} - \theta_0^{\text{reference}}\|_2}, \qquad \text{median}(\rho_{128}) = 2.318. \tag{29}$$

C.4 Inherited decay and closed-loop forcing

Identical gradients imply

$$\delta m_K = \beta_1^K \delta m_0, \qquad \delta v_K = \beta_2^K \delta v_0. \tag{30}$$

With each branch following its own gradients, the exact recurrences are

$$\delta m_h = \beta_1^h \delta m_0 + (1-\beta_1) \sum_{k=1}^{h} \beta_1^{h-k} \delta g_k, \tag{31}$$

$$\delta v_h = \beta_2^h \delta v_0 + (1-\beta_2) \sum_{k=1}^{h} \beta_2^{h-k} \big[(g_k^F)^2 - (g_k^C)^2\big]. \tag{32}$$

The first terms describe inherited state; the sums describe state generated by divergent gradients. Appendix C.5 decomposes their update effects.

C.5 Closed-Loop Dynamics and Long-Horizon Evaluation

Endpoint-clamp protocol. The duration experiment fixes $K \in \{1, 2, 4, 8, 16\}$. For the first K updates it keeps the fault moments, executes the natural branch, and then replaces the output-head parameter endpoint with the paired moment-coherent endpoint. After update K the clamp is released. Every K receives the same post-release event-time horizon. The post-release residual falls from roughly 84% at $K = 1$ to 1.6% at $K = 16$ (Figure 13A). The exact common-gradient contraction in Equation 30 explains why longer co-evolution reduces inherited state, while the natural recurrence Equations 31–32 in Appendix C.4 explain why one corrected endpoint is insufficient.

Exact update-channel decomposition. Let $\mathcal{A}(\xi, g)$ be the adaptive output-head update excluding decoupled weight decay, with optimizer state ξ. For the natural fault and moment-coherent branches,

$$\begin{aligned} \Delta u_t &= H_t + F_t + P_t, \\ H_t &= \mathcal{A}(\xi_t^F, g_t^C) - \mathcal{A}(\xi_t^C, g_t^C), \\ F_t &= \mathcal{A}(\xi_t^F, g_t^F) - \mathcal{A}(\xi_t^F, g_t^C), \\ P_t &= -\eta\lambda(\theta_t^F - \theta_t^C). \end{aligned} \tag{33}$$

H_t changes moment state under the reference gradient, F_t changes the gradient at fixed fault state, and P_t is the remaining weight-decay term. The three terms telescope to the exact update difference.

Reference-gradient replay supplies g_t^C to the fault branch and suppresses parameter-conditioned gradient adaptation. It exposes a much larger open-loop response than the natural trajectory (Figure 13B). In four paths, the state-mediated and gradient-adaptation channels have negative cosine at all 60/60 measured nonzero lags; the combined adaptive-channel gain is below one at 58/60 (panel C). Natural feedback is therefore opposing over these paths, but not fast enough to erase the accumulated displacement.

Long-horizon evaluator. The 1,024-update extension preserves target weights, moment sources, clock, scheduler, RNG, and future stream. The evaluator contains eight fixed shards, 48 length-256 sequences per shard, and 255 next-token targets per sequence: 384 sequences and 97,920 positions.

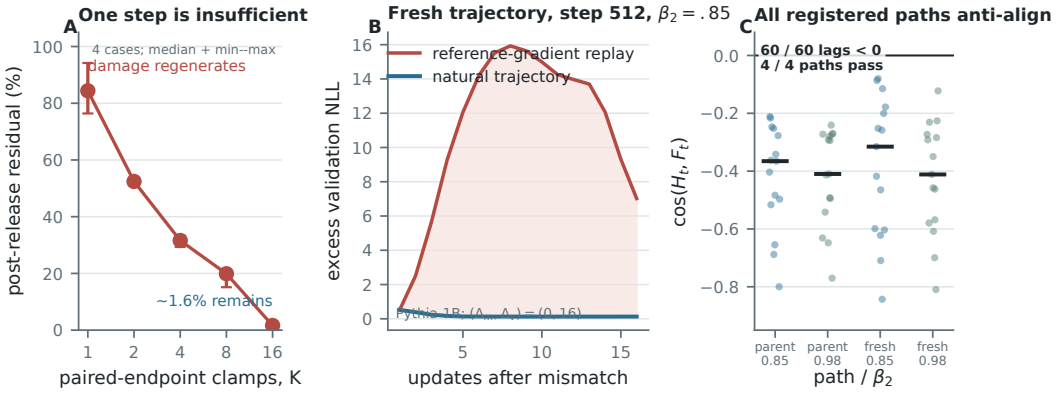

Figure 13: **Bounded interventions distinguish state regeneration from gradient feedback.** (A) One corrected parameter endpoint is insufficient; longer clamping lets the hidden moments relax under shared endpoints. (B) Reference-gradient replay exposes the state-mediated open-loop response for one representative path. (C) Measured natural adaptation anti-aligns with that channel at all 60 measured nonzero lags across four paths.

Each continuation processes 2.09 million next-token positions. The evaluator is shared by eight checkpoints from two training trajectories.

At update one, median absolute NLL is 59.31 for new-m/old-v and 6.37 for FULL. At update 1,024 they are 6.267 and 6.160; the median signed excess is 0.1068. Updates 128–1,024 contain median 32.95% of signed trapezoidal response area, and updates 512–1,024 contain 9.90%. The expanded versus original 0–128 evaluator has slope 1.004, concordance 0.9999, and identity MAE 0.022. Removing token types selected on disjoint shards by the top 1% or 5% of positive response mass retains median 68.41% and 55.82% of signed mass. The long-horizon response remains visible with wider evaluation coverage and after removing the most responsive token types.

C.6 A COHERENT-MOMENT ENVELOPE: INVARIANCE AND REPAIR

An invariant for uncorrected moments. For fixed $0 \le \beta_1 < 1$, $0 < \beta_2 < 1$, and $\beta_1^2 < \beta_2$, define

$$C = \frac{(1-\beta_1)^2}{(1-\beta_2)(1-\beta_1^2/\beta_2)}, \qquad \mathcal{E}_C = \{(m, v) : v_i \ge 0,\ m_i^2 \le C v_i \text{ for every } i\}. \tag{34}$$

With $m^+ = \beta_1 m + (1-\beta_1)g$ and $v^+ = \beta_2 v + (1-\beta_2)g^2$, weighted Cauchy–Schwarz yields

$$(m^+)^2 \le \left[\frac{\beta_1^2 C}{\beta_2} + \frac{(1-\beta_1)^2}{1-\beta_2}\right][\beta_2 v + (1-\beta_2)g^2] = C v^+. \tag{35}$$

Thus $\mathcal{E}_C$ is forward invariant for arbitrary future gradients, including a repaired branch's own gradients, in exact arithmetic. Zero-initialized coherent histories satisfy the tighter finite-history constant $C_t = C[1 - (\beta_1^2/\beta_2)^t] \le C$. The chosen betas give $C = 19/14$. This is the same-source specialization of the coverage argument; a covered pair with a nonzero source gap uses the larger gap-dependent bound in Equation 6, and can violate this particular coherent envelope.

Projection and its implementation. For fixed valid v, the coordinatewise Euclidean projection is

$$\widetilde{m}_i = \operatorname{sign}(m_i)\min\{|m_i|, \sqrt{C v_i}\}, \qquad \widetilde{v}_i = v_i. \tag{36}$$

It preserves sign and healthy coordinates, and forces $m_i = 0$ where $v_i = 0$. Nonfinite moments or negative v fail validation. The implementation compares in FP64 chunks and rounds repaired caps inward, leaving healthy coordinates unchanged. Subsequent finite-precision updates are subject to rounding. Bias correction uses the preserved clock, and decoupled decay is outside the adaptive-moment constraint. A sign-reversed first moment can satisfy the envelope while giving an ascent direction, so it is not a convergence or NLL guarantee. Empirically, resetting m alone and resetting both moments give a second illustration: both are inside the envelope, but retaining v materially changes their response.

The four-corner summary is in Table 11.

Completing the source-order comparison. The new-m/old-v stress comparison uses step 512, age 64, for both seeds. Table 11 summarizes all four source combinations on the shared grid $0, 1, 2, 4, 8, 16, 32, 64$. The four source combinations share a 64-update horizon. Both the reversed source order and old coherent pair stay near FULL compared with the large new-m/old-v response.

Figure 14 shows the stress trajectories; Table 10 supplies all healthy-state controls.

Table 10: All 24 envelope-block records: mean signed excess NLL versus FULL over 64 updates. Four envelope actions are exact no-op aliases. Envelope/reset act once; adaptive clipping is applied at every update.

Initial state	Action	Seed 1	Seed 2
FULL	Unmodified	+0.00000	+0.00000
FULL	Envelope	+0.00000	+0.00000
FULL	Reset m	-0.00904	-0.00862
FULL	Adaptive clip	-0.00060	+0.00000
Old coherent	Unmodified	+0.00724	+0.00888
Old coherent	Envelope	+0.00724	+0.00888
Old coherent	Reset m	-0.00315	-0.00592
Old coherent	Adaptive clip	+0.00613	+0.00888
New m / old v	Unmodified	+19.41069	+24.31780
New m / old v	Envelope	+0.00744	+0.00962
New m / old v	Reset m	-0.00315	-0.00592
New m / old v	Adaptive clip	-0.01104	-0.01041

One-time repairs and continuous clipping. The bounded block contains FULL, the old coherent pair, and the stress state, each with unmodified, one-time envelope projection, one-time m reset, or direct adaptive-update clipping (Table 10). Unlike projection/reset, clipping acts at *every* one of the 64 updates. Its fixed threshold is the L2 norm of FULL's initial bias-corrected moment ratio before the next gradient; each actual post-gradient adaptive displacement is clipped to that norm, with decoupled decay retained exactly.

Before intervention, stress violations occupy 2.619% and 2.722% of coordinates but account for 99.9985% and 99.9989% of the squared *stored-moment* adaptive ratio. These are pre-gradient diagnostics, not measurements of the actual first optimizer step. The unmodified curves reach NLL peaks of 192.77 and 263.67 on the evaluation grid and remain above FULL by 1.161 and 1.340 at update 64. Projection removes the large transient (Figure 14), yet m reset has lower signed mean excess in both seeds, while projection has smaller endpoint excess. Projection leaves FULL and the old coherent pair exactly unchanged in both seeds. Resetting m also lowers FULL's short-window signed mean. Projection selectively acts on envelope violations, while reset changes admissible and violating states alike.

Table 11: Mean excess NLL on the common 64-update grid at step 512, age 64, with fixed current weights and clock. The four corners use matched matrix and stress records.

Moment sources	Seed 1	Seed 2
m_t, v_t	+0.00000	+0.00000
m_s, v_s	+0.00724	+0.00888
m_s, v_t	+0.01181	+0.01141
m_t, v_s	+19.41069	+24.31780

At step 512, age 64, new m/old v yields 64-update mean excess NLL $19.41/24.32$ across the two seeds, versus $0.00724/0.00888$ for the old coherent pair on the same grid. One-time projection reduces it to $0.00744/0.00962$, and resetting m to $-0.00315/-0.00592$. Among envelope-admissible states, keeping v gives 512-update mean excess -0.00176, compared with $+0.02973$ after removing both moments (Table 24).

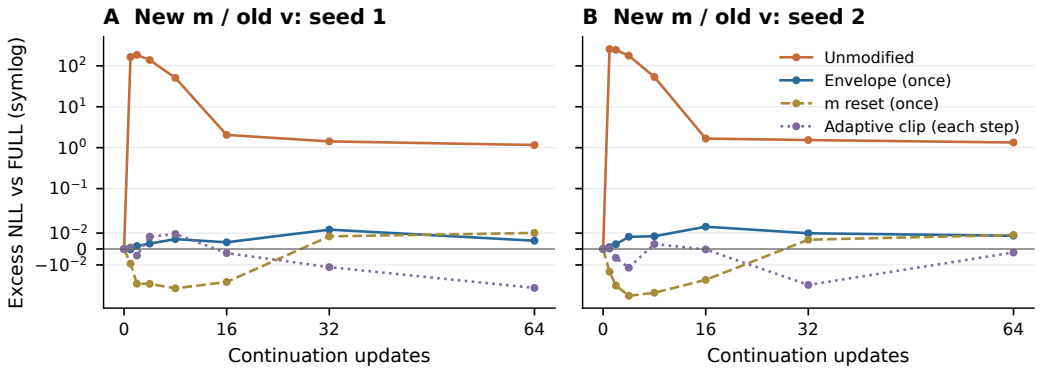

Figure 14: **Coverage-guided constructions remove a large source-mismatch transient.** Two initialization seeds at step 512, age 64. One-time envelope projection and one-time m reset both suppress the new-m/old-v response. The continuous update-clipping control acts at every step. Signed symmetric-log axes (linear threshold 0.02) resolve both the transient and near-FULL curves.

C.7 Recovery information: state choice and finite-range uncertainty

The coverage envelope describes coherent source states. We use it to compare how much of FULL's adaptive-update sequence can be recovered after one moment is omitted. The analysis isolates source information by supplying the same future gradients to all branches. Retaining v bounds the lost numerator history, normalized by its denominator, with the decaying factor ρ^k, where $\rho = \sqrt{q}$. This same rate controls the reset input under natural gradient feedback in Appendix C.8.

Model and recovery information. Fix a scalar source (m_0, v_0) at optimizer clock $t \geq 2$, with $v_0 > 0$. For $k = 1, \ldots, H$, let

$$m_k = \beta_1 m_{k-1} + (1-\beta_1) g_k, \quad v_k = \beta_2 v_{k-1} + (1-\beta_2) g_k^2,$$
$$u_k = -\eta b_k \frac{m_k}{\sqrt{v_k} + e_k}, \qquad b_k = \frac{\sqrt{1-\beta_2^{t+k}}}{1-\beta_1^{t+k}}, \quad e_k = \epsilon\sqrt{1-\beta_2^{t+k}}. \tag{37}$$

Here $0 < \beta_1, \beta_2 < 1$, $\rho = \beta_1/\sqrt{\beta_2} = \sqrt{q} < 1$, $\eta > 0$ is constant, and weight decay is zero. Write $a = 1-\beta_2$, $b_- = \sqrt{1-\beta_2^{t+1}}$, $b_+ = (1-\beta_1^{t+1})^{-1}$, $A = \eta b_+$ and $M = (1-\beta_1)/(1-\rho)$, so $b_- \leq b_k \leq b_+$. The constant C in Equation 8 is $(1-\beta_1)^2/[a(1-q)]$; its finite-history version is $C_t = C(1-q^t)$.

A causal recovery rule D knows the retained moment report I, source parameter, clock, hyperparameters, horizon and declared source class. At step k it observes $g_{1:k}$ and outputs a displacement d_k. Discarded history is unavailable, including through past-training replay; reports contain ordinary moment values. Define

$$L_H(D; s, g) = \sum_{k=1}^{H} |d_k - u_k(s, g)|, \qquad \mathfrak{R}_H^{\text{keep } j} = \inf_{D \text{ causal}} \sup_{\substack{s \in \mathcal{S}_{l,u} \\ g \in \mathbb{R}^H}} \mathbb{E}_D L_H(D; s, g), \quad I = j(s), \tag{38}$$

where $j = m$ or v, and $\mathcal{S}_{l,u} = \{(m, v) : l \leq v \leq u,\ m^2 \leq C_t v\}$. Internal randomness is independent of the hidden source. We also use the conditional risk $\mathcal{R}_H^*(m; [l,u], \epsilon)$ when m is known and only $v \in [l,u]$ is hidden. All gradients are finite real numbers; the worst-case sequence may depend on H. These risks measure recovery of FULL's updates. Natural-training NLL is evaluated in Section 6.

