WHAT AGES A GRADIENT? WHEN OPTIMIZER TIME AND GRADIENT-FIELD RELEVANCE DIVERGE

Anonymous authors
Paper under double-blind review

ABSTRACT

AdamW assigns age by counting optimizer updates, but equal update counts can correspond to very different amounts of learning-state change. We ask whether elapsed optimizer time is sufficient to predict how well a historical gradient direction remains aligned with the current gradient field. Fixed held-out probes keep text identity constant, isolating gradient-field evolution from minibatch variation. On Pythia-160M, elapsed steps and accumulated learning rate leave a reproducible predictive gap: OBJECTIVE, a loss-normalized gradient–update work diagnostic, reduces held-out prediction error by **9.87%** and **7.48%** in two environments and transfers between them without refitting. The gap survives a new seed and both shorter and longer first-moment memory; under the longer-memory setting, a scale-normalized parameter-motion diagnostic reaches **+16.00%**. A post-discovery decomposition shows that learning state captures substantially more of the residual: a four-variable state summary raises held-out R^2 from **0.588/0.621** to **0.741/0.731**. Parameter motion and training progress explain part of this signal, while temporal misalignment erodes the remainder. For actual stochastic gradients, relevance also reflects alignment at gradient birth: unconditional state transfer is weak, while a prespecified birth-conditioned comparison is positive. Together, these results show that fixed-timescale optimizer memory compresses training histories that can have different present-day gradient relevance because they traverse different learning states.

1 INTRODUCTION

AdamW can assign a gradient an age without inspecting it: it simply counts optimizer updates. For the first moment, with $m_0 = 0$,

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1)\tilde{g}_t, \tag{1}$$

$$m_{t+1} = (1 - \beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \tilde{g}_k. \tag{2}$$

Thus a gradient observed at step k receives a weight proportional to β_1^{t-k} at step t. In this sense, one update is one unit of age. This exponential weighting is a memory rule, however, not a measured law of relevance; the full AdamW update also depends on the current optimizer state.

Learning need not tick uniformly. Twenty nearly redundant updates and twenty updates that substantially change the model have the same chronological age $t - k$, yet a historical direction need not remain equally aligned with the current gradient field in those two cases. Recent work shows that old gradients can remain useful and that optimizer memory can be made adaptive (Pagliardini et al., 2025; Topollai & Choromanska, 2026; Hong & Qu, 2026). Those studies ask how gradients should be retained. We ask a prior question: *what should count as "older" in the first place?*

The question is not how to keep gradients longer. It is whether elapsed optimizer time is a sufficient coordinate for relevance. We turn that question into a measurement problem. At checkpoints k and t, fixed probes evaluate the same held-out text and measure the directional alignment between its historical and current gradients, isolating gradient-field evolution from changes in minibatch identity. After elapsed steps and accumulated learning rate are known, we ask what still

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.

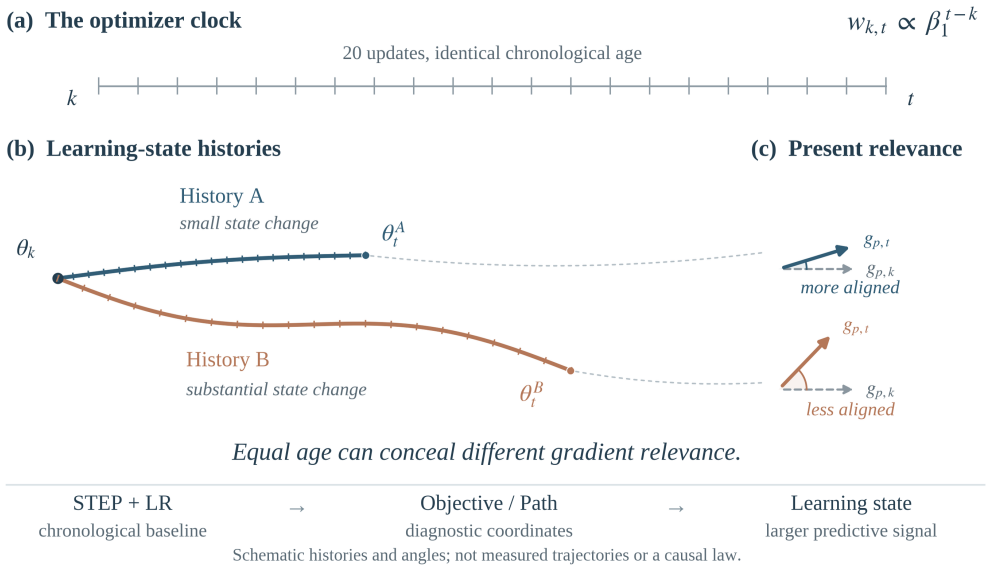

Figure 1: **Equal optimizer age can correspond to different gradient relevance.** (a) AdamW assigns age and first-moment weight through elapsed updates. (b) Two histories share an initial checkpoint and update count but differ in learning-state change. (c) Historical/current gradient alignment can differ. The lower strip links the STEP–LR baseline, OBJECTIVE/PATH diagnostics, and the larger state signal. Histories and angles are schematic, not measured trajectories or a causal law.

predicts the present relevance of a historical direction. Here, relevance means cross-time directional alignment, not measured training speedup.

Elapsed optimizer time is not sufficient. We use two diagnostic coordinates to expose what STEP–LR misses: OBJECTIVE, which summarizes loss-normalized gradient–update work, and PATH, which summarizes scale-normalized parameter motion. In two Pythia-160M environments, OBJECTIVE adds held-out predictive value beyond a flexible STEP–LR baseline and transfers between environments without refitting. A third environment preserves the gap, but PATH is more stable there, and the same phenomenon survives controlled changes to the first-moment timescale. Across these settings, the stable observation is the predictive gap itself rather than any one diagnostic coordinate. **At a descriptive level, fixed-timescale memory therefore assigns the same age to training histories that can have different present-day gradient relevance because they traverse different learning states.**

Learning state explains much of the larger residual. Decomposing OBJECTIVE reveals stronger contributions from gradient–update alignment and inverse loss, and a four-variable state summary explains substantially more held-out variation than STEP–LR alone. Parameter motion and absolute training progress absorb part of the gain but do not remove it. When the same state values are deliberately shifted away from the intervals they originally describe, the remaining predictive value decays.

Actual gradients add variation at birth. Fixed probes intentionally remove minibatch identity, whereas a training gradient also depends on the minibatch that produced it. In a prospective bridge experiment, unconditional state augmentation is weak for actual gradients, while conditioning on the gradient's alignment with the held-out field at birth yields a positive secondary effect. The comparison separates two sources of variation: how a gradient aligns when it is created, and what learning state the model traverses afterward.

We contribute a measurement framework for gradient age. Figure 1 summarizes the argument: measure the gap left by STEP/LR, stress it across trajectories and memory timescales, decompose the gap into learning-state variables, and finally test where the fixed-probe abstraction changes for actual gradients.

Concept	Role in the paper
Elapsed time	STEP/ LR: elapsed updates / accumulated learning rate over an interval.
Diagnostics	OBJECTIVE: loss-normalized gradient–update work; PATH: scale-normalized cumulative parameter motion.
Learning state	Four summaries: gradient norm, update norm, absolute alignment, and inverse loss.
Gain	Reduction in held-out prediction error relative to a fitted baseline.