Reachable sources. The finite-history envelope is exact. For zero-initialized moments after t gradients $h_1, \ldots, h_t$, set $z_j = \sqrt{a\beta_2^{t-j}} h_j$ and $w_j = (1-\beta_1)\beta_1^{t-j}/\sqrt{a\beta_2^{t-j}}$. Then $v_0 = \|z\|^2$, $m_0 = w^\top z$ and $\|w\|^2 = C_t$. For $t \geq 2$, a sphere of radius $\sqrt{v_0}$ projects onto the whole interval

$[-\sqrt{C_t v_0}, \sqrt{C_t v_0}]$ along w. Thus every state in $\mathcal{S}_{l,u}$ is attainable by a finite history. At $\epsilon = 0$, one can choose a representative with $h_1 \neq 0$; for $t = 2$ the two representatives cannot both have first coordinate zero and the same projection, and for $t > 2$ the nonextremal slice is a sphere not contained in that hyperplane. Extremal representatives are parallel to w. The earlier parameter initialization can be chosen to give the same source parameter; the source class does not fix that earlier initialization.

The hidden-v witnesses below admit a stronger construction. Set $h_1 = B > 0$, $h_2 = -\beta_1 B$, and all later $h_j = 0$. Then $m_0 = 0$ and $v_0 = a(\beta_2 + \beta_1^2)\beta_2^{t-2}B^2$. At $\epsilon = 0$, the first bias-corrected adaptive displacement is independent of B, and all later ones are zero. Different B therefore realize any fixed $l, u > 0$ from the same earlier parameter initialization and with the same source parameter and clock. Their past gradient histories differ.

Exact recovery with v retained. Condition on an observed $v > 0$ and $|m_0| \leq \sqrt{C_t v}$. For any $\epsilon \geq 0$, the exact minimax risk is

$$K_H(v, \epsilon) = \sqrt{C_t v} \sum_{k=1}^{H} \frac{\eta b_k \beta_1^k}{\beta_2^{k/2}\sqrt{v} + e_k}. \tag{39}$$

REBUILD_M attains it. Its difference from FULL in the numerator is $-\beta_1^k m_0$, its second moment is identical, and $v_k \geq \beta_2^k v$, giving the upper bound. For the lower bound, take $m_0 = \pm\sqrt{C_t v}$ and $g_{1:H} = 0$. The two FULL update sequences are opposite and have distance $2K_H$; they give any recovery rule identical observations. The triangle inequality makes its error at least K_H in one world. Coupling a randomized rule with the same random seed and taking expectations proves the same lower bound. At $\epsilon = 0$,

$$\mathfrak{R}_H^{\text{keep } v} = \eta\sqrt{C_t} \sum_{k=1}^{H} b_k \rho^k = \Theta(1). \tag{40}$$

The exact value also holds for endpoint parameter error: the upper bound follows by summation, and the witness updates have constant sign.

The conditional radius when v is uncertain. With m_0 known, write $p_k = \beta_1^k m_0 + (1 - \beta_1)\sum_{j=1}^{k} \beta_1^{k-j} g_j$, $s_k = a\sum_{j=1}^{k} \beta_2^{k-j} g_j^2$ and $U_k(z) = -\eta b_k p_k/(\sqrt{\beta_2^k z + s_k} + e_k)$. The exact conditional radius is

$$\mathcal{R}_H^*(m_0; [l, u], \epsilon) = \frac{1}{2} \sup_g \sum_{k=1}^{H} |U_k(l) - U_k(u)|. \tag{41}$$

For each observed gradient prefix, $U_k(z)$ lies between its endpoint values. Their midpoint is a causal rule whose error is at most half the endpoint gap. The same hidden endpoint attains this half-gap at every step. Conversely, the two endpoint worlds are indistinguishable, so their sequence distance lower-bounds twice the worst error of any rule, for every fixed g. Taking the supremum proves the formula, also for randomized rules. This midpoint is a witness for the unrestricted causal class, not generally one Adam initialization; Equation 52 below analyzes the actual single-state implementation.

Tight joint dependence on horizon and range. Let $R = u/l > 1$ and define

$$J(R) = \int_1^\infty \left(S^{-1/2} - (S + R - 1)^{-1/2}\right)^2 dS = 4 \log\cosh\left(\frac{\log R}{4}\right). \tag{42}$$

For $m_0 = 0$ and $\epsilon = 0$,

$$\mathcal{R}_H^*(0; [l, u], 0) \asymp \min\{H, \sqrt{HJ(R)}\}. \tag{43}$$

Constants depend only on $\beta_1, \beta_2, t, \eta$, not on H, l, R. An explicit upper bound, valid also for $\epsilon \geq 0$, is

$$B_H(R) = \frac{A}{2} \min\left\{\sqrt{C}H, \frac{M}{\sqrt{a}}\sqrt{HJ(R)}\right\}. \tag{44}$$

To prove it, normalize $x_j = g_j/(\sqrt{l}\beta_2^{j/2})$, $S_k = 1 + a\sum_{j\le k} x_j^2$, $P_k = (1-\beta_1)\sum_{j\le k}\rho^{k-j}x_j$ and $h_R(S) = S^{-1/2} - (S+R-1)^{-1/2}$. At $\epsilon = 0$, the endpoint gap is $\eta b_k |P_k| h_R(S_k)$. Since S_k increases and h_R decreases,

$$\sum_k |P_k| h_R(S_k) \le M \sum_j |x_j| h_R(S_j) \le M\sqrt{H\sum_j x_j^2 h_R(S_j)^2} \le \frac{M}{\sqrt{a}}\sqrt{HJ(R)}. \tag{45}$$

The last inequality uses $ax_j^2 = S_j - S_{j-1}$ and the right-endpoint rectangle bound for the decreasing function h_R^2. Also $P_k^2 \le C(S_k - 1)$, bounding each same-sign endpoint gap by $A\sqrt{C}$. Equation 41 gives Equation 44. Positive ϵ only decreases the reciprocal-denominator gap. For $\delta = R - 1$, an antiderivative in Equation 42 is $\log S + \log(S+\delta) - 4\,\mathrm{arsinh}\,\sqrt{S/\delta}$. Its endpoints give the stated expression, with $J(R) \sim (\log R)^2/8$ near one and $J(R) \sim \log R$ at infinity.

Two explicit positive-gradient witnesses give the lower bound. For $1 < R \le 2$, choose $x_k = 1/\sqrt{aH}$. Then $S_k \le 2$, $P_k \ge (1-\beta_1)x_k$, and $h_R(S_k) \ge (R-1)/[2(R+1)^{3/2}]$, yielding

$$L_H^{\mathrm{small}}(R) = \frac{\eta b_-(1-\beta_1)(R-1)}{4\sqrt{a}(R+1)^{3/2}}\sqrt{H} \le \mathcal{R}_H^*(0;[l,u],0). \tag{46}$$

For $R \ge 2$, take $r = R^{1/H}$, $S_k = r^k$ and $x_k = \sqrt{(r-1)r^{k-1}/a}$. Here $S_k \le R$ and $h_R(S_k) \ge (1-\sqrt{2/3})/\sqrt{S_k}$, so

$$L_H^{\mathrm{large}}(R) = \frac{\eta b_-(1-\beta_1)(1-\sqrt{2/3})}{2\sqrt{a}} H\sqrt{1-R^{-1/H}} \le \mathcal{R}_H^*(0;[l,u],0). \tag{47}$$

On $1 < R \le 2$, $J(R)$ is uniformly comparable to $(R-1)^2$ and is less than one. On $R \ge 2$, use $1 - e^{-z} \ge (1-e^{-1})\min\{z,1\}$ and $J(R) \le \log R$. These facts prove the joint order in Equation 43. Both witnesses have same-sign update gaps, so the endpoint-parameter minimax risk has the same order.

Comparison on the same source class. For known nonzero m_0, splitting the inherited and new-gradient numerators adds at most

$$I_H(m_0, R) = \frac{\eta|m_0|}{2\sqrt{l}}(1 - R^{-1/2})\sum_{k=1}^{H} b_k\rho^k \tag{48}$$

to $B_H(R)$. This follows by maximizing the reciprocal gap at zero new second-moment mass. When $m_0^2 \le C_0 l$ for fixed C_0, this term is $O(R-1)$ near one and $O(1)$ for $R \ge 2$, and is absorbed by the joint order. Choosing the witnesses' signs to agree with m_0 preserves the lower bound.

For a report m_0 from $\mathcal{S}_{l,u}$, its feasible interval is $[l_m, u]$, where $l_m = \max\{l, m_0^2/C_t\}$. Then $m_0^2 \le C_t l_m$ and $u/l_m \le R$, giving a uniform upper bound. The $m_0 = 0$ subclass supplies the lower bound. Consequently

$$\mathfrak{R}_H^{\mathrm{keep}\ m} \asymp \min\{H, \sqrt{HJ(R)}\}, \qquad \mathfrak{R}_H^{\mathrm{keep}\ v} = \eta\sqrt{C_t}\sum_{k=1}^{H} b_k\rho^k. \tag{49}$$

This proves Proposition 1. If $l_m = u$, the conditional v is known exactly. At $R = 1$, keeping m therefore recovers FULL exactly; the comparison concerns nonzero source uncertainty and its horizon dependence, rather than a ranking at every finite horizon. In the lower witnesses, future gradients decay with the inherited second moment, maintaining relative denominator uncertainty despite decreasing absolute moment differences.

Positive ϵ and the finite window. Let $\epsilon > 0$, $m_0 = 0$ and $e_* = \epsilon b_-$. For every $0 \le T \le H$, with $B_0 = 0$,

$$\mathcal{R}_H^*(0;[l,u],\epsilon) \le B_T(R) + \frac{A\sqrt{C}(u-l)}{4e_*^2}\sum_{k=T+1}^{H}\beta_2^k. \tag{50}$$

Indeed $p_k^2 \le Cs_k$, and differentiating $U_k(z)$ gives $|\partial_z U_k(z)| \le A\sqrt{C}\beta_2^k/(2e_*^2)$. The mean-value theorem and Equation 41 bound the tail; Equation 44 bounds the first T terms. Taking $T = 0$ proves

a uniform-in-H bound at fixed l, u, ϵ. The bounded inherited term in Equation 48 extends this cutoff to known nonzero m_0 and the unified class. Conversely, if $\epsilon \le \sqrt{l\beta_2^H}$, the normalized witnesses have $E_k = e_k/(\sqrt{l}\beta_2^{k/2}) \le 1$ and $S_k \ge 1$. Their positive-ϵ reciprocal gap is at least one quarter of the zero-ϵ gap, because

$$\frac{\sqrt{S_k}\sqrt{S_k+R-1}}{(\sqrt{S_k}+E_k)(\sqrt{S_k+R-1}+E_k)} \ge \frac{1}{4}. \tag{51}$$

Thus Equations 46–47 retain their finite-window lower bounds up to this factor.

C.7.1 Relative precision and the actual recovery rule

Suppose $m_0^2 \le C_t v_0$ and a saved $\widehat{v} > 0$ satisfies $|\widehat{v} - v_0| \le \chi v_0$, $0 \le \chi < 1$. For a single Adam filter initialized at $(0, \widehat{v})$, with the original clock and common future gradients, the following bound holds for $\epsilon \ge 0$:

$$\begin{aligned}\sum_{k=1}^{H} |u_k^{\mathrm{R},\widehat{v}} - u_k^{\mathrm{FULL}}| \le A\sqrt{C_t}\frac{\rho(1-\rho^H)}{1-\rho} \\ + \frac{A\chi(1-\beta_1)}{2(1-\chi)^{3/2}(1-\rho)}\sqrt{\frac{H}{2a}}.\end{aligned} \tag{52}$$

The first term comes from omitting m_0, the second from approximating v_0. For the latter, normalize by the true v_0 as above and put $\delta = (\widehat{v} - v_0)/v_0$, $E_k = e_k/(\sqrt{v_0}\beta_2^{k/2})$. The two new-gradient denominators are $\sqrt{S_k} + E_k$ and $\sqrt{S_k + \delta} + E_k$. For $r_E(S) = (\sqrt{S} + E)^{-1}$, $|r'_E(S)| \le 1/(2S^{3/2})$ and every intervening point is at least $(1-\chi)S_k$. Hence the reciprocal gap is at most $\chi/[2(1-\chi)^{3/2}S_k^{3/2}]$. The convolution and energy argument now gives

$$\sum_k \frac{|P_k|}{S_k^{3/2}} \le M\sqrt{H\sum_j \frac{x_j^2}{S_j^3}} \le M\sqrt{\frac{H}{2a}}, \tag{53}$$

using $a\sum_j x_j^2/S_j^3 \le \int_1^\infty S^{-3}dS = 1/2$. Adding the inherited numerator bound proves Equation 52.

The relative-precision report corresponds to $l = \widehat{v}/(1+\chi)$, $u = \widehat{v}/(1-\chi)$ and $R_\chi = (1+\chi)/(1-\chi)$. At $\epsilon = 0$, with m omitted, its minimax risk is

$$\mathcal{R}_H^{\mathrm{relative}} \asymp 1 + \min\{H, \sqrt{HJ(R_\chi)}\}. \tag{54}$$

For the upper bound use the midpoint of the two zero-m endpoint filters and add Equation 40. The class contains both the fixed-v opposite-m pair and the zero-m interval-v pair. Their lower bounds have maximum at least half their sum, proving Equation 54. When $\chi \le \chi_{\max} < 1$, $J(R_\chi) \asymp \chi^2$ uniformly, giving $\Theta(1 + \chi\sqrt{H})$; the actual rule in Equation 52 attains this order.

This corollary concerns one-time moment encoding at the same source parameters, followed by ordinary arithmetic. Its relative-error condition is coordinatewise; zeroing tiny positive values need not satisfy $\chi < 1$. An exactly zero v_0 in the coherent class implies $m_0 = 0$ and is reproduced by exact zero encoding. Appendix D.4 evaluates complete BF16 and block-wise 8-bit checkpoints, including their parameter encoding, through matched NLL comparisons.

C.8 Loss Perturbation and Response to Complete First-Moment Reset

The information analysis identifies which history is lost. Here the question is how that loss changes validation NLL during natural continuation. The connection has three parts: coverage bounds the history-normalized reset input; joint parameter–moment dynamics propagate it with gradient feedback; validation sensitivity reads out the resulting displacement. Equation 68 expresses this connection as a signed input–response decomposition. An exact compensated state handles complete deletion of m without treating the reset itself as infinitesimal. Under the stated local conditions, complete deletion produces $O(\eta)$ parameter displacement and loss perturbation over $h \le T/\eta$, with an $O(\eta^2)$ remainder in the signed response expansion. Both branches use the same batch and randomness maps G_k evaluated at their own parameters. All statements in this subsection concern exact arithmetic.

Setup and exact compensation. Fix $0 < \beta_1 < \sqrt{\beta_2} < 1$, $\epsilon > 0$, the source clock t_0, and a source (θ_0, m_0, v_0) satisfying $m_{0,i}^2 \le C v_{0,i}$, with $C \ge C_\infty = (1-\beta_1)^2/[(1-\beta_2)(1-\beta_1^2/\beta_2)]$. Write $\rho = \beta_1/\sqrt{\beta_2}$, $a = 1-\beta_1$, $b = 1-\beta_2$, and

$$A_k = \eta_k \frac{\sqrt{1-\beta_2^{t_0+k}}}{1-\beta_1^{t_0+k}}, \qquad e_k = \epsilon\sqrt{1-\beta_2^{t_0+k}}. \tag{55}$$

Assume $0 \le \eta_k\lambda \le 1$. For fixed positive scales s_i, set $u = m/s$ and $r = \sqrt{v}/s$, with all vector quotients coordinatewise. Define

$$\begin{aligned} u^+ &= \beta_1 u + aG_k(\theta)/s, & r^+ &= \sqrt{\beta_2 r^2 + b(G_k(\theta)/s)^2}, \\ T_{k,\theta}(x) &= (1-\eta_k\lambda)\theta - A_k\frac{u^+ - \beta_1^k m_0/s}{r^+ + e_k/s}, & T_{k,y}(x) &= (u^+, r^+), \quad x = (\theta, y). \end{aligned} \tag{56}$$

The FULL map F_k omits the compensator $\beta_1^k m_0/s$. Start both lifted branches at $x_0 = (\theta_0, m_0/s, \sqrt{v_0}/s)$.

Lemma (complete-reset representation). The parameter path of $x_k^R = T_k(x_{k-1}^R)$ is exactly REBUILD_M. Its physical first moment is $su_k^R - \beta_1^k m_0$. Indeed, $\widetilde{m}_k = m_k^{R,\mathrm{phys}} + \beta_1^k m_0$ starts at m_0 and obeys the ordinary first-moment recurrence with $G_k(\theta_{k-1}^R)$. The second-moment recurrence and source clock are unchanged. Thus the parameter update is the ordinary lifted update plus $f_k = A_k\beta_1^k m_0/(\sqrt{v_k^R} + e_k)$. Since $v_k^R \ge \beta_2^k v_0$,

$$\max_i \sum_{k=1}^H |f_{k,i}| \le \overline{A}\kappa_0 \frac{\rho(1-\rho^H)}{1-\rho}, \qquad \kappa_0 = \max_{i:v_{0,i}>0} \frac{|m_{0,i}|}{\sqrt{v_{0,i}}} \le \sqrt{C}, \tag{57}$$

where $\overline{A} = \max_{k\le H} A_k$ and an empty maximum is zero. Zero-v_0 coordinates have zero inherited m_0. The physical reset may be order one, while its cumulative parameter input is order η when $\eta_k = \eta\ell_k$ and $0 \le \ell_k \le \overline{\ell}$.

First-order response along FULL. Using the source and FULL trajectory, define

$$b_k = T_k(x_{k-1}^F) - x_k^F, \qquad J_k = DT_k(x_{k-1}^F), \qquad z_0 = 0, \quad z_k = J_k z_{k-1} + b_k. \tag{58}$$

Here J_k differentiates the compensator's denominator as well as the ordinary update. This source- and reference-conditioned linear response uses neither the recovered trajectory nor its gradients. Joint-state signed propagation is also used for minibatch-gradient shocks (Liu & Li, 2026); here compensation treats a complete moment reset.

Uniform local conditions. Fix $T > 0$, the source, betas, scales, and a FULL-centered tube of radius $R_* > 0$ uniformly over the family $H \le T/\eta$. On this tube suppose

$$\|S^{-1}(G_k(\theta) - G_k(\theta_{k-1}^F))\|_\infty \le L_g\|\theta - \theta_{k-1}^F\|_\infty, \qquad \max_i \frac{s_i}{\sqrt{v_{k,i}^F} + e_k} \le D, \tag{59}$$

where $S = \mathrm{diag}(s_i)$. The maps T_k are C^2 throughout the tube, with positive square-root arguments. In block infinity norms require

$$\|J_{\theta\theta}\| \le 1 + \eta a_0, \quad \|J_{\theta y}\| \le \eta b_0, \quad \|J_{y\theta}\| \le c_0, \quad \|J_{yy}\| \le \sigma < 1, \tag{60}$$

and Taylor remainders

$$\|N_{k,\theta}(d)\|_\infty \le \eta M_\theta\|d\|_\infty^2, \qquad \|N_{k,y}(d)\|_\infty \le M_y\|d\|_\infty^2. \tag{61}$$

All constants are independent of η, k, H in this family. At nonnegative reference root states, J_{yy} is diagonal with entries β_1 and $\beta_2 r_{k-1,i}^F/r_{k,i}^F$, bounded by $\sigma = \max(\beta_1, \sqrt{\beta_2}) < 1$. The parameter row carries the factor $A_k = O(\eta)$; no contraction of parameter dynamics is required. For clipped gradients, G_k includes the clipping operation: the smooth conditions apply to tubes within a fixed clipping branch.

Small lifted displacement from a full reset. The envelope is invariant for arbitrary gradient sequences because $(\beta_1 m + a g)^2 \le (\beta_1^2 C/\beta_2 + a^2/b)(\beta_2 v + b g^2) \le C(\beta_2 v + b g^2)$. It therefore holds for the auxiliary filter driven by the recovered branch. If $E_h = \max_{j \le h} \|\theta_j^R - \theta_j^F\|_\infty$, the common initial filters and the reverse triangle inequality for weighted gradient histories give

$$|\delta \widetilde{m}_{k,i}| \le L_g s_i (1-\beta_1^k) E_{k-1}, \qquad |\delta \sqrt{v_{k,i}}| \le L_g s_i \sqrt{1-\beta_2^k}\, E_{k-1}. \tag{62}$$

For auxiliary and FULL values denoted by A, F, respectively, the identity

$$\frac{m_A}{r_A + e} - \frac{m_F}{r_F + e} = \frac{m_A - m_F}{r_F + e} + \frac{m_A}{r_A + e}\frac{r_F - r_A}{r_F + e}$$

with unscaled roots bounds the adaptive difference by $L_g D(1 + \sqrt{C}) E_{k-1}$. Let $\overline{\gamma}$ bound A_k/η_k, $Q = \overline{\ell}\overline{\gamma}\kappa_0 \rho/(1-\rho)$, and $K_0 = \overline{\ell}\overline{\gamma} L_g D(1+\sqrt{C})$. Unrolling the nonexpansive coordinatewise decay and using Equation 57 yields

$$E_h \le Q\eta + \eta K_0 \sum_{k \le h} E_{k-1} \le Q\eta e^{K_0 T} = D_\theta \eta, \qquad \|x_h^R - x_h^F\|_\infty \le D_x \eta, \tag{63}$$

where $D_x = \max(1, L_g) D_\theta$. For sufficiently small η with $D_x \eta < R_*$, a first-exit induction closes this same tube. This step derives small auxiliary-state displacement rather than assuming a small physical reset. The coordinatewise input sum is retained before taking its maximum, as required by Equation 57.

Uniform response theorem and proof. Under Equations 59–61, there exist $\eta_* > 0$ and constants independent of η, h such that for all $0 < \eta \le \eta_*$ and $0 \le h \le \lfloor T/\eta \rfloor$,

$$\|\theta_h^R - \theta_h^F - z_{h,\theta}\|_\infty \le C_* V_T \eta^2, \quad K = a_0 + \frac{b_0 c_0}{1-\sigma}, \quad C_* = D_x^2 \left(M_\theta + \frac{b_0 M_y}{1-\sigma} \right), \tag{64}$$

where $V_T = (e^{KT} - 1)/K$ for $K > 0$ and $V_T = T$ for $K = 0$. To prove it, subtract z_k from the exact Taylor recurrence and denote parameter and moment residual norms by p_k, q_k. Then

$$\begin{aligned} p_k &\le (1 + \eta a_0) p_{k-1} + \eta b_0 q_{k-1} + M_\theta D_x^2 \eta^3, \\ q_k &\le c_0 p_{k-1} + \sigma q_{k-1} + M_y D_x^2 \eta^2. \end{aligned} \tag{65}$$

With $P_k = \max_{j \le k} p_j$, geometric summation gives $q_{k-1} \le [c_0 P_{k-1} + M_y D_x^2 \eta^2]/(1-\sigma)$, hence $P_k \le (1 + \eta K) P_{k-1} + C_* \eta^3$. Summing from zero proves Equation 64; substituting back also bounds the moment residual by $[c_0 C_* V_T + M_y D_x^2]\eta^2/(1-\sigma)$. The fast-row remainder is order η^2, but its geometric forgetting and order-η feedback into parameters prevent accumulation into an order-η parameter error over $O(1/\eta)$ updates.

Loss-perturbation bound and signed response. If $\|\nabla L(\theta_h^F)\|_1 \le G_L$ and $|w^\top \nabla^2 L\, w| \le M_L \|w\|_\infty^2$ throughout the endpoint tube, Equation 63 and Taylor's theorem immediately give the following corollary:

$$\sup_{0 \le h \le \lfloor T/\eta \rfloor} |L(\theta_h^R) - L(\theta_h^F)| \le G_L D_\theta \eta + \tfrac{1}{2} M_L D_\theta^2 \eta^2. \tag{66}$$

Thus an order-one physical moment deletion has a controlled local loss effect on a growing window. The essential step is obtaining $D_\theta \eta$ from the summable, coverage-controlled reset input and closing the tube, rather than assuming a small state intervention. Combining Taylor's theorem with Equation 64 further yields

$$\left| L(\theta_h^R) - L(\theta_h^F) - \nabla L(\theta_h^F)^\top z_{h,\theta} \right| \le \left(G_L C_* V_T + \tfrac{1}{2} M_L D_\theta^2 \right) \eta^2. \tag{67}$$

The signed linear term is generally order η. Its sign determines that of the actual response whenever its magnitude exceeds the displayed remainder. All bounds concern perturbation relative to FULL, with source- and geometry-dependent constants; they do not establish optimization convergence or uniformly accurate finite-step forecasts for arbitrary training runs.

History input, propagation, and loss sensitivity. Unrolling Equation 58 makes the connection explicit. Let $\Phi_{h,k} = J_h \cdots J_{k+1}$, $\Phi_{h,h} = I$, with Π_θ the parameter projection and ι_θ its embedding into joint state. The reset input has only a parameter component, so

$$\mathcal{K}_{h,k} = \nabla L(\theta_h^F)^\top \Pi_\theta \Phi_{h,k} \iota_\theta, \qquad \Delta L_h = \sum_{k=1}^{h} \mathcal{K}_{h,k} b_{k,\theta} + r_h, \quad |r_h| \le C_L \eta^2, \tag{68}$$

where C_L is the coefficient in Equation 67. Here $b_{k,\theta} = A_k \beta_1^k m_0 / (\sqrt{v_k^F} + e_k)$. As $v_k^F \ge \beta_2^k v_0$, the input bound applies to this reference input too. Consequently, for $h \ge 1$,

$$|\Delta L_h| \le \Gamma_h B_h + C_L \eta^2, \qquad \Gamma_h = \sum_i \max_{1 \le k \le h} |\mathcal{K}_{h,k,i}|, \quad B_h = \overline{A} \kappa_0 \frac{\rho(1-\rho^h)}{1-\rho}. \tag{69}$$

Indeed, bound each coordinate's input sum before summing its maximal gain. Coverage determines B_h; the signed kernel $\mathcal{K}_{h,k}$ retains direction, gradient feedback, and validation sensitivity. This separation connects the structural bound to the directional and persistence experiments.

Finite-step nonlinear readout. For finite step sizes one can instead evaluate $\widehat{\Delta} L_h = L(\theta_h^F + z_{h,\theta}) - L(\theta_h^F)$, retaining validation-loss curvature exactly, and enclose the remaining parameter approximation error as follows.

Choose FULL-centered boxes of radii R_k and whole-box Hessian bounds $|\partial_a \partial_b T_{k,i}| \le M_{k,iab}$. Set $w_0 = 0$, $\Phi_{h,j}$ as above, and recursively

$$d_j = |z_{j-1}| + w_{j-1}, \qquad n_{j,i} = \tfrac{1}{2} \sum_{a,b} M_{j,iab} d_{j,a} d_{j,b}, \qquad w_h = \sum_{j=1}^{h} |\Phi_{h,j}| n_j. \tag{70}$$

If $|z_k| + w_k \le R_k$ for all $k \le H$, then $|x_h^R - x_h^F - z_h| \le w_h$. Indeed the exact residual is $\sum_j \Phi_{h,j} N_j$; induction verifies each box before bounding N_j. Absolute values are taken after the signed matrix product, while unknown remainders at different times are still summed in worst case. This uses standard variational enclosure principles (van der Hoeven, 2018). For approximating parameters $\theta^P = \theta^F + z_\theta$, logit derivative bounds U_{ji} on $\theta^P \pm w_\theta$ imply logit radii $e_j^{\text{out}} = \sum_i U_{ji} w_{\theta,i}$. For binary NLL with label fraction y_j, a response interval is

$$\widehat{\Delta} L_h \ \pm \ \frac{1}{N} \sum_j e_j^{\text{out}} \max_{t \text{ in logit interval}} |\operatorname{sigmoid}(t) - y_j|. \tag{71}$$

The role of coverage. For fixed positive ϵ, a simpler input budget is $\overline{A} \|m_0\|_\infty \beta_1 / [e_*(1-\beta_1)]$, where $e_* = \epsilon \sqrt{1 - \beta_2^{t_0+1}}$. Both this budget and B_h are valid, so their minimum can be used in Equation 69. Their scale dependence differs: under $(m_0, v_0) \mapsto (c m_0, c^2 v_0)$, $c > 0$, κ_0 and B_h are unchanged, while the epsilon-only budget grows as c/ϵ. Coverage therefore supplies a history-normalized input budget. The propagation gain and local geometric constants retain their own reference dependence. The argument specializes signed-response analysis and fast–slow sensitivity estimates (Liu & Li, 2026; Cothren et al., 2026) to complete moment deletion. Small-learning-rate momentum comparisons over $O(1/\eta)$ windows have also been studied for SGD (Wang et al., 2024).

Nonempty uniform conditions and remainder order. A scalar logistic family supplies a concrete example. Take betas $(.9, .95)$, $\lambda = .1$, $t_0 = 128$, $|\theta_0| < .35$, $v_0 > .06$, $\eta \le .004$, $T = .512$, and alternating future label fractions $.4, .6$. The envelope with $C = 19/14$ bounds the FULL path by $|\theta_k| < .95$. Adjacent gradients then differ by more than .1987, so one has magnitude above .09935 and $v_k \ge .05 \cdot .95 \cdot (.09935)^2 > .000468$ for $k \ge 2$; the first step has $v_1 \ge .95 v_0$. This supplies a compact reference set with a fixed smooth neighborhood and uniform constants. At sufficiently small η, Equation 63 closes that neighborhood. For a nonzero coherent m_0, the first reset step has $\delta\theta_1 = z_{1,\theta} = \eta c \ne 0$. Logistic validation curvature is positive, so its linear NLL readout has remainder $\frac{1}{2} L''(\theta_0) c^2 \eta^2 + O(\eta^3)$. Thus the uniform linear-readout remainder cannot generally be $o(\eta^2)$.

Numerical checks of the response construction. Scalar logistic and three-parameter tanh-network checks cover 14 finite-window configurations; eight additional scalar runs hold $T = .512$ fixed as η decreases from .004 to .0005 and H grows from 128 to 1,024. The two source-wise log–log slopes of maximum parameter approximation error are 1.921 and 1.940. Finite-step enclosures use whole-box interval derivatives; these floating-point evaluations are numerical checks of the analytic construction, rather than formally rounded execution certificates. The distinction between the two NLL readouts is visible in a scalar run with $\eta = .001, H = 1024$: for validation label fraction .5, maximum actual response is 1.5815×10^{-6}, linear-readout error is 1.7708×10^{-6}, and nonlinear-readout error is 2.7689×10^{-12}. For label fraction .3 on the same paths these are 7.5431×10^{-4}, 1.7717×10^{-6}, and 1.3120×10^{-8}, respectively. The finite-step interval uses the nonlinear readout and its residual radius, not the linear term alone.

Componentwise form of the response. The first-order response in Equation 58 has the following componentwise form, without a dense joint-state Jacobian. In unscaled $(\theta, m, r = \sqrt{v})$ coordinates, write $d_k = r_k^F + e_k$, $c_k = \beta_1^k m_0$, and $\dot{g}_k = DG_k(\theta_{k-1}^F) z_{k-1,\theta}$. Then

$$
\begin{aligned}
z_{k,m} &= \beta_1 z_{k-1,m} + (1-\beta_1)\dot{g}_k, \\
z_{k,r} &= \frac{\beta_2 r_{k-1}^F \odot z_{k-1,r} + (1-\beta_2) g_k^F \odot \dot{g}_k}{r_k^F}, \\
z_{k,\theta} &= (1-\eta_k\lambda) z_{k-1,\theta} - A_k \left(\frac{z_{k,m}}{d_k} - \frac{(m_k^F - c_k) \odot z_{k,r}}{d_k^2} \right) + A_k \frac{c_k}{d_k}.
\end{aligned} \tag{72}
$$

There is one gradient-map Jacobian–vector product per update, a training-loss Hessian–vector product when G_k is an ordinary gradient. Computing this linear response and bounding derivatives over the entire tube are separate tasks.