2 MEASURING RELEVANCE WITHOUT NAMING A CLOCK

2.1 HISTORICAL GRADIENT-FIELD RELEVANCE

A fixed probe measures field evolution while holding text identity constant. Let a fixed held-out batch p have gradients $g_{p,k} = g_p(\theta_k)$ and $g_{p,t} = g_p(\theta_t)$ at checkpoints k and t. Their difference comes only from the model checkpoints. We compare both gradients in the current Adam geometry,

$$P_t = \operatorname{diag}\Big((\sqrt{\hat{v}_t} + \varepsilon_{\text{adam}})^{-1}\Big). \tag{3}$$

With $\|g\|_{P_t} = \sqrt{g^\top P_t g}$, the primary response is

$$R_{k,t}^{\text{dir,currentP}} = \frac{g_{p,t}^\top P_t g_{p,k}}{\|g_{p,t}\|_{P_t}\,\|g_{p,k}\|_{P_t} + \varepsilon_{\text{metric}}}. \tag{4}$$

Positive values indicate aligned directions and values near zero indicate weak alignment. We call this *current-P geometry*. Raw and reference-P controls give the same qualitative conclusion (Figure 4).

2.2 CHRONOLOGICAL BASELINE AND DIAGNOSTIC COORDINATES

STEP and LR define the chronological baseline; OBJECTIVE and PATH are diagnostics for what that baseline misses. All four summarize what happened between k and t. We write

$$A_{k,t}^c = \sum_{i=k}^{t-1} q_i^c. \tag{5}$$

For STEP, $q_i^{\text{step}} = 1$; for LR, $q_i^{\text{lr}} = \eta_i$. Relative PATH measures scale-normalized loss-driven parameter motion, while OBJECTIVE measures loss-normalized absolute gradient–update work:

$$q_i^{\text{path}} = \sqrt{\sum_\ell \frac{\|u_i^{\text{loss},\ell}\|_2^2}{\|\theta_i^\ell\|_2^2 + \varepsilon_{\text{path}}}}, \tag{6}$$

$$q_i^{\text{obj}} = \frac{|(g_i^{\text{raw}})^\top u_i^{\text{loss}}|}{L_i + \varepsilon_{\text{obj}}}. \tag{7}$$

Here g_i^{raw} is the pre-clipping training gradient, u_i^{loss} is the signed learning-rate-scaled increment reconstructed from the post-update Adam moments while excluding decoupled weight decay, and L_i is the training loss. PATH accumulates update lengths relative to parameter scale rather than endpoint distance. OBJECTIVE accumulates work magnitude, so the absolute value does not distinguish beneficial from harmful update directions. These are training-interval predictors; the response is measured on held-out probes.

The baseline is deliberately flexible in chronological time. Rich STEP–LR uses identical five-knot cubic splines for elapsed steps and accumulated learning rate, including their interaction,

$$\Phi_{SL} = [\phi_s, \phi_l, \operatorname{vec}(\phi_s \otimes \phi_l)]. \tag{8}$$

A scalar candidate adds only its own spline. We report the relative reduction in held-out squared prediction error,

$$\Delta = \frac{L_{\text{base}} - L_{\text{aug}}}{L_{\text{base}}}. \tag{9}$$

A 10% GAIN therefore means 10% lower prediction error than the fitted baseline; it does *not* mean 10% faster training. Inputs and targets use train-only normalization. Each feature set selects its Ridge strength from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ on validation, and test/confirmation never retune transforms or regularization.

Table 1: Fixed-probe evidence across follow-ups. GAIN is held-out direction/current-P prediction-error reduction over rich STEP–LR. The last column counts how many of the six recorded consistency criteria are met. Original is a post-discovery richer-baseline follow-up locked before fit; Second, Third, and $\beta_1 = 0.85$ are prospective relative to their new outcomes.

Setting	β_1	Candidate	Gain	90% SI	Test	Criteria met
Original	.90	Objective	**9.87%**	[6.51,13.24]%	4/4	**6/6**
Original	.90	Path	5.21%	[2.24,8.17]%	4/4	6/6
Second	.90	Objective	**7.48%**	[7.05,7.59]%	4/4	**6/6**
Second	.90	Path	4.04%	[0.33,7.75]%	3/4	5/6‡
Third	.90	Objective	4.66%	[0.26,6.76]%	2/4	5/6†
Third	.90	Path	**5.01%**	[0.78,9.23]%	4/4	**6/6**
Third seeds	.85	Objective	7.12%	[2.47,9.38]%	3/4	6/6
Third seeds	.85	Path	4.49%	[2.15,6.84]%	4/4	6/6

†Third Objective meets the SI, survival, evidence-group, confirmation-mean, and confirmation-sign criteria; its limiting item is 2/4 positive test bursts. ‡Second Path meets five criteria; its limiting item is 2/4 positive confirmation bursts. Additional longer-memory check ($\beta_1 = 0.95$, same Third seeds): Path **+16.00%** (90% SI [12.74,24.92]%), Objective +3.23% ([-6.30,12.77]%).

2.3 TRAINING, HELD-OUT EVALUATION, AND CONSISTENCY CHECKS

Each environment contains matched training schedules under fixed initialization and data order. We train a randomly initialized 12-layer Pythia-160M architecture (Biderman et al., 2023) (hidden size 768, approximately 162.3M parameters) on a frozen FineWeb-Edu sample (Penedo et al., 2024). Each run uses sequence length 2048, 16 sequences per update, and 1,500 updates (49.152M tokens). AdamW uses $\beta_1 = 0.90$, $\beta_2 = 0.95$, $\varepsilon = 10^{-8}$, weight decay 0.1, global clipping 1.0, and BF16. Each environment contains four paired schedules: 0.5×, 1.0×, and 1.5× cosine schedules, plus a linear alternate.

Train, validation, test, and confirmation refer to the diagnostic response model, not separate language-model phases. A default measurement window contains 64 checkpoints. Sixteen fixed burst positions are split four each across fitting, regularization selection, testing, and confirmation. Probes D0–D3 follow the frozen train/validation/test assignments; C0–C1 are used only for zero-refit confirmation. Each probe contains four fixed 2048-token blocks. Appendix A records the construction and full timeline.

The repeated counts describe consistency across time, lag, and parameter block. Five survival levels define historical lags through β_1^{t-k}; they are not probabilities of successful learning. Seven evidence groups are parameter blocks (embedding, selected attention/MLP layers, and output). Thus "4/4 test", "3/5 survival", and "4/7 groups" count positive gains across held-out positions, lags, and blocks. The latter two are repeated measurements rather than additional independent trials.

The six-part criterion set is a consistency checklist, not a significance test. We resample burst positions while keeping all four paired schedules together. Two thousand pre-generated non-circular moving-block resamples of length two yield a 90% finite-design stability interval (SI), which summarizes stability within the four held-out positions. A comparison satisfies the full criterion set only if the SI lower bound is positive; at least 3/4 test positions, 3/5 survival levels, and 4/7 groups are positive; the confirmation mean is positive; and at least 3/4 confirmation positions are positive. Appendix A reports all components.

The follow-ups deliberately separate replication, changed settings, and robustness. Second and Third change initialization and data order. The $\beta_1 = 0.85/0.95$ checks reuse Third's initialization/order and change first-moment decay, with lags recomputed from the same survival targets. The recorded $\beta_1 = 0.95$ protocol also widens the measurement windows to 91 checkpoints. Original is post-discovery but locked before its richer-baseline fit; Second, Third, and $\beta_1 = 0.85$ were specified before their new outcomes; $\beta_1 = 0.95$ is additional robustness. Architecture and actual-gradient tests retain their recorded primary/secondary roles (Appendices D and E.1).