D Recovery Applications

D.1 Direct Recovery under Native Training Conditions

Five-policy comparison at 1B. We use OLMo-2-0425-1B ingredient 1 at stage-2 step 5,000, with no adaptation prefix or additional warmup. All five policies retain the original effective batch of 512×4096 tokens, the original learning-rate schedule (multiplier 1.0), and the saved training-data cursor. The next batches follow the original file order, shuffle, and repetition filtering. FULL, REBUILD_M, REBUILD_V, V_FROM_M, and REBUILD_MV use the state definitions in Section 6, retaining the native moments selected by each policy and constructing omitted moments once on recovery. The inherited bias-correction clock, CE+z objective, and gradient clipping are shared.

Each policy undergoes a real process kill after checkpoint persistence, then disk-only recovery for 128 updates, processing 268.4 million tokens per policy. Gradient accumulation on a single GPU reproduces the original effective batch, retaining microbatch size 4 and the original rank-based sample grouping. Training uses FP32 master parameters and moments with BF16 computation.

The primary endpoint is validation NLL at update 128 on the 11 published OLMo PPL sources listed in Table 13. We take the first 32 sequences of length 4096 per source (352 total), with equal source weights. The first four sequences per source form a fixed 44-sequence subset evaluated throughout recovery. FineWeb-512 supplies an additional validation measurement; 64 unused sequences from the subsequent native training batch form a separate training-distribution probe.

We compute paired 95% percentile bootstrap intervals with 10,000 draws (seed 20260920). For the primary set, sequence pairs are resampled within each source and the 11 source means receive equal weight. FineWeb and the native-data probe resample sequence pairs within their respective sets. Table 13 further breaks down the REBUILD_M–FULL primary endpoint by source.

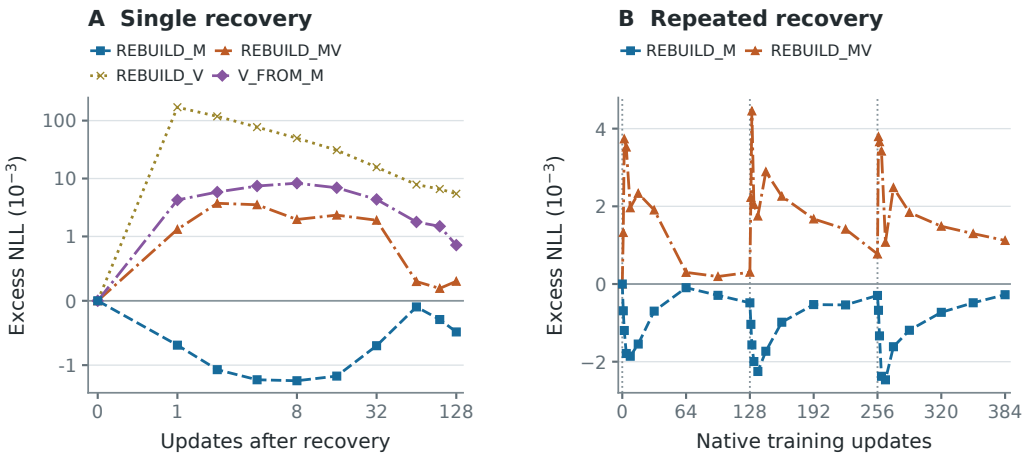

Figure 15: **Native 1B recovery differences relative to FULL.** (A) Single recovery: each reduced-state policy minus FULL, with a symmetric-log vertical axis and linear threshold 1 in the displayed 10^{-3} units. (B) Repeated recovery: REBUILD_M and REBUILD_MV minus FULL on a linear vertical axis; dotted lines mark faults at updates 0/128/256. Both panels use the same fixed 44 validation sequences, with zero denoting FULL at each update. Absolute trajectories, including FULL, appear in Figure 5; the 352-sequence primary estimates and intervals are in Tables 12 and 14.

Figure 5A shows all five absolute-NLL trajectories, including FULL, with an inset enlarging the near-FULL region. The two main-text panels use separate linear vertical ranges. In the single-recovery panels (Figures 5A and 15A), the horizontal axis is linear through update 1 and logarithmic thereafter; the repeated-recovery panels use linear update axes.

Table 12: Five-policy recovery under native training conditions at update 128. NLL is in nats/token; differences from FULL and paired 95% intervals are in 10^{-3} nats/token (lower is better). Payloads are 17.819 GB for FULL, 11.879 GB for REBUILD_M, REBUILD_V and V_FROM_M, and 5.940 GB for REBUILD_MV.

Evaluation	Policy	NLL	$\Delta \times 10^3$	95% interval
Primary (352)	FULL	2.472503	—	—
	REBUILD_M	2.472203	-0.300	$[-0.479, -0.098]$
	REBUILD_V	2.478181	$+5.679$	$[+5.172, +6.303]$
	V_FROM_M	2.473652	$+1.149$	$[+0.940, +1.382]$
	REBUILD_MV	2.473705	$+1.202$	$[+0.611, +2.206]$
FineWeb (512)	FULL	2.748819	—	—
	REBUILD_M	2.748798	-0.021	$[-0.219, +0.182]$
	REBUILD_V	2.753636	$+4.817$	$[+4.405, +5.237]$
	V_FROM_M	2.749762	$+0.942$	$[+0.753, +1.130]$
	REBUILD_MV	2.749609	$+0.790$	$[+0.482, +1.159]$
Native probe (64)	FULL	1.856089	—	—
	REBUILD_M	1.855869	-0.219	$[-0.402, -0.033]$
	REBUILD_V	1.859780	$+3.691$	$[+3.204, +4.195]$
	V_FROM_M	1.856883	$+0.795$	$[+0.580, +0.991]$
	REBUILD_MV	1.856340	$+0.251$	$[+0.027, +0.483]$

Retaining v preserves recovery quality under these native training conditions. REBUILD_M reduces payload from 17.819 to 11.879 GB (33.3%) and has a primary endpoint difference of -0.000300 nats/token from FULL. Its endpoint NLL is also lower than FULL on FineWeb and the native-data probe (Table 12). At the same payload, zeroing v yields $+0.005679$ primary excess NLL; constructing $v = m^2/C$ reduces this to $+0.001149$, removing 79.8% of the excess. Removing both moments gives $+0.001202$, while saving two thirds of the payload. Thus envelope construction repairs much of the zero-v damage, and retaining the actual second-moment history further improves recovery.

The early response shows the same distinction (Figure 15A). At the first update on the fixed 44-sequence set, constructing v reduces the excess NLL from 0.170240 to 0.004265 (97.5%). REBUILD_M is below FULL at all nine post-recovery measurements, with differences from -0.001859 to -0.000095 nats/token. On the full primary set, FULL's mean NLL changes from 2.471676 to 2.472503; on the native training-distribution probe, it decreases from 1.856750 to 1.856089.

Table 13: Per-source primary validation at update 128 (32 sequences per source). Differences and paired 95% intervals are in 10^{-3} nats/token.

Source	FULL	REBUILD_M	$\Delta \times 10^3$	95% interval
c4_en	2.80365	2.80346	-0.196	$[-0.659, +0.181]$
dolma_books	2.63394	2.63353	-0.410	$[-0.841, +0.017]$
dolma_common-crawl	2.99163	2.99134	-0.288	$[-0.616, +0.037]$
dolma_pes2o	2.11511	2.11498	-0.132	$[-0.383, +0.105]$
dolma_reddit	3.07996	3.08121	$+1.249$	$[-0.068, +3.099]$
dolma_stack	1.33212	1.33203	-0.096	$[-0.481, +0.295]$
dolma_wiki	2.19223	2.19205	-0.180	$[-0.385, +0.023]$
ice	2.72533	2.72439	-0.942	$[-1.371, -0.520]$
m2d2_s2orc	2.87628	2.87464	-1.643	$[-1.991, -1.326]$
pile	2.21884	2.21803	-0.807	$[-1.628, -0.179]$
wikitext_103	2.22843	2.22857	$+0.141$	$[-0.210, +0.504]$

Scaling native recovery to 7B. We compare FULL and REBUILD_M at the OLMo-2-7B stage-1 step-500,000 checkpoint, keeping the original 1024×4096-token effective batch, token-based learning-rate schedule (multiplier 1.0), and saved training cursor. No adaptation prefix or additional warmup is used. Gradient accumulation on two GPUs preserves the original two-sequence micro-batches: the sample groups of the original 512 ranks are assigned across the two execution ranks. The inherited optimizer clock, repetition filtering, CE+z objective, clipping, and FP32 states/BF16

computation are unchanged. Each policy undergoes a real process kill and disk-only recovery, followed by 128 updates (536.9 million tokens).

The primary endpoint uses the fixed 512-sequence OLMo held-out validation set, with length 4096 per sequence. We resample sequence pairs with 20,000 percentile-bootstrap draws (endpoint seed 20261050). The fixed first 64 sequences supply the entire plotted trajectory; the full 512-sequence evaluations at updates 0, 64, and 128 are subsampled to those same sequence indices for the figure.

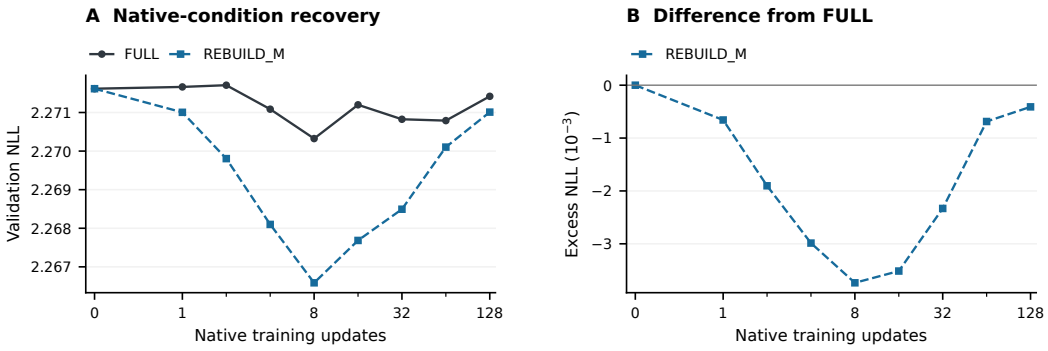

Figure 16: **Native 7B recovery.** Original batch, learning-rate schedule, and data cursor, with 128 updates after recovery. Both panels use the same fixed 64 sequences: absolute NLL (left) and REBUILD_M minus FULL (right). Horizontal axes are linear through update 1 and logarithmic thereafter. The full 512-sequence endpoint and its paired interval appear in Table 1.

On the full primary set, FULL changes from 2.272919 to 2.272864 NLL, while REBUILD_M ends at 2.272495. The paired endpoint difference is -0.000369 nats/token, with 95% interval $[-0.000611, -0.000138]$. Payload falls from 87.584 to 58.389 GB. Thus the same state choice retains near-FULL recovery quality at 7B under the original training batch, schedule, and data progression.

D.2 Repeated Recovery under Native Training Conditions

To compare retained history across repeated recoveries, we continue the native 1B FULL, REBUILD_M, and REBUILD_MV branches from Appendix D.1 to 384 updates. Faults occur at updates 0, 128, and 256 in every branch. Each event writes the policy's payload, terminates the process, and reloads from disk; REBUILD_M zeroes m, REBUILD_MV zeroes both moments, and FULL retains both. The first 128 updates are the same trajectories reported above. Thereafter each branch continues from its own parameters and history, retaining the original batch, learning-rate schedule, data order, filtering, and clocks, with no added warmup. The full window processes 805.3 million tokens per branch.

We retain the 352-sequence primary evaluation (32 per published source), with the fixed 44-sequence subset for recovery curves. For each fault, the early statistic is the trapezoidal mean of excess NLL at relative updates 0/1/2/4/8/16 over the first 16 updates. It measures the evolving branch difference, including any difference already present at the fault. Primary and early intervals use paired within-source resampling, equal source weights, 10,000 draws, and seed 20260920 for each contrast, as in the single-recovery comparison.

Figure 5B in the main text shows the absolute recovery curves; Figure 15B shows their differences from FULL. Table 14 reports the primary 352-sequence endpoints and early-response intervals.

Table 14: Repeated native 1B recovery. Differences from FULL and paired 95% intervals are in 10^{-3} nats/token. Endpoints use 352 sequences; early responses integrate the fixed 44-sequence difference over 0–16 updates after each fault and divide by 16.

Measurement	Policy	ΔNLL	95% interval
Endpoint 128	REBUILD_M	-0.300	$[-0.479, -0.098]$
Endpoint 128	REBUILD_MV	$+1.202$	$[+0.611, +2.206]$
Endpoint 256	REBUILD_M	-0.515	$[-0.646, -0.381]$
Endpoint 256	REBUILD_MV	$+0.994$	$[+0.826, +1.168]$
Endpoint 384	REBUILD_M	-0.427	$[-0.582, -0.272]$
Endpoint 384	REBUILD_MV	$+1.176$	$[+1.005, +1.361]$
Early, fault 0	REBUILD_M	-1.574	$[-1.812, -1.346]$
Early, fault 0	REBUILD_MV	$+2.414$	$[+2.043, +2.773]$
Early, fault 128	REBUILD_M	-1.878	$[-2.262, -1.494]$
Early, fault 128	REBUILD_MV	$+2.330$	$[+1.895, +2.769]$
Early, fault 256	REBUILD_M	-1.951	$[-2.346, -1.551]$
Early, fault 256	REBUILD_MV	$+2.269$	$[+1.804, +2.741]$

REBUILD_M remains close to FULL throughout the repeated-recovery comparison. Its primary differences at updates 128/256/384 are $-0.000300/-0.000515/-0.000427$ nats/token, compared with $+0.001202/+0.000994/+0.001176$ for REBUILD_MV. FULL's primary NLL falls from 2.472503 at update 128 to 2.471318 at update 384. The early comparison has the same ordering at all three faults: REBUILD_M's mean excess is between -0.001951 and -0.001574, whereas REBUILD_MV's is between $+0.002269$ and $+0.002414$. Retaining v therefore improves both early recovery and the measured endpoints relative to storing parameters alone, at payloads of 11.879 versus 5.940 GB (FULL: 17.819 GB).

Table 15: Additional 1B evaluations at update 384. NLL is in nats/token; differences and paired 95% intervals are in 10^{-3} nats/token.

Set	Policy	NLL	ΔNLL	95% interval
FineWeb (512)	FULL	2.748497	0	—
	REBUILD_M	2.747965	-0.532	$[-0.740, -0.330]$
	REBUILD_MV	2.749199	$+0.702$	$[+0.381, +0.989]$
Native probe (64)	FULL	1.719742	0	—
	REBUILD_M	1.719555	-0.187	$[-0.536, +0.199]$
	REBUILD_MV	1.720744	$+1.002$	$[+0.530, +1.531]$

Additional endpoint evaluations use the fixed FineWeb-512 set and a new 64-sequence native-data probe beyond the 384-update training window. The latter replaces the single-recovery probe, whose samples enter the extended training stream. These intervals resample paired sequences within each set (10,000 draws, seed 20260920). Table 15 gives both estimates and intervals. The matched fault positions here isolate history choice; the Pythia experiments in Appendix D.5 instead use smaller payloads to increase checkpoint frequency at matched planned write budgets.

D.3 Pretrained-State Recovery: Configuration and Results

The source-wise recovery curves are shown in Figure 17.