Figure 2: **Original/Second richer-baseline follow-up.** (a,b) Representative age-seven current-P relevance. Faint lines show all four schedules; bold lines equally average schedules, then average up to five anchors within each window. Gains below use all recorded lags and rich STEP–LR. (c) Objective after Path conditioning and Original-to-Second transfer without refitting, renormalization, or recalibration. Intervals are frozen 90% finite-design stability intervals.

3 STEP AND LEARNING RATE DO NOT FULLY PREDICT GRADIENT RELEVANCE

3.1 REPLICATION UNDER A NEW INITIALIZATION AND DATA ORDER

Second asks whether the gap is tied to a single training realization. Relative to rich STEP–LR, OBJECTIVE reduces held-out prediction error by **9.87%** in Original and **7.48%** in the prespecified Second trajectory (Table 1). All four held-out test positions are positive in both environments, and both comparisons satisfy all six prespecified criteria.

The relationship transfers across environments without refitting. A response model fitted in Original and applied directly to Second—without refitting, renormalization, or recalibration—retains an 8.23% gain. OBJECTIVE also remains informative after conditioning on Relative PATH, with gains of 7.74% and 5.83% in Original and Second. The additional contribution of PATH after OBJECTIVE is smaller (2.84%/2.36%), with the Second interval crossing zero.

Original and Second provide different kinds of evidence. Original is a post-discovery richer-baseline follow-up whose fit specification and criteria were locked before the follow-up outcome; Second is a prespecified replication on a new initialization and data order. We keep those roles explicit rather than treating the two rows as equivalent evidence.

3.2 THE GAP SURVIVES COORDINATE AND MEMORY-TIMESCALE CHANGES

Third asks whether the finding depends on OBJECTIVE as a privileged coordinate. In Third, OBJECTIVE remains positive on average (+4.66%; SI [0.26,6.76]%) but has mixed test-burst signs (2/4 positive), so it meets five of six criteria. Independently prespecified PATH is **+5.01%**, with 4/4 test bursts, 5/5 survival levels, 7/7 evidence groups, and a positive confirmation mean; it meets all six. Because the candidates were prespecified separately, we report them separately. The contrast is the result: non-step structure persists while the scalar that exposes it most stably can change.

The β_1 perturbations ask whether the gap is tied to the default memory timescale. Shortening memory to $\beta_1 = 0.85$ on the Third initialization/data order gives OBJECTIVE +7.12% and PATH +4.49%, both satisfying all six criteria. Lengthening memory to $\beta_1 = 0.95$ produces the strongest PATH effect in the paper: **+16.00%** with SI [12.74,24.92]%. OBJECTIVE remains positive on average (+3.23%) but its interval crosses zero. The fixed-probe gap therefore survives on both sides of the default $\beta_1 = 0.90$ timescale, while the most stable scalar changes. The phenomenon is more stable than the coordinate.

The OPT stress test asks how far the phenomenon carries across architecture. OPT-125M (Zhang et al., 2022) reproduces the gap through PATH; the cross-architecture evidence is therefore carried by PATH rather than the Pythia learning-state summary. The prespecified PATH secondary yields **+4.81%** with SI [+0.15,+9.48]% and satisfies all six criteria. The state-block primary has a positive mean but an interval spanning zero. Appendix E.1 gives the complete audit.

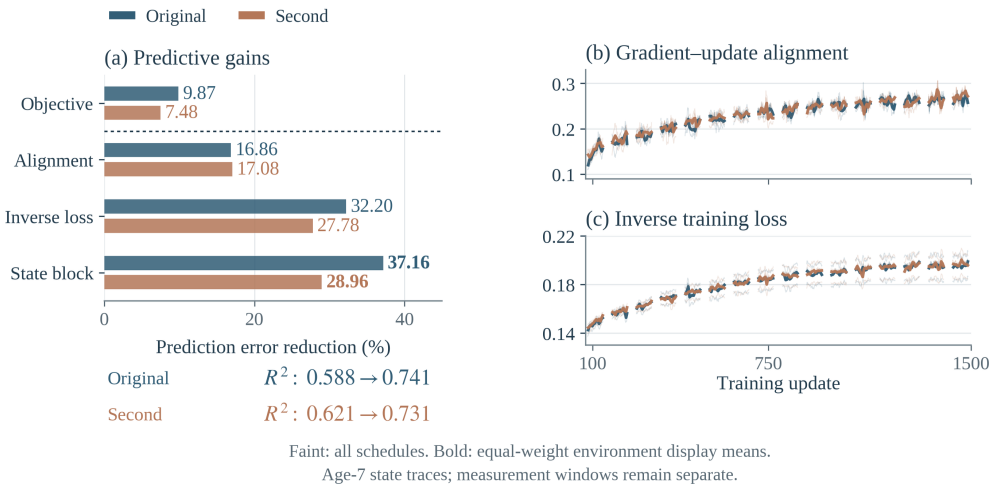

Figure 3: **Objective opens into a larger optimization-state signal.** (a) Gains over rich STEP–LR: the dashed rule separates canonical Objective from post-discovery state components. The state block is strongest; R^2 uses the frozen training-mean predictor. (b,c) Age-seven alignment and inverse-loss traces; faint/bold lines use the display convention of Figure 2. These show state evolution, not predictive-gain trajectories. Artifact controls are in Appendix B.

4 LEARNING STATE EXPLAINS MUCH OF THE MISSING STRUCTURE

Objective factorizes into four quantities that naturally describe optimization state. At a single update,

$$q_i^{\mathrm{obj}} = \frac{\|g_i^{\mathrm{raw}}\| \, \|u_i^{\mathrm{loss}}\| \, |\cos(g_i^{\mathrm{raw}}, u_i^{\mathrm{loss}})|}{L_i + \varepsilon_{\mathrm{obj}}}. \tag{10}$$

This motivates a four-variable learning-state summary (reported as the *State block* in tables) containing interval gradient norm, loss-update norm, mean absolute gradient–update alignment, and mean inverse loss. The decomposition is post-discovery, but its audit plan was locked before the corresponding relevance outcomes were inspected. Appendix B also checks simple LR-normalization and clipping explanations; those artifact features add little, while recomputing OBJECTIVE with clipping included preserves the signal.

The strongest individual terms are state-like. Mean absolute gradient–update alignment gives gains of 16.86% and 17.08% in Original/Second. Mean inverse loss,

$$S_{1/L}(k,t) = \frac{1}{t-k} \sum_{i=k}^{t-1} \frac{1}{L_i + \varepsilon_{\mathrm{obj}}}, \tag{11}$$

reaches 32.20%/27.78%. The full four-variable summary reaches **37.16%/28.96%**. Against a frozen training-mean predictor, held-out R^2 rises from **0.588/0.621** for rich STEP–LR to **0.741/0.731** with the learning-state summary.

Most of the OBJECTIVE signal is already represented in the learning-state summary. Adding frozen OBJECTIVE after that summary reduces held-out prediction error by a further 1.79%/3.43%; the Second stability interval crosses zero. The decomposition therefore points to learning state as the broader source of predictive structure.

5 CAN SIMPLER EXPLANATIONS REMOVE THE STATE SIGNAL?

Three simpler explanations could account for a large post-discovery state gain. The STEP/LR baseline might be underfit; inverse loss might merely track training progress; or temporal autocorrelation might make any nearby state summary look predictive. We test these alternatives using the stored interval table only, without retraining a neural model.

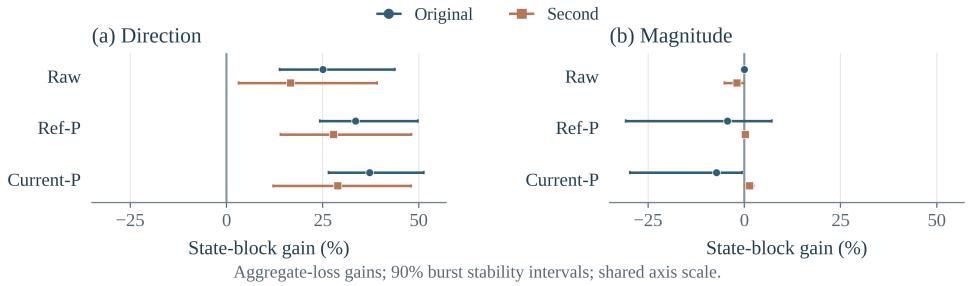

Figure 4: **Post-discovery geometry diagnostic.** Circles denote Original and squares Second. Points show aggregate-loss state-block gains over rich STEP–LR; bars are recorded 90% burst stability intervals. Both panels share an axis scale. Directional gains are positive across raw, reference-P, and current-P geometries; magnitude gains are small, mixed, or interval-unstable.

5.1 MORE STEP/LR CAPACITY DOES NOT ERASE IT

Increasing only the chronological baseline capacity leaves the state signal large. We expand the STEP and LR spline bases from five knots to eight and twelve, keep candidate/state splines at five knots, and preserve the same train/validation/test roles and validation-selected Ridge family. The state-block gain remains above 30% in both environments under both higher-capacity controls, while OBJECTIVE remains positive. The residual is therefore not an artifact of the specific five-knot STEP/LR basis (Appendix C).

5.2 PATH AND TRAINING PROGRESS EXPLAIN MUCH, BUT NOT ALL

Parameter motion and absolute training progress absorb much of the state gain but leave a positive residual. We augment the baseline with Relative PATH and endpoint progress t, thereby providing both how far the parameters moved and where the interval sits in training. The state-block gain falls from **37.16%/28.96%** to **12.94%/4.12%**. Here t is absolute progress, whereas STEP is $t - k$. These overall audit gains are relative reductions of hierarchically averaged losses rather than arithmetic means of the per-burst percentages below.

The residual is positive at every held-out position in both environments. Original's four gains are 18.31%, 4.07%, 8.60%, and 11.97%; Second's are 5.43%, 3.03%, 4.48%, and 2.15%. All eight positions and all nine legal length-two moving-block combinations in each environment are positive.

5.3 PROGRESSIVELY MISALIGN THE STATE, AND THE RESIDUAL ERODES

A locality test keeps the state values fixed while breaking their temporal correspondence. Within each measurement sequence, we circularly shift the four state features together, wrapping values from the end to the start. This preserves the observed values and their marginal distribution while changing which interval each state describes. We then measure predictive value as a function of shift distance.

Predictive value decreases as the state is moved farther from its original interval. We sweep 19 phases $\rho = j/20$, $j = 1, \ldots, 19$, within each trajectory/burst/role/survival/evidence sequence ordered by target step. The four features shift jointly by $\lfloor n\rho + 0.5 \rfloor$, clipped to $[1, n-1]$ for sequence length n. Controls and targets remain aligned, and the aligned regularization strength (0.001 in both environments) is fixed for every shifted fit. Very small shifts retain signal, as expected under temporal smoothness, but the direction-averaged residual generally declines with circular distance. At half phase, **12.94%/4.12%** becomes $-0.34\%/-4.16\%$. Individual shift directions need not decline monotonically.

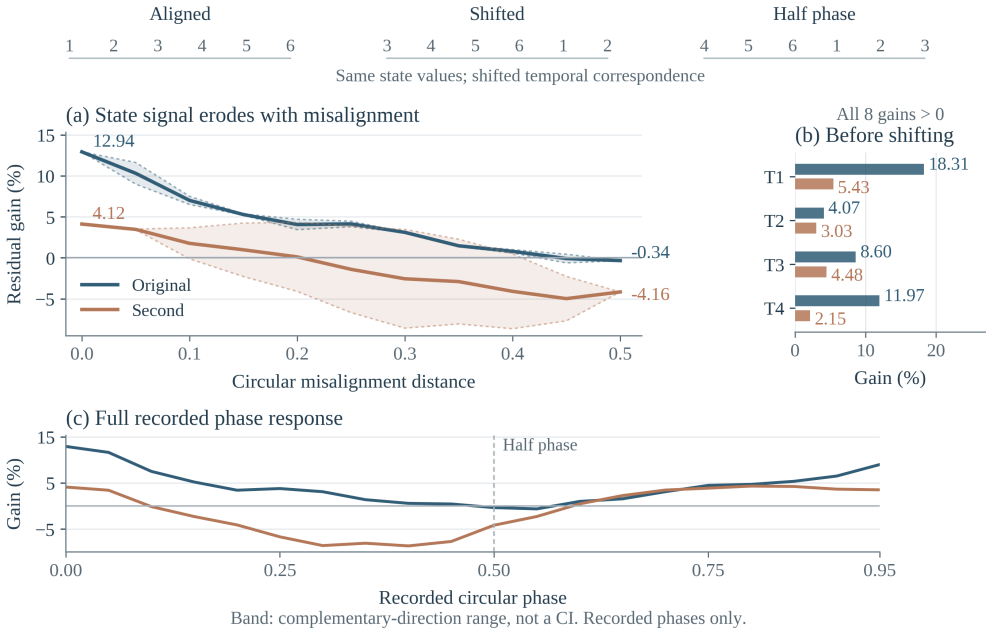

Figure 5: **Misalign the state in time, and the residual erodes.** (a) Curves average complementary shift directions; bands span their gains, not confidence intervals. Half phase is a singleton: $-0.34\%/-4.16\%$. (b) Separately computed test-position gains before shifting. (c) The full recorded phase response before folding. Curves use aggregate-loss reductions (Appendix F), not arithmetic means of the gains in (b).

6 Actual Gradients Add Variation Present at Birth

Fixed probes isolate field evolution, but actual training gradients also vary with the minibatch that produced them. A schematic fixed-probe description is

$$R_{k,t} = F(t-k, \text{optimization state}). \tag{12}$$

The prospective actual-gradient bridge restores minibatch variation by replacing the historical fixed-probe vector with the clipped training gradient $\tilde{g}_k$ that entered AdamW, while retaining a held-out probe gradient at the current endpoint. The response remains normalized current-P alignment, not realized long-horizon loss improvement.

Birth alignment measures how the training gradient initially related to the held-out field. It is the signed alignment between the held-out probe gradient and clipped training gradient at checkpoint k, both measured in the pre-update P_k geometry (Appendix F). The prespecified secondary adds this birth-alignment feature to Step–LR before testing the same learning-state summary. "Birth-conditioned" refers to this change in regression baseline: birth alignment is added before the same learning-state summary is evaluated. The implementation name `birthP` is given in Appendix F.

Conditioning on birth alignment reveals a positive secondary effect that is absent unconditionally. The birth-conditioned secondary yields **+4.83%** average gain, SI [3.20,5.73]%, with 4/4 positive test positions and 6/7 positive groups, although survival-level support is 2/5. The distinct unconditional primary yields -0.89%. Appendix D reports both comparisons, confirmation outcomes, and the prospectively recorded source-environment change to Third.

The bridge explains one way fixed-probe and actual-gradient relevance can diverge. An actual gradient inherits minibatch-dependent alignment at birth before any subsequent state exposure occurs. The unconditional primary remains negative, while the positive birth-conditioned secondary has mixed support across survival levels; a train-defined descriptive stratification finds the strongest residual at intermediate birth alignment (Appendix D).