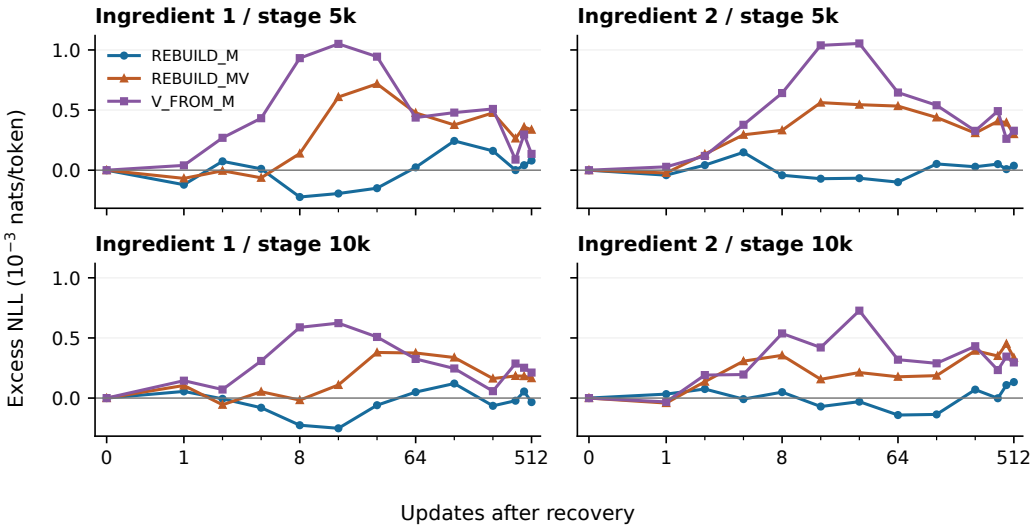

Figure 17: **Coverage-guided recovery and the value of retained history.** Fixed-64-document curves at 1B stages 5k/10k: keeping v (REBUILD_M), constructing it from m (V_FROM_M), or removing both moments (REBUILD_MV). Table 2 also includes REBUILD_V.

Source states and continuation settings. The 1B sources are OLMo-2-0425-1B ingredient 1/2 checkpoints (Allen Institute for AI, 2025) at stage-2 steps 5k and 10k, each with 1,484,916,736 parameters. The ingredients share earlier training ancestry. A fresh 256-update common FineWeb continuation (Penedo et al., 2024) produces each recovery anchor. We use effective batch 4, sequence length 4096, and the native scheduler multiplied by 0.1 during both the prefix and recovery. The 7B sources are unaveraged OLMo-2-1124-7B stage-1 checkpoints (Allen Institute for AI, 2024) at 500k and 800k from one training history. They enter recovery directly, using effective batch 16, sequence length 4096, and the original token-based scheduler multiplied by 0.1. The batch and schedule were fixed from FULL-only continuation diagnostics before the strategy comparisons and are shared by all policies without extra warmup. Appendix D.8 reports these diagnostics and recovery comparisons with the original schedule.

Computation and clocks. Parameters and optimizer moments use FP32 with BF16-autocast computation, math SDPA attention, and whole-layer activation recomputation. The 1B runtime is single-device; 7B uses two-rank fully sharded data parallelism. The original optimizer groups, betas, epsilon, weight decay, clipping, and bias-correction clocks are inherited. The 1B continuation objective is shifted cross-entropy. The 7B runtime retains the native cross-entropy plus z-loss (coefficient 10^{-5}), instance filtering, and normalization. At 7B recovery update j, the trainer step is $s+j$, the optimizer step is $s+1+j$, and the token clock is $s \cdot 1024 \cdot 4096 + j \cdot 16 \cdot 4096$. The original one-step trainer/optimizer offset is preserved.

1B data and evaluation. The public FineWeb dataset shard uses revision `9bb295ddab0e05d785b879661af7260fed5140fc`. Canonical-document hash buckets separate training and validation; fixed salted ordering selects 512 validation documents, each represented by a 1024-token window with the official Dolma2 tokenizer (523,776 prediction targets). The training stream contains 3,072 nonoverlapping 4096-token blocks: 1,024 for the prefix and 2,048 for recovery, with order seeds 2026091201/2026091202 for the two ingredients. Training and validation are disjoint at canonical-document level. Every policy at a source uses the same data order, RNG initialization, and continuation clock.

7B data and evaluation. A fixed seed (2026091507) samples 8,192 complete 4096-token training sequences without replacement from the configured native data recipe, retaining repeated source occurrences in that recipe. Native instance filtering remains active. This produces a fixed continuation stream drawn from the original recipe. Validation uses 512 packed 4096-token sequences, equally

stratified as closely as integer counts allow over eleven official held-out sources, in a fixed round-robin order. All strategies and both stages share this stream and evaluator. The 1B and 7B absolute NLLs use their respective evaluation sets.

The published OLMo PPL files use the `v3_small_dolma2-tokenizer` release, with file paths `SOURCE/val/part-0-00000.npy` under that release directory. The eleven source identifiers are listed in Table 13. The direct native-condition 1B experiment in Appendix D.1 uses the same release and source set, with its own balanced sequence selection.

State policies and real faults. The four 1B anchors each compare FULL, REBUILD_M, REBUILD_MV, REBUILD_V, and V_FROM_M, plus the matched BF16 pair described below (28 branches). The two 7B anchors each compare the same five core policies (ten branches). FULL serializes (θ, m, v); REBUILD_M serializes (θ, v) and sets $m = 0$; REBUILD_MV saves only θ and resets both moments; REBUILD_V saves (θ, m) and sets $v = 0$; V_FROM_M saves (θ, m) and reconstructs raw $v_i = m_i^2/C$ using the inherited parameter-group betas. Bias correction then uses the restored clock. After saving, one unsaved update precedes an actual SIGKILL; fresh workers restore only the declared disk payload and metadata and perform all 512 recovery updates. FULL roundtrip and per-policy next-update checks precede the comparisons. Saved RNG, data position, scheduler, and clocks are restored, and batches and learning rates match across paired policies.

Payload and evaluation metrics. The main precision uses 12 bytes per parameter for FULL, 8 for either one-moment strategy, and 4 for REBUILD_MV, plus serialization metadata. Actual 1B payloads are 17.819, 11.879, and 5.940 GB, respectively; the corresponding 7B payloads are 87.584, 58.389, and 29.195 GB. These are serialized checkpoint payloads, not training-memory measurements. Full-set evaluations occur at updates 0, 128, 256, 384, 448, and 512; the same first 64 evaluation items are additionally evaluated at 1, 2, 4, 8, 16, 32, and 64. The late statistic is the per-item paired difference weighted $0.25/0.50/0.25$ at 384/448/512. Intervals use 2,000 paired bootstrap resamples: document resampling for 1B (seed 2026091701), and packed-sequence resampling within source strata for 7B (seed 2026091508). They quantify evaluation-item variation at each source state. Tables 16–18 provide the complete 1B contrasts and absolute FULL trajectories; Table 19 provides the complete five-policy 7B comparison.

At 7B, zeroing v while retaining m produces early excess NLL up to $0.002234/0.001427$ across the sampled updates 1–64 at 500k/800k. Constructing $v = m^2/C$ suppresses this transient and reduces the late excess by $87.8\%/93.3\%$ relative to zero initialization. REBUILD_MV gives late excess $0.000133/0.000190$ with two-thirds smaller payloads; retaining v brings these differences to $0.000044/-0.000087$. The comparison connects envelope-guided construction, retained history, and the storage–quality trade-off at the larger scale.

Table 16: 1B late excess NLL and paired-document bootstrap intervals (10^{-3} nats/token).

Source	Policy	Δ	95% CI
I1 / 5k	REBUILD_M	+0.036	[-0.029, +0.103]
I1 / 5k	REBUILD_MV	+0.246	[+0.173, +0.318]
I1 / 5k	REBUILD_V	+22.106	[+20.394, +23.916]
I1 / 5k	V_FROM_M	+0.180	[+0.104, +0.252]
I2 / 5k	REBUILD_M	-0.013	[-0.069, +0.041]
I2 / 5k	REBUILD_MV	+0.256	[+0.179, +0.329]
I2 / 5k	REBUILD_V	+37.503	[+34.825, +40.584]
I2 / 5k	V_FROM_M	+0.273	[+0.208, +0.344]
I1 / 10k	REBUILD_M	-0.005	[-0.065, +0.051]
I1 / 10k	REBUILD_MV	+0.174	[+0.111, +0.236]
I1 / 10k	REBUILD_V	+13.059	[+11.971, +14.288]
I1 / 10k	V_FROM_M	+0.111	[+0.045, +0.178]
I2 / 10k	REBUILD_M	+0.095	[+0.047, +0.144]
I2 / 10k	REBUILD_MV	+0.232	[+0.171, +0.292]
I2 / 10k	REBUILD_V	+14.927	[+13.424, +16.839]
I2 / 10k	V_FROM_M	+0.241	[+0.177, +0.304]

Table 17: 1B endpoint excess NLL and paired-document bootstrap intervals (10^{-3} nats/token).

Source	Policy	Δ	95% CI
I1 / 5k	REBUILD_M	+0.010	[-0.070, +0.088]
I1 / 5k	REBUILD_MV	+0.266	[+0.179, +0.358]
I1 / 5k	REBUILD_V	+20.739	[+19.148, +22.473]
I1 / 5k	V_FROM_M	+0.167	[+0.079, +0.259]
I2 / 5k	REBUILD_M	+0.008	[-0.068, +0.084]
I2 / 5k	REBUILD_MV	+0.187	[+0.095, +0.274]
I2 / 5k	REBUILD_V	+34.903	[+32.377, +37.909]
I2 / 5k	V_FROM_M	+0.286	[+0.203, +0.372]
I1 / 10k	REBUILD_M	+0.030	[-0.050, +0.110]
I1 / 10k	REBUILD_MV	+0.186	[+0.106, +0.266]
I1 / 10k	REBUILD_V	+12.250	[+11.185, +13.432]
I1 / 10k	V_FROM_M	+0.114	[+0.030, +0.203]
I2 / 10k	REBUILD_M	+0.102	[+0.032, +0.172]
I2 / 10k	REBUILD_MV	+0.210	[+0.133, +0.291]
I2 / 10k	REBUILD_V	+14.077	[+12.642, +15.892]
I2 / 10k	V_FROM_M	+0.213	[+0.131, +0.298]

Table 18: Absolute FULL validation NLL at all full-set evaluations (nats/token). Each model uses its own fixed evaluation set.

Source	0	128	256	384	448	512
1B / I1 / 5k	2.733005	2.732096	2.730942	2.730730	2.730349	2.730230
1B / I2 / 5k	2.733402	2.732477	2.731192	2.730358	2.730286	2.729888
1B / I1 / 10k	2.732680	2.731840	2.730816	2.730371	2.730006	2.729727
1B / I2 / 10k	2.733326	2.732468	2.731077	2.730137	2.729999	2.729761
7B / 500k	2.272912	2.271874	2.271070	2.271980	2.271380	2.271192
7B / 800k	2.232247	2.232166	2.231572	2.231799	2.231845	2.231433

Table 19: 7B paired differences from FULL (10^{-3} nats/token), with source-stratified packed-sequence bootstrap intervals.

Stage	Policy	Late	95% CI	Step 512	95% CI
500k	REBUILD_M	+0.044	[-0.017, +0.107]	+0.001	[-0.092, +0.088]
500k	REBUILD_MV	+0.133	[+0.062, +0.203]	+0.020	[-0.065, +0.103]
500k	REBUILD_V	+0.827	[+0.652, +1.036]	+0.632	[+0.456, +0.832]
500k	V_FROM_M	+0.101	[+0.043, +0.159]	-0.005	[-0.081, +0.067]
800k	REBUILD_M	-0.087	[-0.143, -0.035]	+0.108	[+0.033, +0.176]
800k	REBUILD_MV	+0.190	[+0.135, +0.248]	+0.231	[+0.163, +0.295]
800k	REBUILD_V	+0.478	[+0.376, +0.596]	+0.662	[+0.549, +0.789]
800k	V_FROM_M	+0.032	[-0.028, +0.088]	+0.153	[+0.074, +0.224]

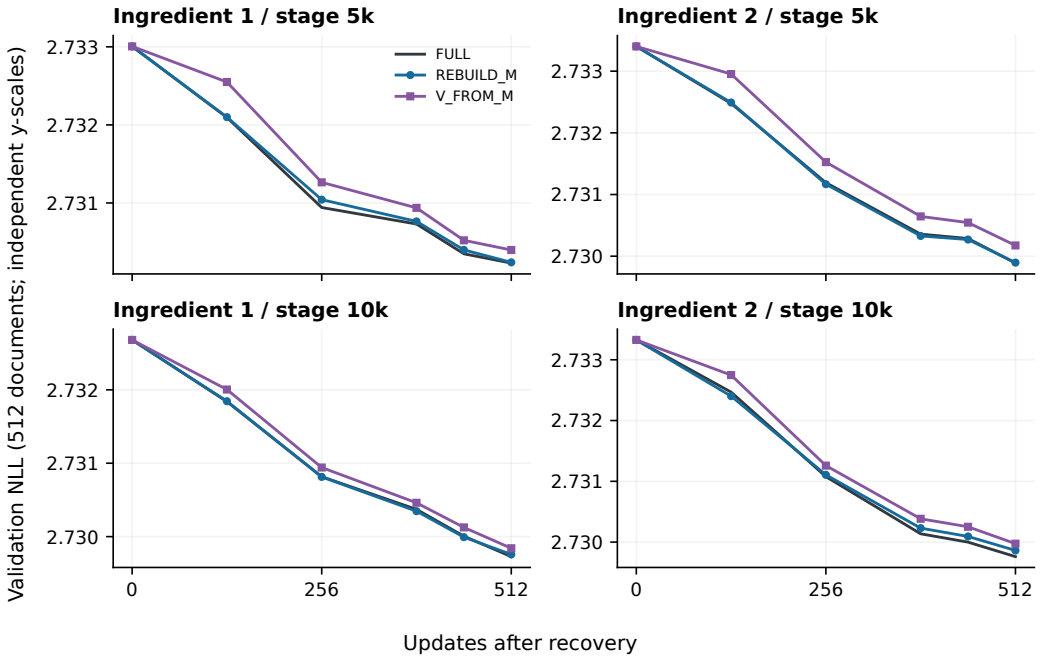

Figure 18: **Selective recovery alongside improving 1B continuation.** All points use the full 512-document evaluator. FULL, REBUILD_M, and V_FROM_M are shown for all four source states. Panels have independent vertical scales to resolve within-state changes.

D.3.1 Envelope reconstruction and retained history

Reconstruction at equal payload. V_FROM_M constructs a state on the envelope boundary without storing second-moment tensors or extra per-parameter information. At the 5k/10k states, its late NLL is 0.000116–0.000286 above REBUILD_M; the paired-document intervals in Table 20 resolve this additional benefit of actual v. Figure 20 shows the complete policy responses. REBUILD_V's larger zero-denominator transient is tabulated separately.

First recovery update. The reconstruction initializes $\widetilde{v}_i = m_i^2/C$ once. Given the first resumed gradient g_i, AdamW first updates the raw moments as

$$m_i^+ = \beta_1 m_i + (1-\beta_1) g_i, \qquad v_i^+ = \beta_2 \frac{m_i^2}{C} + (1-\beta_2) g_i^2.$$

The adaptive parameter update uses m_i^+ and v_i^+, with bias correction at the inherited clock and the original ϵ; decoupled weight decay is unchanged. Thus the envelope-boundary initialization enters the ordinary AdamW recurrence before the first parameter update.

Table 20: V_FROM_M minus REBUILD_M at equal payload (10^{-3} nats/token), with paired-document intervals for the late contrast.