7 Related Work

Adaptive-memory optimizers change retention; our question precedes retention design. Adam and AdamW use exponential moment memories (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). AdEMAMix mixes fast and slow EMAs (Pagliardini et al., 2025); Adaptive Memory Momentum adjusts momentum online (Topollai & Choromanska, 2026); and Activation-Keyed Momentum makes forgetting direction-aware (Hong & Qu, 2026), extending earlier multi-timescale or adaptive-forgetting variants (Lucas et al., 2019; Graepel & Schraudolph, 2002). These methods modify retention; we measure the coordinate that defines effective age.

Gradient-staleness work changes how delay is represented. Faster Asynchronous SGD uses moving averages of gradient statistics instead of elapsed updates (Odena, 2016), while Gap-Aware mitigation uses endpoint parameter discrepancy to penalize stale gradients (Barkai et al., 2020). Our PATH instead accumulates scale-normalized update lengths along the interval. Overshoot evaluates gradients at momentum-shifted weights so that the resulting direction is intended to remain useful for future updates (Kopal et al., 2025).

Recent analyses study different consequences of optimizer memory. Sweeney (2026) studies shuffle-order effects under fixed-clock memory, while Liu and Li (2026) trace delayed loss effects of localized minibatch perturbations in AdamW. Our contribution is upstream: with the optimizer fixed, we measure whether elapsed optimizer time is sufficient for historical relevance and what trajectory state predicts the residual.

Our state summaries are predictors, not curvature measurements. Adaptive Edge-of-Stability work studies the preconditioned Hessian and its stability threshold (Cohen et al., 2022); our response is historical alignment with the current gradient field. The inverse-loss and alignment summaries therefore do not identify an edge-of-stability mechanism.

8 Limitations

The present evidence is limited to small decoder-only Transformers (Pythia-160M and OPT-125M), one data regime, 1,500-update horizons, and AdamW. First-moment stresses vary β_1 while β_2, clipping, weight decay, and other settings remain fixed. They test sensitivity to the memory timescale, not scale or cross-optimizer universality. Broader probe families, longer training, and other optimizer settings remain to be evaluated.

The uncertainty quantification is also finite-design. Each environment has four held-out burst positions; SIs summarize this finite design, and the consistency checklist is neither a population guarantee nor a multiplicity correction. Higher-knot controls test sensitivity within the spline family. The state decomposition remains post-discovery, OPT support varies by predictor, actual-gradient predictivity depends on birth conditioning, and whole-schedule extrapolation has a repeatable low-LR boundary (Appendix C). These constraints define the scope of the measurement result.

9 Conclusion

One optimizer update is one unit of AdamW's memory age, but it is not a complete unit of gradient-field change. OBJECTIVE exposes a reproducible predictive gap beyond STEP–LR in Original/Second and transfers without refitting. New trajectories and shorter or longer first-moment memory preserve the gap, even though the most stable scalar can change. A prespecified OPT PATH comparison provides a second-architecture check.

The larger signal is learning state, with birth alignment adding a complementary source of variation for actual gradients. On Pythia, a four-variable state summary explains substantially more predictive structure; parameter motion and training progress absorb part of it, and temporal misalignment erodes the locally aligned remainder. The actual-gradient bridge further points to birth alignment as a conditioning variable. **At a descriptive level, fixed-timescale optimizer memory assigns the same age to training histories that can have different present-day gradient relevance. Measuring that mismatch provides a concrete basis for testing state-aware memory rules.**

AI USE STATEMENT

Generative AI tools were used extensively as research-assistance tools, including for brainstorming and refinement of hypotheses, feedback on experimental methodology, assistance with code implementation and debugging, interpretation and stress-testing of experimental results, and drafting and editing of the manuscript. All experimental outputs, numerical results, citations, analyses, and AI-assisted scientific suggestions remain the responsibility of the author(s), who take full responsibility for the methodology, claims, results, and final content of the paper.

REPRODUCIBILITY STATEMENT

Appendix A records probe construction, burst assignments, lag mapping, transforms, fitting, aggregation, and all consistency components. Appendices B–E.1 give the decomposition, stress tests, actual-gradient comparisons, full OPT audit, and analysis chronology. Numerical evidence is reported with its original prospective or post-discovery status. All training, measurement, and analysis code developed for this study will be released as open source, together with experiment configurations and scripts used to generate the reported figures and tables. Repository link: https://anonymous.4open.science/r/gradient-age-repro-D17D/.

REFERENCES

S. Barkai, I. Hakimi, and A. Schuster. Gap-aware mitigation of gradient staleness. ICLR, 2020.

S. Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. ICML, 2023.

J. M. Cohen et al. Adaptive gradient methods at the edge of stability. arXiv:2207.14484, 2022.

T. Graepel and N. N. Schraudolph. Stable adaptive momentum for rapid online learning in nonlinear systems. ICANN, 2002.

E. Hong and G. Qu. Activation-keyed momentum: An anisotropic momentum update via the delta rule. arXiv:2608.19491v2, 2026.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. ICLR, 2015.

J. Kopal, M. Gregor, S. de Leon-Martinez, and J. Simko. Overshoot: Taking advantage of future gradients in momentum-based stochastic optimization. arXiv:2501.09556, 2025.

K. Liu and S. Li. Finite-horizon input-output dynamics of minibatch perturbations in AdamW. arXiv:2608.19762, 2026.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. ICLR, 2019.

J. Lucas, S. Sun, R. S. Zemel, and R. Grosse. Aggregated momentum: Stability through passive damping. ICLR, 2019.

A. Odena. Faster asynchronous SGD. arXiv:1601.04033, 2016.

M. Pagliardini, P. Ablin, and D. Grangier. The AdEMAMix optimizer: Better, faster, older. ICLR, 2025.

G. Penedo et al. The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv:2406.17557, 2024.

J. Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise. arXiv:2606.29554, 2026.

K. Topollai and A. E. Choromanska. Adaptive memory momentum via a model-based framework for deep learning optimization. AISTATS, 2026.

S. Zhang et al. OPT: Open pre-trained transformer language models. arXiv:2205.01068, 2022.

A Frozen Measurement Protocol and Criteria Audit

This appendix preserves the recorded protocol, comparison roles, and analysis chronology.

A.1 Probe, survival, evidence, and response construction

Probes. Six deterministic held-out probes are frozen before the formal outcomes: D0–D3 for train/validation/test and C0–C1 only for zero-refit confirmation. Each consists of four fixed 2048-token blocks selected by manifest order from its held-out document pool. Token hashes are recorded in the probe manifest; Table 2 gives the frozen assignments.

Survival mapping. For first-moment decay β_1, lag ℓ has chronological survival β_1^ℓ. Target levels $\{0.8, 0.5, 0.2, 0.05, 0.01\}$ are mapped by

$$\ell(s) = \left\lceil \frac{\log s}{\log \beta_1} \right\rceil, \qquad s_{\text{actual}} = \beta_1^{\ell(s)}. \tag{13}$$

The pair log retains both target and realized survival. The $\beta_1 = 0.85/0.95$ checks recompute lags from the same targets rather than reuse default-β_1 lags.

Evidence groups. E0 is the embedding (`gpt_neox.embed_in`); E1/E2/E3 are attention at layers 1/5/10; E4/E5 are MLP at layers 1/10; E6 is the LM head (`embed_out`). The consistency threshold is four of seven positive groups. The OPT-specific E6 mapping appears in Appendix E.1.

Transforms, fits, and aggregation. Scalar features use train-only weighted median/IQR normalization and five-knot cubic splines; targets use train-only weighted mean/standard deviation. Response coefficients are group-specific. One Ridge strength per endpoint/geometry/feature set is shared across groups and chosen on validation from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Test and confirmation never retune these choices. Standardized-target cell losses are averaged equally over survival levels and groups within each run/burst, converted to relative reductions, and then averaged equally over paired runs and held-out burst positions.

Post-discovery state summaries. The state block combines the two logged interval norm summaries with

$$S_{\cos}(k,t) = \frac{1}{t-k} \sum_{i=k}^{t-1} |\cos(g_i^{\text{raw}}, u_i^{\text{loss}})|, \tag{14}$$

$$S_{1/L}(k,t) = \frac{1}{t-k} \sum_{i=k}^{t-1} \frac{1}{L_i + \varepsilon_{\text{obj}}}. \tag{15}$$

The Original/Second decomposition is post-discovery; the later OPT state-block primary was specified before OPT outcomes.

A.2 Frozen burst timeline

The default schedule has 64 checkpoints per window; the $\beta_1 = 0.95$ amendment is detailed in Appendix F.

A.3 Frozen richer-baseline criteria audit

A formal follow-up satisfies the full prespecified criterion set only if all six recorded conditions hold: SI lower bound > 0; test sign consistency $\geq 3/4$; at least 3/5 survival levels positive; at least 4/7 evidence groups positive; confirmation mean > 0; and confirmation sign consistency $\geq 3/4$. Original is post-discovery relative to the initial discovery but was locked before its richer-baseline fit; Second, Third, and $\beta_1 = 0.85$ are prospective relative to their new outcomes.

Table 2: Frozen burst timeline. Step ranges are inclusive. Roles are Tr = train, V = validation, Te = test, C = confirmation. Probe assignments are frozen with the split.

Pos.	Steps	Role	Probe	Pos.	Steps	Role	Probe
0	75–138	Tr	D0	8	795–858	Tr	D2
1	165–228	V	D1	9	885–948	V	D3
2	255–318	Te	D2	10	975–1038	Te	D0
3	345–408	C	C0	11	1065–1128	C	C0
4	435–498	Tr	D1	12	1155–1218	Tr	D3
5	525–588	V	D2	13	1245–1308	V	D0
6	615–678	Te	D3	14	1335–1398	Te	D1
7	705–768	C	C1	15	1425–1488	C	C1

Table 3: Recorded richer-baseline criteria audit for direction/current-P. Gain, SI lower bound, and confirmation mean are percentages. All settings use $\beta_1 = 0.90$ except the last two rows, which reuse Third initialization/data order at $\beta_1 = 0.85$.

Setting / candidate	Gain	SI low	Test	Surv.	Groups	Conf. mean	Conf. sign / criteria
Original / Objective	**9.87**	6.51	4/4	4/5	7/7	10.32	4/4; **6/6**
Original / Path	5.21	2.24	4/4	5/5	7/7	6.58	4/4; 6/6
Second / Objective	**7.48**	7.05	4/4	5/5	7/7	12.92	3/4; **6/6**
Second / Path	4.04	0.33	3/4	4/5	7/7	2.93	2/4‡; 5/6
Third / Objective	4.66	0.26	2/4†	3/5	4/7	3.93	3/4; 5/6
Third / Path	**5.01**	0.78	4/4	5/5	7/7	5.04	3/4; **6/6**
Third .85 / Objective	7.12	2.47	3/4	4/5	7/7	6.04	3/4; 6/6
Third .85 / Path	4.49	2.15	4/4	5/5	7/7	4.88	4/4; 6/6

†Third Objective: test sign 2/4 is limiting. ‡Second Path: confirmation sign 2/4 is limiting. Both meet five of six criteria. Additional $\beta_1 = 0.95$ robustness: Objective +3.23%, SI [-6.30,12.77]%; Path **+16.00%**, SI [12.74,24.92]%.

A.4 Why the Second-environment stability interval is narrow

The narrow SI is a property of the frozen block-resampling design, not a large-sample confidence interval. Second-environment Objective gains at the four test positions are 1.746%, 13.002%, 1.103%, and 14.074%. Under the frozen non-circular block length of two, the three contiguous block means are 7.374%, 7.053%, and 7.589%. The nine legal two-block resample means are their pairwise averages,

$$\begin{bmatrix} 7.374 & 7.213 & 7.481 \\ 7.213 & 7.053 & 7.321 \\ 7.481 & 7.321 & 7.589 \end{bmatrix} \%. \tag{16}$$

Thus the finite-design resampling support itself lies near [7.05,7.59]%. It summarizes contiguous-block stability rather than dispersion across the four individual burst gains.

Table 4: Confirmation-position outcomes. Values are paired held-out gain (%). The frozen confirmation-sign threshold is at least three of four positive positions.

Environment	Candidate	Conf.1	Conf.2	Conf.3	Conf.4	Mean
Original	Objective	9.90	5.81	0.16	25.40	10.32
Original	Path	1.97	6.61	4.11	13.62	6.58
Second	Objective	-3.16	22.51	9.74	22.59	12.92
Second	Path	4.18	-4.38	-1.03	12.95	2.93

B AUDIT DETAILS

The audit reconstructs the same frozen predictions with a different aggregation order. It was locked before outcome inspection and reconstructs OBJECTIVE exactly from 99,008 pair rows per environment. Canonical estimates average run-by-burst relative reductions; the audit averages absolute losses before forming the relative reduction. This gives 10.37%/6.86% versus the canonical 9.87%/7.48%. These descriptive audit estimates do not replace Table 1; the later actual-gradient extension is reported separately in Appendix D.

Table 5: Post-discovery audit for direction/current-P. Values are gains over the listed base; SI is the audit 90% stability interval.

Augmentation	Orig. gain	Orig. SI	Second gain	Second SI
Objective / LR	-1.80%	[-7.21,5.48]%	0.24%	[-4.47,3.96]%
Clip exposure	1.33%	[-0.59,6.04]%	1.29%	[-2.91,10.25]%
Clipped Objective	10.80%	[9.50,12.56]%	7.00%	[6.38,7.78]%
Grad–update alignment	16.86%	[7.05,30.13]%	17.08%	[5.86,29.72]%
Inverse-loss state	32.20%	[24.51,43.11]%	27.78%	[13.46,43.55]%
State component block	**37.16%**	[26.50,51.29]%	**28.96%**	[12.15,48.02]%
Objective after state	1.79%	[0.83,4.95]%	3.43%	[-1.91,4.23]%

Base: rich Step–LR, except the final row, which uses rich Step–LR plus the state block.

C VERIFIED OFFLINE ROBUSTNESS AND STRESS TESTS

All analyses in this section use the exported 198,016-row interval table with no GPU replay. The CSV export was checked against the frozen Parquet source with row order, column order, null masks, and all floating-point values preserved exactly.

The Path+t residual is positive under every frozen held-out block combination. All four test-position gains are positive in both environments, and all nine legal length-two moving-block means are positive. The systematic temporal check uses 19 common circular phases (5%–95%) and freezes the aligned $\alpha = 10^{-3}$ before refitting shifted response coefficients. Tiny shifts retain information, while larger circular distances erase most of the residual; half phase gives $-0.34\%/-4.16\%$.

Whole-schedule extrapolation has a reproducible low-LR boundary. Leave-one-trajectory-out tests show the same pattern in all three $\beta_1 = 0.90$ environments: the 0.5× cosine holdout has negative mean gain, while 1.0×, 1.5×, and linear holdouts are positive. Ordered as 0.5×/1.0×/1.5×/linear, gains are -42.35/34.81/60.03/22.38% in Original, -10.43/32.00/53.98/20.26% in Second, and -18.51/37.86/54.53/31.96% in Third. A feature-support diagnostic shows larger nearest-neighbor and Mahalanobis distances for 0.5× than 1.0×/linear, but 1.5× is comparably separated and still transfers positively; support shift alone therefore does not explain the asymmetry. This identifies low-LR whole-schedule extrapolation as a boundary of the pooled prespecified claim.