Source	Late	95% CI	Step 512
I1 / 5k	+0.144	[+0.058, +0.229]	+0.158
I2 / 5k	+0.286	[+0.210, +0.365]	+0.278
I1 / 10k	+0.116	[+0.039, +0.193]	+0.085
I2 / 10k	+0.146	[+0.069, +0.220]	+0.111

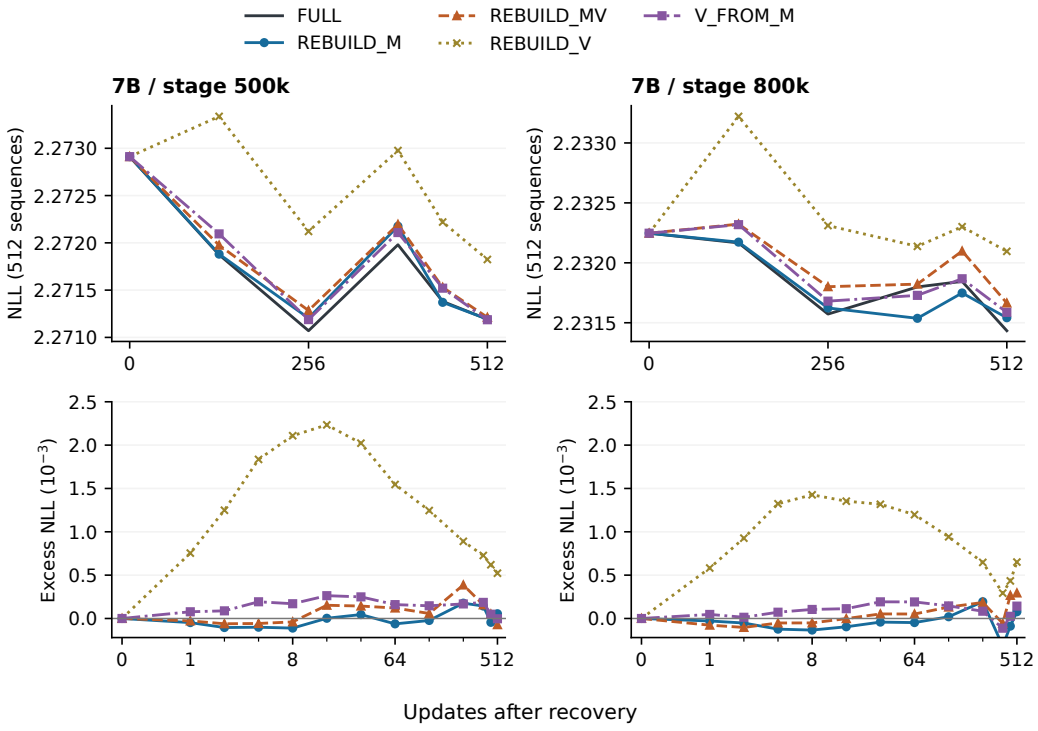

Figure 19: **State construction and retained history at 7B.** All five policies are shown. Top: absolute NLL on all 512 held-out packed sequences, with independent vertical scales across stages. Bottom: differences from FULL on the same first 64 sequences at every recorded time, with a shared vertical scale. Envelope construction suppresses the transient caused by zeroing v; REBUILD_M stays close to FULL. The late primary comparison is in Table 19.

Table 21: REBUILD_V excess NLL on the same first 64 documents at every displayed update (nats/token).

Source	1	8	64	512
I1 / 5k	+0.045454	+0.062898	+0.046821	+0.025516
I2 / 5k	+0.768712	+0.166045	+0.074321	+0.032102
I1 / 10k	+0.032442	+0.040707	+0.028990	+0.015945
I2 / 10k	+0.084660	+0.044454	+0.023668	+0.011676

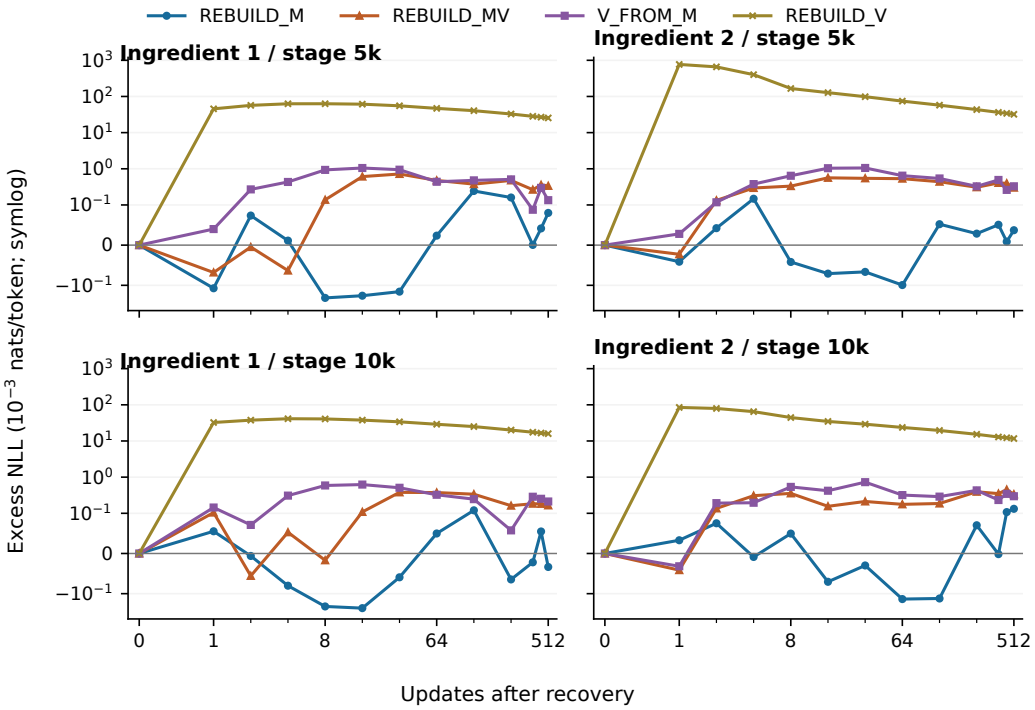

Figure 20: **Envelope reconstruction suppresses the zero-v transient.** All four 1B states and four reduced-state policies, each relative to FULL on the same first 64 documents. Symmetric-log vertical scale with linear threshold 0.1 in the displayed 10^{-3} units; horizontal scale is linear up to one update, logarithmic thereafter.

D.4 Composition with low-precision checkpoint storage

State selection determines which moment histories are retained, while encoding determines their stored representation. We combine REBUILD_M with BF16 storage and block-wise 8-bit moment quantization at the same four 1B anchors, comparing each with full-state recovery under the same encoding.

Matched BF16 checkpoint storage. FULL_BF16 stores parameters and both moments in BF16 (six bytes per parameter); REBUILD_M_BF16 stores parameters and v in BF16 (four bytes) and initializes $m = 0$. Both restore into the unchanged FP32-parameter, BF16-autocast runtime. Checkpoint storage precision changes; training arithmetic does not. The primary pair is REBUILD_M_BF16 minus FULL_BF16; Table 22 also gives FULL_BF16 minus FULL to expose the storage-precision effect. All four sources use the same fault protocol, 512-update horizon, data, clocks, and late weights as their core comparisons. The further one-third saving is 8.910 to 5.940 GB, with matched late NLL differences between -0.000060 and $+0.000033$.

Table 22: Matched BF16 checkpoint comparisons (10^{-3} nats/token). RM denotes REBUILD_M. FULL_{BF16} – FULL isolates the storage-precision change; RM_{BF16} uses FULL_{BF16} as its reference.

Source	Comparison	Late	95% CI	Step 512
I1 / 5k	FULL_{BF16} – FULL	+0.054	[-0.013, +0.118]	+0.035
I1 / 5k	RM_{BF16} – FULL_{BF16}	-0.060	[-0.123, +0.000]	-0.036
I2 / 5k	FULL_{BF16} – FULL	+0.087	[+0.008, +0.177]	+0.078
I2 / 5k	RM_{BF16} – FULL_{BF16}	-0.008	[-0.092, +0.064]	+0.017
I1 / 10k	FULL_{BF16} – FULL	+0.020	[-0.032, +0.070]	-0.005
I1 / 10k	RM_{BF16} – FULL_{BF16}	+0.020	[-0.026, +0.068]	+0.037
I2 / 10k	FULL_{BF16} – FULL	+0.055	[-0.004, +0.118]	+0.036
I2 / 10k	RM_{BF16} – FULL_{BF16}	+0.033	[-0.019, +0.085]	+0.049

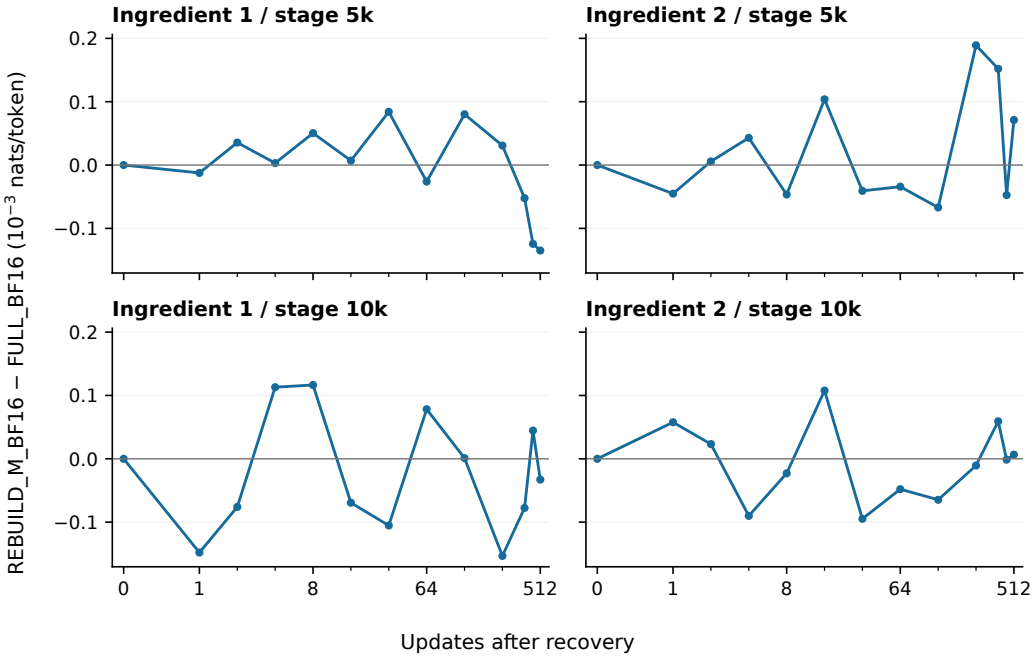

Figure 21: **Selective storage composes with BF16 checkpointing.** REBUILD_M_BF16 minus FULL_BF16 on the same first 64 documents at every time. The 512-document late and endpoint comparisons appear in Table 22.

Matched 8-bit moment storage. We apply the dynamic block-wise quantizer of Dettmers et al. (2022) using bitsandbytes 0.49.1, with blocks of 256 elements, signed codes for m and unsigned codes for v. Parameters are saved in BF16; retained moments use 8-bit codes with FP32 block scales and explicit codebooks, without nested quantization. FULL_8BIT saves both moments, whereas REBUILD_M_8BIT saves only v and reconstructs $m = 0$ once. Both load into the same training runtime as the BF16 pair and use the same anchors, 512-update recovery protocol, data, clocks, and evaluations.

Including scales, codebooks, and metadata, the actual payload falls from 5.987 to 4.478 GB, an additional 25.2% reduction. Across the four anchors, REBUILD_M_8BIT minus FULL_8BIT has late NLL differences from -0.000053 to $+0.000013$ nats/token. Table 23 reports the paired late and endpoint differences with 2,000 paired-document percentile bootstrap resamples (seed 2026091701), using the same late weights as Table 22.

Table 23: Composition with block-wise 8-bit moment storage. F_8 and R_8 denote FULL_8BIT and REBUILD_M_8BIT; R_{16} denotes REBUILD_M_BF16. Their actual payloads are 5.987, 4.478, and 5.940 GB, respectively. NLL differences are in 10^{-3} nats/token (lower is better); intervals are paired-document 95% bootstrap intervals.

Source	Comparison	Late	95% CI	Step 512	95% CI
I1 / 5k	$R_8 - F_8$	-0.053	[-0.115, +0.007]	-0.029	[-0.107, +0.054]
I2 / 5k	$R_8 - F_8$	-0.034	[-0.092, +0.025]	-0.004	[-0.081, +0.078]
I1 / 10k	$R_8 - F_8$	+0.013	[-0.039, +0.068]	-0.025	[-0.100, +0.053]
I2 / 10k	$R_8 - F_8$	-0.049	[-0.096, -0.001]	-0.041	[-0.108, +0.030]
I1 / 5k	$F_8 - R_{16}$	+0.043	[-0.017, +0.106]	+0.024	[-0.055, +0.105]
I2 / 5k	$F_8 - R_{16}$	-0.053	[-0.115, +0.011]	-0.082	[-0.170, +0.002]
I1 / 10k	$F_8 - R_{16}$	-0.005	[-0.055, +0.045]	-0.002	[-0.070, +0.066]
I2 / 10k	$F_8 - R_{16}$	-0.006	[-0.056, +0.041]	+0.044	[-0.031, +0.114]

The table also compares FULL_8BIT with REBUILD_M_BF16 at nearby payloads of 5.987 and 5.940 GB. Their late differences range from -0.000053 to $+0.000043$ nats/token. Together, the matched BF16 and 8-bit comparisons show that choosing which history to retain adds storage savings to either encoding while keeping late recovery quality close to its full-state counterpart.

D.5 Selective Checkpoint Application: Protocol and Complete Results

D.5.1 Independent training histories and evaluation

Initialization and data provenance. The application uses GPTNeoXForCausalLM with the Pythia-1B architecture (Black et al., 2022; Biderman et al., 2023) and $1{,}011{,}781{,}632$ randomly initialized parameters. Initialization seeds 2026091001/2026091002 and order keys 2026091011/2026091012 define two trajectories. The 18,000 background blocks and 48 validation blocks are exact copies of the Pythia-tokenized FineWeb selection (Penedo et al., 2024). The construction takes one deterministic 256-token window per selected canonical document and separates document-hash role buckets. Training and validation have zero exact block intersection; document-level provenance is retained from the parent construction. Each order uses two explicit permutation epochs, yielding 4,500 available optimizer updates without wraparound.

Numerical recipe and outcome. Parameters, gradients, and moments are FP32; attention is eager, dropout and TF32 are disabled, and deterministic algorithms are enabled. A microbatch of one accumulates to batch eight at sequence length 256. AdamW uses constant learning rate 0.0003, betas $(.9, .95)$, $\epsilon = 10^{-8}$, weight decay 0.1, and neither fused nor foreach kernels; there is no warmup or gradient clipping. Validation uses all 48 sequences and $48 \times 255 = 12{,}240$ next-token targets. NLL is computed from FP64-cast logits of the FP32 forward pass. Additional summaries use the first 16 and remaining 32 validation sequences.

The matrix primary statistic is

$$Y^{\mathrm{app}}_{p,e,d} = \frac{1}{512}\,\mathrm{Trap}_{h\in\{0,16,64,128,256,512\}}\big[L_{p,e,d}(t+h) - L_{F,e}(t+h)\big]. \tag{73}$$

Here Trap is the trapezoidal integral, not an unweighted average of sampled losses. The complete matrix evaluation records also include the early updates $1, 2, 4, 8, 32$. The primary 512-update statistic uses only the grid displayed in Equation 73; the matched 64-update four-corner comparison uses the recorded grid $0, 1, 2, 4, 8, 16, 32, 64$, including S2, without interpolation. Positive values are excess NLL; negative values are retained as signed differences. S0 restores the entire state at $s = t - d$ and executes $d + 512$ actual updates. Other policies execute 512 from current weights and clock. Exact replay makes S0 zero at common logical clocks but does not remove its extra compute.

Table 24 summarizes the policies.

Complete matrix and nonmonotonicity. Table 25 contains every tested cell. S3 and S5 have four distinct seed–stage interventions each because their selected tensors do not depend on d. At zero m, S3 improves on S4 in 7/8 age rows, but seed 2 at step 2048, age 16 reverses the comparison: $+0.00213$ versus -0.00200. Age 64 also improves over age 16 for S4 in two of four seed–stage pairs.

Table 24: State-value matrix: signed mean excess NLL over 512 logical updates (lower is better). Seed columns average nested stages and ages; the range is not a confidence interval. S3/S5 age rows are duplicate interventions.