Reciprocal transfer and cross-trajectory matching provide additional descriptive checks. In the reconstructed interval-table pipeline, no-refit transfer gives 8.07% for Original→Second and 10.89% for Second→Original, with the latter positive at all four target test bursts. Exact (k, t) cross-trajectory pairs passing lower-quartile cutoffs on absolute LR/Path differences yield inverse-loss-oriented relevance differences of +0.0127/+0.0126, with 3/4 positive bursts in each environment.

C.1 SIGNED WORK, NUMERICAL CONSTANTS, AND EVIDENCE BLOCKS

The stored logs also permit direct checks of several implementation choices tied to the state decomposition. For signed work, define the descent-work coordinate

$$q_i^{\mathrm{signed}} = -\frac{(g_i^{\mathrm{raw}})^\top u_i^{\mathrm{loss}}}{L_i + \varepsilon_{\mathrm{obj}}}. \tag{17}$$

All 12,000 logged Original/Second update terms are positive under this sign convention, and their 198,016 stored interval sums satisfy $A_{k,t}^{\mathrm{signed}} = A_{k,t}^{\mathrm{obj}}$ exactly (maximum absolute difference 0). The

Table 6: Post-discovery robustness and stress tests. All values are gains (%). Objective-stress values come from the audit reconstruction and do not replace the canonical frozen 9.87%/7.48% estimates.

	Objective stress		State stress	
Step/LR basis	Original	Second	Original	Second
5-knot richSL	10.37	6.86	37.16	28.96
8-knot Step/LR	10.55	6.85	36.28	30.23
12-knot Step/LR	9.01	3.29	36.66	31.63
State-only local-residual controls				
Rich Step–LR + Path + t	**12.94 / 4.12**			
Common phase 0.50 (fixed α)	-0.34 / -4.16			

literal no-absolute dot-product coordinate is therefore only the global sign reversal of the same scalar on these trajectories, so the signed-work comparison gives the same fitted prediction result. On these trajectories, sign therefore adds no independent variation to this diagnostic.

The state result is also insensitive to numerical epsilons over ranges far wider than those used in the paper. Re-fitting the same stored-data pipeline while varying $\varepsilon_{\mathrm{norm}}$ from 10^{-15} to 10^{-3} keeps the state-block gain within 37.114–37.160% in Original and 28.880–28.957% in Second. Reconstructing inverse-loss features from the per-step logs while varying $\varepsilon_{\mathrm{obj}}$ from 0 to 10^{-2} gives 37.160–37.164% and 28.957–28.960%, respectively. The selected Ridge strengths are unchanged across these checks.

The signal is not concentrated in a single prespecified evidence block. At the default settings, all seven group-specific test gains are positive: 19.84–58.81% in Original and 10.59–58.93% in Second. Removing one evidence group at a time from the held-out aggregation leaves gains of 34.18–39.71% and 26.18–31.86%. These checks use the seven evidence blocks of the response design; finer per-tensor regrouping would require a new measurement pass.

Table 7: Additional stored-data sensitivity checks for the post-discovery state block. Ranges report held-out gain (%) over rich Step–LR. No neural model is retrained.

Check	Original	Second
$\varepsilon_{\mathrm{norm}} \in [10^{-15}, 10^{-3}]$	37.114–37.160	28.880–28.957
$\varepsilon_{\mathrm{obj}} \in [0, 10^{-2}]$	37.160–37.164	28.957–28.960
Per-evidence-group gains (7 groups)	19.84–58.81	10.59–58.93
Leave-one-group-out aggregate	34.18–39.71	26.18–31.86

D Prospective Actual-Gradient Boundary Test

The bridge replaces the historical fixed-probe vector with the clipped training gradient that actually entered AdamW. In the Third environment, the response at anchor t is the normalized current-P first-order alignment

$$U^{\mathrm{AG}}_{k,t,p,e} = \frac{(g^{(e)}_{p,t})^\top P^{(e)}_t \tilde{g}^{(e)}_k}{\sqrt{(g^{(e)}_{p,t})^\top P^{(e)}_t g^{(e)}_{p,t}}\sqrt{(\tilde{g}^{(e)}_k)^\top P^{(e)}_t \tilde{g}^{(e)}_k} + \varepsilon_{\mathrm{metric}}}. \tag{18}$$

This is a normalized local first-order alignment proxy, not a long-horizon training benefit. Before any bridge outcome, the planned Second clean-room replay was prospectively changed to Third to match the fresh trajectory family of the fixed-probe extension. The endpoint, state block, splits, six-part consistency criteria, and distinct unconditional primary / birth-conditioned secondary remained unchanged.

The bridge shows conditional rather than unconditional transfer of the fixed-probe state pattern. The unconditional locality control gives aligned gain -0.89%, half-phase gain -17.57%, and a reported difference of +16.69 percentage points; it tests sensitivity to temporal correspondence but does not change the primary decision. The positive birth-conditioned secondary suggests an interaction between birth alignment and subsequent state exposure, with mixed support across survival strata.

Table 8: Actual-gradient boundary test. Both comparisons add the same state block. `birthP` conditions on held-out alignment at gradient birth.

Quantity	Unconditional primary	Prespecified secondary
Baseline	Rich Step–LR	Rich Step–LR + birthP
Added features	State block	State block
Mean gain	**-0.89%**	**+4.83%**
90% SI	[-4.74,1.52]%	[3.20,5.73]%
Positive test positions	1/4	4/4
Positive survival levels	2/5	2/5
Positive evidence groups	3/7	6/7
Confirmation mean	-7.11%	+0.43%
Positive confirmation positions	1/4	3/4

D.1 Birth-alignment stratification

To examine that interaction without fitting separate models inside favorable subsets, we keep the globally fitted birth-conditioned baseline and state-augmented models fixed. Low/middle/high birth-alignment strata are defined by train-only tertiles separately within each evidence group and then applied unchanged to test and confirmation rows. This avoids conflating birth-alignment scale with evidence-group identity. The resulting gains are descriptive held-out comparisons of the same two fitted models.

Table 9: State-block gain conditional on similar birth alignment. Strata use train-only within-evidence-group tertiles. The interval is the 90% moving-block range obtained from the frozen test-burst resamples; no model is refit within a stratum.

Birth stratum	Test gain	90% interval	Positive test	Confirmation gain
Low	+3.60%	[-1.44,8.64]%	3/4	-0.22%
Middle	**+10.86%**	**[4.94,16.79]%**	**4/4**	**+9.38%**
High	+5.71%	[0.45,10.97]%	4/4	+5.74%

The pattern is non-monotone: the additional state signal is strongest in the middle birth-alignment stratum, remains positive in the high stratum, and is weaker in the low stratum. Thus birth condition modulates how much predictive structure remains for subsequent state exposure, rather than acting only as a uniform offset. The stratification is descriptive; the bridge result remains the +4.83% birth-conditioned comparison above.

E Reproducibility Boundary

Recorded decisions are not retroactively changed by later analysis or presentation. Post-discovery analyses remain post-discovery; the Original richer-baseline follow-up remains post-discovery relative to the initial discovery; prospective Second/Third/changed-setting/bridge outcomes retain their recorded statuses. The archive contains the frozen protocol/configuration, burst plan and split, probe manifest with token hashes, pre-generated resampling indices, validation/replication manifests, and SHA-256 manifests. These provenance records establish chronology; they do not turn later audits or stress tests into prespecified hypotheses.