Policy	Restored tensors	Seed 1	Seed 2	Cell range
S0	(θ_s, m_s, v_s) + replay	+0.00000	+0.00000	+0.00000, +0.00000
S1	(θ_t, m_s, v_s)	+0.00341	+0.00302	-0.00052, +0.00923
S2	(θ_t, m_s, v_t)	+0.00363	+0.00485	-0.00382, +0.01047
S3	$(\theta_t, 0, v_t)$	-0.00430	+0.00078	-0.00531, +0.00213
S4	$(\theta_t, 0, v_s)$	-0.00031	+0.00363	-0.00493, +0.00818
S5	$(\theta_t, 0, 0)$	+0.03044	+0.02903	+0.02021, +0.03785

Table 25: All 48 matrix cells: 512-logical-update mean excess NLL relative to paired FULL. S3 and S5 construct the same initial recovery state at both source ages.

Seed	Stage	Age	S0	S1	S2	S3	S4	S5
1	512	16	+0.00000	+0.00264	+0.00760	-0.00328	-0.00151	+0.02843
1	512	64	+0.00000	+0.00520	+0.00520	-0.00328	+0.00751	+0.02843
1	2048	16	+0.00000	+0.00322	-0.00173	-0.00531	-0.00231	+0.03245
1	2048	64	+0.00000	+0.00259	+0.00347	-0.00531	-0.00493	+0.03245
2	512	16	+0.00000	-0.00052	-0.00382	-0.00057	+0.00818	+0.02021
2	512	64	+0.00000	+0.00010	+0.00586	-0.00057	+0.00330	+0.02021
2	2048	16	+0.00000	+0.00923	+0.01047	+0.00213	-0.00200	+0.03785
2	2048	64	+0.00000	+0.00327	+0.00689	+0.00213	+0.00506	+0.03785

Thus the results favor retaining useful second-moment history without making temporal proximity a monotone quality rule. Retaining old m in S1 instead of zeroing it in S4 wins four age rows and loses four: the preferred first-moment action varies with the recovery state.

D.5.2 Durable inventories, fault schedules, and byte accounting

Saved tensor allocation. All policies start from the same complete step-2048 anchor. Their subsequent publication intervals in committed logical updates are:

Policy	Parameter/core	First moment	Second moment
S0	Complete atomic state every 48 updates		
S1	24	96	96
S2	40	80	40
S3	32	omitted; reset on recovery	32
S4	24	omitted; reset on recovery	48
S5	16	omitted; reset on recovery	omitted; reset on recovery

Each schedule plans 30 parameter-sized tensor writes over 480 uninterrupted updates, excluding the common anchor. The parameter/core payload also preserves the optimizer clock, parameter groups and name mapping, scheduler, RNG, and next data position. Omitting a moment never resets the Adam clock. Unlike the controlled matrix, E2E recovery loads the latest durable parameter version, so its actual moment ages and replay costs are determined by the publication schedule and fault history rather than fixed at 16 or 64.

Publication, reclamation, and recovery. All jointly due components are serialized, file-synchronized, and renamed before one atomic HEAD manifest publishes the transaction. The loader checks lineage, parameter schema, optimizer-group identity, payload sizes, and hashes, then reads only those durable files and constructs zero tensors for missing moments using the parameter shapes. The newest two versions per required component are retained. Once all required component heads exist, a lightweight policy can retire the common full anchor. Only obsolete experiment-owned blobs are reclaimed, after durable replacement, while the append-only version ledger remains. Peak bytes are measured before reclamation. Byte counters cover serialized checkpoint blobs, including their in-file metadata and buffers; separate JSON manifests, operational logs, and physical filesystem allocation are not counted. GiB denotes 2^{30} bytes.

The single fault occurs after committed call 239; repeated faults occur after 119, 239, and 359. Each is consumed once even if loading rolls back logical progress. The worker is actually killed and replaced by a fresh disk-only process. Runs stop at 480 net logical updates or 768 actual updates, whichever comes first; all 36 attain logical step 2528 within this cap. Every policy's initial full-anchor write and load are charged. The 12 fault-free controls have identical terminal states within each seed, and all four interrupted S0 terminal states exactly match their corresponding fault-free reference.

Quality and costs are reported jointly. Table 26 includes all policies for one and repeated faults; Figure 22 includes every trajectory. S3's single-fault endpoint differences are -0.00211 and $+0.00159$; repeated faults yield -0.00501 and $+0.00312$. S5 trades still lower storage for consistently larger interrupted endpoint loss. S0 repeats 47/71 updates under single/repeated faults; S3 repeats 15/71. Actual publication totals, not planned rates, yield the near-equal written payload range 124.3855–124.3871 GiB across all 36 runs.

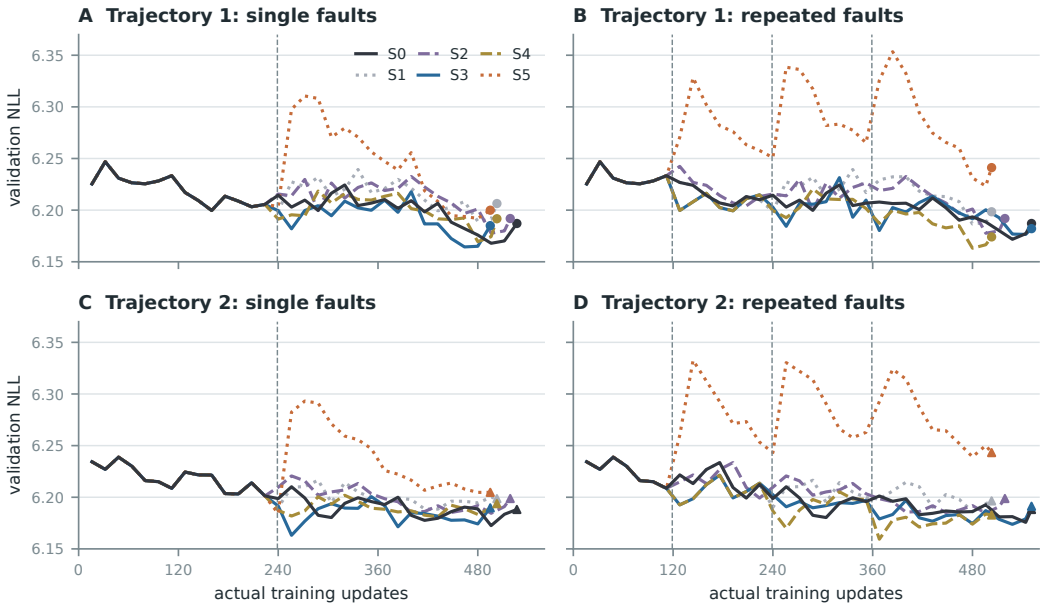

Figure 22: **Recovery policies under single and repeated faults.** Rows are initialization seeds; columns are single and repeated faults. All axes use the same scales, and horizontal time is actual committed training updates, not wall-clock time or net logical progress. Vertical dashed lines mark the fixed fault calls. The six no-fault curves coincide exactly and are omitted here. Endpoints are separately marked and all attain logical step 2528.

D.5.3 Quality targets and measured recovery time

Time-to-quality criterion. For each seed, the target equals the mean NLL of the final three S0 no-fault checks plus 0.001. Time to quality requires three consecutive qualifying checks strictly after the last fault. If the no-fault baseline cannot meet that criterion within its observed horizon, the whole seed's time-to-quality comparison is unavailable. The targets are 6.17765269 and 6.18280103. Seed 1's baseline attains the criterion; seed 2's longest baseline streak has only two checks, so all its TTQ entries remain unavailable. For seed 1, only S0/S3 attain the single-fault target, and only S4 attains it under repeated faults. Other runs are right-censored at their last actual observation.

Table 26: Every interrupted policy: endpoint excess NLL relative to fault-free FULL at logical step 2528. Columns 1/2 denote initialization seeds. Peak/read columns use repeated faults and include the common full anchor.

Policy	Single 1	Single 2	Repeated 1	Repeated 2	Peak GiB	Read GiB
S0	+0.00000	+0.00000	+0.00000	+0.00000	33.92	45.23
S1	+0.01930	+0.01048	+0.01134	+0.00830	33.92	45.23
S2	+0.00480	+0.01062	+0.00480	+0.01062	33.92	45.23
S3	-0.00211	+0.00159	-0.00501	+0.00312	22.62	33.92
S4	+0.00460	+0.00565	-0.01300	-0.00536	22.62	33.92
S5	+0.01285	+0.01660	+0.05409	+0.05507	15.08	22.62

Quality at fixed time budgets. Operational time sums training, snapshot/state copying, durable publication and hashing/reclamation, and each loader's spawn-to-ready cost. Evaluation and common experimental anchor preparation are reported separately. Time budgets are $0.5/1/1.5$ times that seed's S0 no-fault operational endpoint: 1325.99 and 1216.47 seconds at the $1\times$ level. We use the last evaluation within a cap, retaining both its NLL and logical progress (Table 27). All $1.5\times$ caps exceed the observed run endpoints, so their entries retain the endpoint and explicitly flag that fact.

Table 27: Repeated faults at a time budget equal to each seed's fault-free S0 operational endpoint. We use the last observed evaluation within the budget; progress is net logical updates.

Policy	NLL, seed 1	Progress 1	NLL, seed 2	Progress 2
S0	6.21178	361	6.19855	329
S1	6.23237	377	6.21454	377
S2	6.21338	393	6.19859	329
S3	6.20035	425	6.18226	377
S4	6.19626	377	6.19792	329
S5	6.29437	393	6.31442	377

Timing conditions. The host is shared and its resource occupancy varies across runs; the page cache is retained. Full snapshots additionally compute a model/optimizer/RNG state witness inside the timed path, while partial snapshots omit this operation. Both paths hash the published payload. Table 28 reports the observed operational costs, including these differences.

Table 28: All 36 runs: replay updates, operational time, actual written bytes, and time-to-quality status. Timing components and the quality threshold are defined in the text.

Policy	Seed	Faults	Replay	Op. s	Written GiB	TTQ
S0	1	none	0	1326.0	124.3855	attained
S0	1	single	47	1355.3	124.3855	attained
S0	1	repeated	71	1639.3	124.3855	censored
S0	2	none	0	1216.5	124.3855	unavailable
S0	2	single	47	1362.9	124.3855	unavailable
S0	2	repeated	71	1584.4	124.3855	unavailable
S1	1	none	0	1066.3	124.3863	attained
S1	1	single	23	1218.5	124.3863	censored
S1	1	repeated	23	1552.9	124.3863	censored
S1	2	none	0	1229.1	124.3863	unavailable
S1	2	single	23	1309.3	124.3863	unavailable
S1	2	repeated	23	1401.4	124.3863	unavailable
S2	1	none	0	1058.0	124.3856	attained
S2	1	single	39	1356.7	124.3856	censored
S2	1	repeated	39	1546.2	124.3856	censored
S2	2	none	0	1122.7	124.3856	unavailable
S2	2	single	39	1316.5	124.3856	unavailable
S2	2	repeated	39	1543.4	124.3856	unavailable
S3	1	none	0	1062.5	124.3859	attained
S3	1	single	15	1152.4	124.3859	attained
S3	1	repeated	71	1428.9	124.3859	censored
S3	2	none	0	1015.8	124.3859	unavailable
S3	2	single	15	1301.3	124.3859	unavailable
S3	2	repeated	71	1403.1	124.3859	unavailable
S4	1	none	0	976.6	124.3863	attained
S4	1	single	23	1320.8	124.3863	censored
S4	1	repeated	23	1556.3	124.3863	attained
S4	2	none	0	1116.7	124.3863	unavailable
S4	2	single	23	1313.6	124.3863	unavailable
S4	2	repeated	23	1510.9	124.3863	unavailable
S5	1	none	0	1361.5	124.3871	attained
S5	1	single	15	1143.4	124.3871	censored
S5	1	repeated	23	1534.4	124.3871	censored
S5	2	none	0	1031.3	124.3871	unavailable
S5	2	single	15	1183.8	124.3871	unavailable
S5	2	repeated	23	1412.8	124.3871	unavailable

Table 29: Repeated-fault recovery at 480 net logical updates on the original 48 sequences. Signed endpoint differences are relative to fault-free FULL. Peak inventory and cumulative loaded-file payload include the common anchor and agree across seeds at the shown precision. All six policies appear in Appendix D.5.

Policy	ΔL, seed 1	ΔL, seed 2	Peak GiB	Read GiB
S0	+0.0000	+0.0000	33.92	45.23
S3	-0.0050	+0.0031	22.62	33.92
S4	-0.0130	-0.0054	22.62	33.92
S5	+0.0541	+0.0551	15.08	22.62

D.6 Fresh-Document Reevaluation of Repeated-Fault Terminals

We reevaluate the eight S0/S3/S4/S5 terminals at logical step 2528 from the two fresh-initialization Pythia trajectories. The original 48-sequence NLLs reproduce exactly. The new evaluation uses 2,048 canonical documents, the original Pythia tokenizer, one 256-token window per document (522,240 targets), and the same paired-document bootstrap convention. It evaluates the existing terminals and shares 448 canonical documents with the OLMo evaluation set; the two studies are reported separately.

Table 30: Fresh-document terminal-only reevaluation (2,048 documents): policy minus FULL NLL in nats/token and paired-document intervals.

Seed	Policy	Mean	95% interval
2026091001	S0	$+0.000000$	$[+0.000000, +0.000000]$
2026091001	S3	$+0.000605$	$[-0.001476, +0.002548]$
2026091001	S4	-0.006656	$[-0.008523, -0.004747]$
2026091001	S5	$+0.062811$	$[+0.060539, +0.064972]$
2026091002	S0	$+0.000000$	$[+0.000000, +0.000000]$
2026091002	S3	-0.001504	$[-0.003514, +0.000528]$
2026091002	S4	-0.004429	$[-0.006556, -0.002239]$
2026091002	S5	$+0.050402$	$[+0.048139, +0.052624]$

S3's first interval has an upper endpoint above $+0.001$, while its second straddles zero. S4 has negative mean differences for both seeds, and S5 has large positive differences. These results complement the original trajectory measurements with wider document coverage. Time-to-quality statistics continue to use the original evaluator and threshold.

D.7 Recovery from Durable Files and Crash-Phase Sensitivity

Component sources. The 70M disk-recovery experiment tests how durable file selection affects moment pairing. The 1B application in Appendix D.5 then measures selective saving and checkpoint costs. Source-order comparisons require components to agree on the first three fields of the provenance tuple

$$(\text{lineage id},\ \text{optimizer-config hash},\ \text{parameter/shard-manifest id},\ \text{source step}). \tag{74}$$

If identity validation or admissible source selection fails, the loader uses an atomic/common fallback.

Durability path. The writer uses real PyTorch serialization followed by file `fsync`, atomic rename, and directory `fsync`. Parameters and each moment component have independently published version metadata; a global-atomic control publishes a manifest only after all current components are durable. The crash controller chooses one of 128 fixed phase fractions for a 16-update publication window, observes the phase witness, and sends `SIGKILL`. A new process scans only durable artifacts and records the selected inventory. Schedule fractions are paired, so write order is the only designed difference between the m- and v-priority executions.