Optimizer interventions require a separate experiment. The current evidence supports non-chronological fixed-probe relevance; whether a state-dependent forgetting rule improves training is a distinct intervention question.

E.1 OPT-125M cross-architecture stress

The OPT stress was specified before OPT training. The state block was the primary comparison, with Path and Objective as prespecified secondaries. All used the fixed-probe endpoint, rich Step–LR

baseline, and unchanged six-part checklist. OPT was randomly initialized with its native tokenizer; shared embedding/LM-head parameters were counted once, and E6 used the final LayerNorm. Training and analysis reused the same functions.

Table 10: Complete OPT-125M comparison. All three prespecified comparisons are retained. SI is a 90% finite-design stability interval.

Quantity	State	Path	Objective
Role	Primary	Secondary	Secondary
Mean gain	+14.80%	**+4.81%**	+3.05%
90% SI	[-3.82,33.41]%	[+0.15,+9.48]%	[-3.11,+9.22]%
Positive test positions	3/4	3/4	1/4
Positive survival levels	2/5	5/5	3/5
Positive evidence groups	5/7	7/7	2/7
Confirmation mean	+13.69%	+5.44%	+6.96%
Positive confirmation positions	4/4	4/4	4/4
Criteria met	4/6	**6/6**	3/6

The OPT result reproduces the Step–LR gap through PATH while showing that the Pythia state summary is not architecture-invariant. PATH meets all six criteria. The state primary has a positive mean but an SI spanning zero and only 2/5 positive survival strata; OBJECTIVE also has mixed component support. No architecture-specific redesign of the state block is reported.

Table 11: Analysis chronology and evidence status. The table records when each analysis layer was defined relative to the outcomes it is used to interpret.

Stage	Status	Timing / lock
Original richer-baseline follow-up	Post-discovery follow-up	Defined after the initial discovery, then locked before the richer-baseline fit and follow-up outcome.
Second trajectory validation	Prespecified replication	Same candidate/criteria family applied to an independent initialization and data order.
Third new-seed environment	Prespecified extension	Inherited pipeline frozen before the new trajectory outcomes.
$\beta_1 = 0.85$ changed setting	Prespecified extension	Same Third initialization/data order; only first-moment decay is changed.
$\beta_1 = 0.95$ longer-memory check	Additional robustness	Same Third initialization/data order; decay 0.95, recomputed lags, and a recorded 91-checkpoint window amendment.
Actual-gradient bridge	Prespecified boundary test	Unconditional primary and birth-conditioned mechanistic secondary recorded separately; primary decision unchanged.
State decomposition audit	Post-discovery interpretation	Audit plan locked before audit relevance outcomes were inspected.
Offline robustness / stress tests	Post-discovery stress test	Defined after the primary result; phase sweep keeps aligned regularization fixed across misalignments.
OPT-125M architecture stress	Prespecified architecture stress	Specified before OPT training: State primary; Path and Objective secondaries; unchanged criteria and analysis functions.

F IMPLEMENTATION AND AGGREGATION DETAILS

Loss-driven update and birth alignment. Update i takes θ_i to θ_{i+1}. The observer records the pre-clipping gradient g_i^{raw}, clips once globally, executes AdamW, and reconstructs the signed loss-only

increment from the updated moments,

$$u_i^{\mathrm{loss}} = -\frac{\eta_i}{1-\beta_1^{i+1}} \frac{m_{i+1}}{\sqrt{v_{i+1}/(1-\beta_2^{i+1})} + \varepsilon_{\mathrm{adam}}}. \tag{19}$$

Operations are coordinatewise. Thus u_i^{loss} includes the learning rate and excludes $-\eta_i \lambda \theta_i$, the decoupled decay term. The observer checks the reconstructed sum against the measured parameter increment. Parameters, gradients, and moment buffers use FP32; the forward pass uses BF16 autocast with gradient scaling disabled.

The birth feature uses the held-out probe and clipped training gradients before update k in the pre-update geometry,

$$B_{k,p,e}^{\mathrm{birthP}} = \frac{(g_{p,k}^{(e)})^\top P_k^{(e)} \tilde{g}_k^{(e)}}{\|g_{p,k}^{(e)}\|_{P_k^{(e)}} \, \|\tilde{g}_k^{(e)}\|_{P_k^{(e)}} + \varepsilon_{\mathrm{metric}}}. \tag{20}$$

This signed feature receives train-only normalization and a spline in the secondary baseline; it does not replace the current-P_t response in Appendix D.

Geometry and squared-error scale. Reference-P uses the fixed `lr_1p0` snapshot at completed update 750; changed settings construct their own reference. Magnitude is the signed natural log-ratio

$$\log \frac{\|g_{p,k}\|_P + \varepsilon_{\mathrm{metric}}}{\|g_{p,t}\|_P + \varepsilon_{\mathrm{metric}}}, \qquad P \in \{I, P_{\mathrm{ref}}, P_t\}. \tag{21}$$

For group e, the diagnostic target is $z_j = (R_j - \mu_{\mathrm{tr},e})/(\sigma_{\mathrm{tr},e} + \varepsilon_{\mathrm{norm}})$ and the observation loss is $(z_j - \hat{z}_j)^2$. The audit reports $R^2 = 1 - L_{\mathrm{model}}/L_{\mathrm{constant}}$, where the constant is the frozen group-specific training mean (zero after standardization), not the test mean. All losses in this comparison use the same hierarchical averaging.

Overall versus per-burst audit gains. Let B_b and A_b be baseline and augmented squared errors after averaging within cells and equally over survival levels, evidence groups, and runs in burst b. The post-discovery Path/progress analysis uses

$$d_b = \frac{B_b - A_b}{B_b}, \qquad G_{\mathrm{audit}} = \frac{\sum_b B_b - \sum_b A_b}{\sum_b B_b} = \sum_b \frac{B_b}{\sum_j B_j} d_b. \tag{22}$$

Thus the overall 12.94%/4.12% and the four per-burst percentages in Figure 5 are not related by a simple arithmetic mean. Saved overall error pairs are 0.258209/0.224794 (Original) and 0.215228/0.206353 (Second). The locality plot folds phases ρ and $1-\rho$: their mean is the point and their range is the vertical bar; half phase is a singleton.

Learning-rate configuration and longer-memory windows. Frozen schedules use 75 warmup updates. Cosine peak/minimum rates are $1.5 \times 10^{-4}/1.5 \times 10^{-5}$, $3 \times 10^{-4}/3 \times 10^{-5}$, and $4.5 \times 10^{-4}/4.5 \times 10^{-5}$; the linear alternate uses the middle pair. The recorded $\beta_1 = 0.95$ amendment uses checkpoint windows $[60 + 90b, 151 + 90b)$ for $b = 0, \ldots, 15$, rather than the default 64-checkpoint windows. Lags [5,14,32,59,90] then have [86,77,59,32,1] pairs per cell. Roles and probes remain assigned by burst ID; histories stay inside each window. Adjacent windows share a boundary checkpoint, and θ_{1500} is measured without an extra training update. This is longer-memory robustness under a modified measurement schedule, not a measurement-matched estimate of the effect of β_1 alone. The frozen configuration uses $\varepsilon_{\mathrm{adam}} = 10^{-8}$ and $\varepsilon_{\mathrm{metric}} = \varepsilon_{\mathrm{path}} = \varepsilon_{\mathrm{obj}} = \varepsilon_{\mathrm{norm}} = 10^{-12}$.