Observed inventory and a same-population metric. Inventory prevalence is measured directly from 256 crash executions. Conditional risk maps these inventories to the measured 70M responses

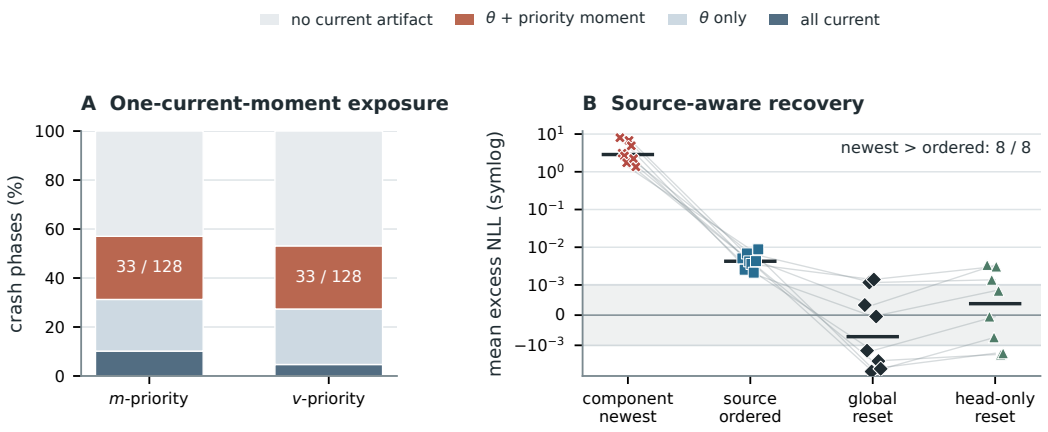

Figure 23: **Actual disk recovery preserves the controlled compatibility effect.** (A) In 256 real kills, each schedule exposes current weights plus its priority moment in 33/128 phases. (B) Across two trajectories with four checkpoints each, component-newest is separated by orders of magnitude from ordered and reset actions; lines connect actions within a stage and black bars mark medians. The symmetric-log axis has a 0.001-nat linear region. Disk-only recovery reproduces all 56 matching parent continuations.

for phases where current parameters are durable. To compare risk and lost progress over the same population, we also report

$$\bar{R}(\pi) = \sum_{u \in \mathcal{U}} p(u) R(\pi(I_u)) \,, \tag{75}$$

where I_u is the crash-visible inventory at phase u, $p(u) = 1/128$ for the primary analysis, and branches without a mapped mismatch contribute zero mismatch risk. Table 31 reports both risk views and computes lost progress over the same 128 phases.

Table 31: Primary random-crash mapping. Both priority schedules contain 33/128 current-θ/priority-moment inventories. "All-phase risk" and lost updates use the same uniform population.

Schedule	Loader	Cond. risk	All-phase risk	Lost upd.
m-priority	component-newest	1.406711	0.802265	6.875
m-priority	source-ordered	0.002082	0.001188	6.875
m-priority	atomic/common	0	0	14.375
v-priority	component-newest	0.001867	0.000992	7.500
v-priority	source-ordered	0.001867	0.000992	7.500
v-priority	atomic/common	0	0	15.250

Within one verified provenance lineage, the source-ordered action is

$$\mathcal{R}(m_{t_m}, v_{t_v}) = \begin{cases} (m_{t_m}, v_{t_v}), & t_m \le t_v, \\ (m_{s^\star}, v_{t_v}), \quad s^\star = \max\{s \in \mathcal{M} : s \le t_v\}, & t_m > t_v, \end{cases} \tag{76}$$

where $\mathcal{M}$ is the durable set of compatible first-moment versions. If no such version exists, or the lineage/config/manifest fields differ, the loader uses an atomic/common fallback.

Recovery from durable state. The closure protocol materializes eight inventories spanning two training trajectories and four checkpoints each. Two actual process kills are used per inventory. Each continuation starts in a fresh process and loads only durable state. State identity, future batches, and evaluation are checked against the paired reference. The full matrix contains 72 branches: two FULL repeats, a disk-copy sham, component-newest and source-ordered actions for both priority directions, and global/head-reset actions for the high-risk direction. All 56 parent-equivalent branches reproduce the parent curve and risk with zero error. Each selected moment-payload pair is 534.1 MiB; the median combined read, hash, and deserialization time on this host is 0.764 seconds for source-ordered branches.

Table 32: End-to-end recovery actions on the high-risk m-newer inventory. Rows summarize the eight checkpoints. Positive-part and catastrophic counts use thresholds 0.001 and 0.01, respectively.

Action	Median	Mean	Endpoint med.	$> .001$	$> .01$
component-newest	2.866482	3.833457	0.690657	8/8	8/8
source-ordered	0.004251	0.004683	0.001317	8/8	0/8
global m reset	-0.000714	-0.001288	0.003786	2/8	0/8
head-only m reset	0.000380	0.000560	0.002038	3/8	0/8

The paired component-newest minus source-ordered mean is 3.860744 on one trajectory and 3.796804 on the other. Global and head reset are slightly below source ordering in integrated risk (median differences -0.005206 and -0.003768). Negative signed risk denotes lower measured NLL than the paired FULL branch.

Relation to recovery interfaces. ByteCheckpoint and DeepSpeed Universal Checkpointing support consistent distributed state reconstruction (Wan et al., 2025; Lian et al., 2025). Here, component-newest selection exposes the effect of combining different moment sources, while source-ordered selection enforces history coverage.

Phase-prior and environment sensitivity. We recompute the inventory map over the same eight 70M environments using three crash-phase priors: here u is the fraction of the persistence schedule elapsed at the crash, and $\mathcal{U}$ contains the 128 sampled crash phases for that schedule.

$$p_\lambda(u) = \frac{\exp[\lambda(u - \frac{1}{2})]}{\sum_{u' \in \mathcal{U}} \exp[\lambda(u' - \frac{1}{2})]}, \qquad \lambda \in \{-2, 0, 2\}. \tag{77}$$

Effective sample sizes are 96.9, 128, and 93.2. Source ordering improves 24/24 environment–prior comparisons; the median and minimum relative reductions are 99.760% and 98.427%, with zero lost-progress change (Figure 24A–B).

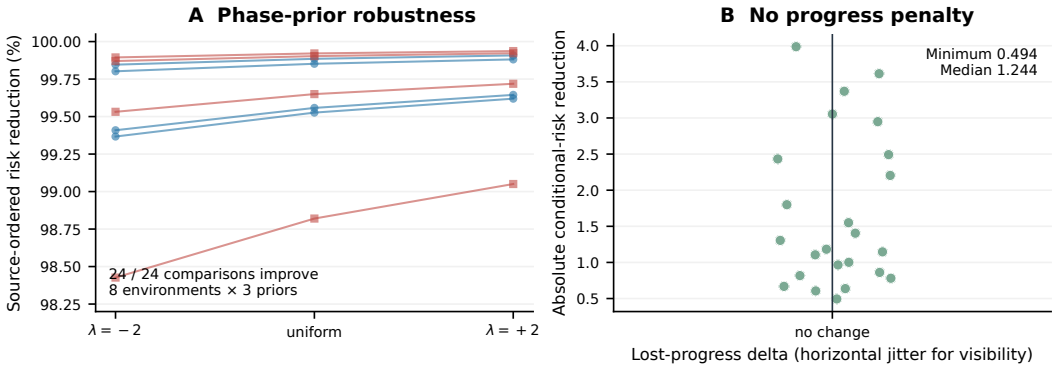

Figure 24: **Source-ordered recovery under different crash-phase priors.** (A) Source-ordered improvement remains above 98.4% for all eight functional environments under three crash-phase priors. Lines connect priors within one environment. (B) All 24 mappings reduce absolute conditional risk with zero expected lost-progress change.

D.8 Continuation Settings and Recovery Sensitivity

Matching the schedule to the continuation batch. The continuation experiments use smaller effective batches than the source pretraining runs to keep the multi-policy comparisons within the available compute budget. We select the batch–schedule combination using FULL alone, then hold it fixed across recovery policies. Table 33 shows the diagnostic from the 1B ingredient-1 10k state, entered directly without the common prefix, and the corresponding 7B 500k state. At 1B, reducing the scheduler multiplier from 1 to 0.1 at batch 4 changes the 128-update NLL movement from $+0.05591$ to -0.01237. Increasing batch to 16 at the original schedule also reduces the rise, to $+0.01937$. At 7B, the lower rate removes most of the batch-4 rise; increasing batch to 16 at that rate yields a decreasing NLL.

Table 33: FULL-only continuation diagnostics: absolute validation NLL (nats/token). B counts 4096-token sequences per update; λ multiplies the source scheduler. Each model uses the same evaluation items across conditions and times. The 1B source is ingredient 1; 7B comparisons use the common first 64 sequences. The 7B short runs end at update 64.

Source	B/λ	N	0	32	64	128
1B / 10k	4 / 1.0	512	2.746802	2.783770	2.791749	2.802709
1B / 10k	4 / 0.1	512	2.746802	2.739037	2.736567	2.734434
1B / 10k	16 / 1.0	512	2.746802	2.759632	2.762992	2.766168
7B / 500k	4 / 1.0	64	2.271613	2.443487	2.489929	–
7B / 500k	4 / 0.1	64	2.271613	2.273946	2.274345	–
7B / 500k	16 / 0.1	64	2.271613	2.271163	2.270865	–

The batch effect is also visible at matched input-token budgets. On the full 512-sequence 7B evaluator, batch 16 at update 16 improves NLL by 0.00310 relative to batch 4 at update 64, with both using the 0.1 schedule multiplier. Each condition processes 1,048,576 input tokens. The native 7B scheduler advances by consumed tokens, while the 1B schedule advances by updates. These diagnostics motivate batch 4 for 1B and batch 16 for 7B, both at multiplier 0.1, as used in the main recovery comparison.

Recovery with the original learning-rate schedule. Table 34 reports the same recovery policies with the unscaled source scheduler and effective batch 4. The 1B experiments use both ingredients at 10k/20k, with a 256-update common prefix at this same unscaled rate; the 7B experiments start directly from 500k/800k. Within each setting, the policies share their recovery anchor, data order, clocks, evaluator, and 512-update horizon. Thus the two settings compare complete continuation configurations: the 1B prefix follows its respective schedule, and the 7B effective batch changes from 4 to 16.

REBUILD_M tracks the broad trajectory of matched FULL under these reduced-batch continuations with the unscaled source schedule, while saving one third of the checkpoint payload. With the original schedule, its four 1B late differences range from -0.000435 to -0.000062 nats/token. At 7B, the late differences are -0.010998 and $+0.003023$, and the 512-update differences are $+0.002275$ and $+0.000125$. Figure 25 shows how the recovered branches follow FULL even when the continuation NLL itself rises. The absolute trajectories show the continuation dynamics; paired differences quantify the additional effect of state reconstruction under each configuration. Together with the improving-continuation results in Table 2, these measurements show that retaining v gives a practical recovery policy across both tested batch–schedule settings.

The 1B alternatives also clarify the value of retained history at the original rate. Reconstructing $v = m^2/C$ reduces the large REBUILD_V late excess from 0.0979–0.5364 to 0.00091–0.00479; retaining the actual v brings recovery still closer to FULL.

Table 34: Recovery with the unscaled source scheduler ($\lambda = 1$, effective batch 4). Entries are differences from matched FULL in 10^{-3} nats/token, using all 512 evaluation items. Late means use the same 384/448/512 weighting as the main comparison. Intervals use 2,000 paired document resamples (1B) or within-source sequence resamples (7B). RM, RV, RMV, and VFM denote REBUILD_M, REBUILD_V, REBUILD_MV, and V_FROM_M.

State	Policy	Late	Late 95% interval	512
1B I1 / 10k	RM	-0.337	$[-0.774, +0.247]$	-0.782
1B I1 / 10k	RV	$+536.377$	$[+519.712, +552.555]$	$+502.795$
1B I1 / 10k	VFM	$+4.522$	$[+3.988, +5.256]$	$+3.334$
1B I1 / 10k	RMV	$+3.235$	$[+2.760, +3.796]$	$+2.657$
1B I1 / 20k	RM	-0.118	$[-0.222, -0.017]$	-0.151
1B I1 / 20k	RV	$+97.907$	$[+92.572, +103.389]$	$+92.447$
1B I1 / 20k	VFM	$+0.910$	$[+0.809, +1.017]$	$+0.842$
1B I1 / 20k	RMV	$+0.614$	$[+0.480, +0.751]$	$+0.541$
1B I2 / 10k	RM	-0.435	$[-0.808, -0.030]$	-0.627
1B I2 / 10k	RV	$+497.361$	$[+481.431, +513.498]$	$+461.071$
1B I2 / 10k	VFM	$+4.790$	$[+4.342, +5.261]$	$+4.219$
1B I2 / 10k	RMV	$+2.861$	$[+2.428, +3.344]$	$+2.463$
1B I2 / 20k	RM	-0.062	$[-0.155, +0.024]$	-0.097
1B I2 / 20k	RV	$+121.414$	$[+116.360, +127.007]$	$+112.899$
1B I2 / 20k	VFM	$+1.020$	$[+0.918, +1.116]$	$+0.990$
1B I2 / 20k	RMV	$+0.614$	$[+0.508, +0.714]$	$+0.590$
7B / 500k	RM	-10.998	$[-14.896, -7.666]$	$+2.275$
7B / 500k	VFM	-10.736	$[-14.999, -6.330]$	$+3.986$
7B / 800k	RM	$+3.023$	$[+1.129, +5.143]$	$+0.125$
7B / 800k	VFM	$+5.722$	$[+3.794, +7.988]$	-1.946

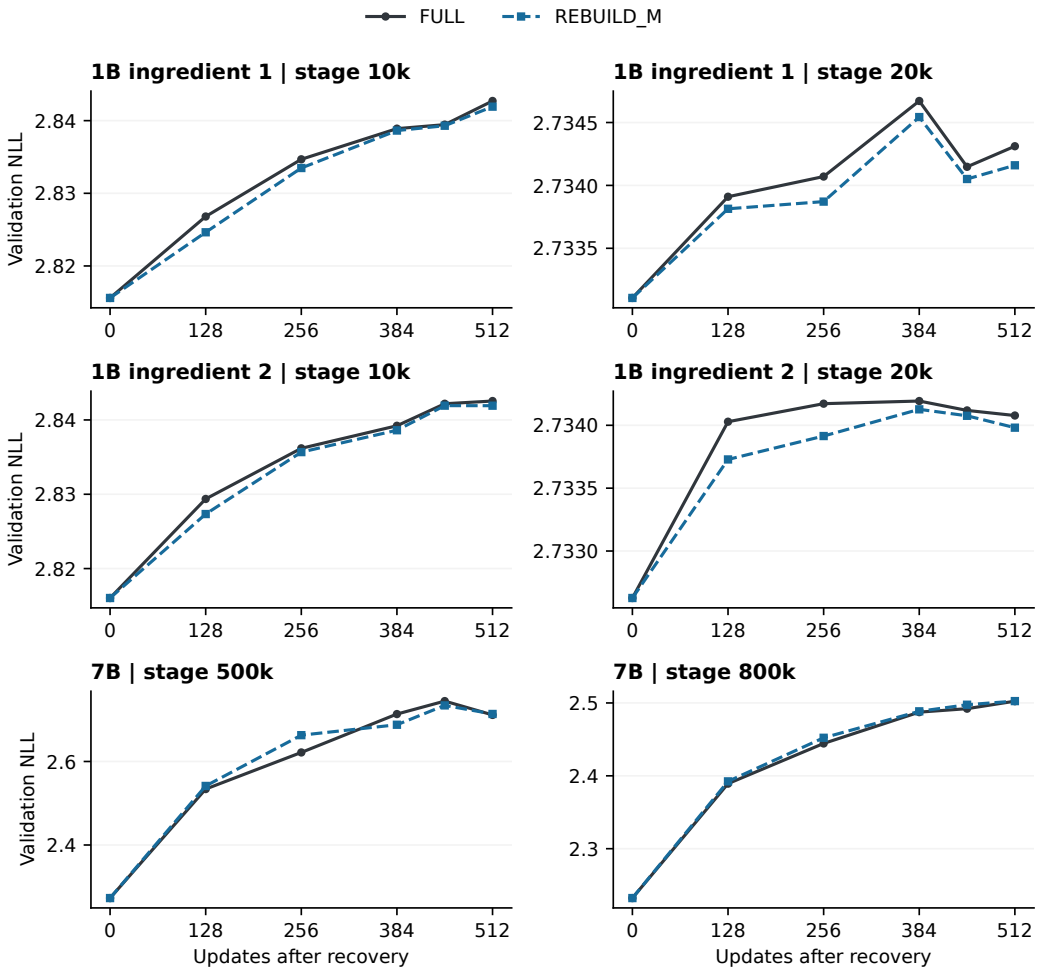

Figure 25: **Selective recovery tracks FULL under the original schedule.** All panels use effective batch 4 and the unscaled source scheduler. Solid lines show FULL and dashed lines show REBUILD_M; every point uses all 512 evaluation items. Each panel has its own vertical scale to show within-state trajectories. Table 34 reports the paired late and endpoint differences.