Reference Accuracy and Predictive Value in Low-Precision Training

Anonymous authors
Paper under double-blind review

Abstract

Direct low-precision write-back can erase nonzero optimizer proposals. A high-precision reference can evaluate that event on a projected state, but aggregate agreement need not imply improvement over simpler controls. In a controlled grid, 55/72 cells have measured and predicted post-initialization crossings spanning $384\times$; 52/55 lie within 15%, while four cells disagree on crossing category. A reused-source NeoX audit passes temporal skill against a constant, yet an outcome-blind secondary comparison favors a historical template (RMSE 0.00360) over the corrected source (0.00438). A prospective Pythia-410M checkpoint-restart stage evaluates 15 holdout targets across three ages. Its arithmetic-margin selector meets study-specific absolute error/coverage requirements on every target but has higher error than always predicting unchanged on the same accepted events. Every unit-gain source correction also loses to its template. Weak residual alignment limits even a separate hindsight-optimal scalar gain to less than 3.4% RMSE improvement per observed comparison, below the fixed 10% requirement; this is a post-outcome geometric result. The complete original protocol already rejects the forecast, and finalization includes a disclosed analysis-binding repair. Matched rounding interventions separately show large fixed-run policy contrasts. These empirical results characterize this reference construction and its controls, rather than a general failure of reference forecasting.

1 Introduction

Blanket fp32 master weights accumulate sub-grid changes (Micikevicius et al., 2018). Nearest rounding can instead leave a stored code unchanged after a nonzero proposal; stochastic rounding, compensated accumulation, and optimizer error feedback are established responses (Gupta et al., 2015; Zamirai et al., 2020; Connolly et al., 2021; Ozkara et al., 2025; Nikdan et al., 2026). Recent fp8 systems extend low precision to gradients, optimizer states, activations, or distributed communication (DeepSeek-AI, 2024; Peng et al., 2023; Xi et al., 2025); fp4 systems extend quantization to native linear layers and backward/update paths (Tseng et al., 2025; Castro et al., 2025; Chmiel et al., 2025; Panferov et al., 2026). We isolate only destination-grid write-back and ask a narrower measurement question: what can a high-precision schedule-pilot trace say before the corresponding low-precision run?

Prior work identifies, analyzes, or exploits update invisibility (Zamirai et al., 2020; Xia et al., 2024; Yu, 2026; Miahi & Belilovsky, 2026; Liang et al., 2026). The open empirical question here is what an offline reference establishes about another trajectory. Three claims require different evidence: the rounding event is correct on its supplied state; the projected reference matches an aggregate target statistic; and the reference improves prediction beyond information already available from simpler controls. Exact code equality establishes the first claim, and measured aggregate agreement establishes the second. Neither establishes incremental value beyond the controls.

We test this distinction from controlled onset prediction to unseen targets. Our sealed protocols are study-specific designs fixed before their corresponding new outcomes; amendments and post-outcome repairs are labeled separately. In Pythia-410M, 96.45–99.77% of accepted events are unchanged, making the constant-unchanged label a strong control. The selector satisfies absolute-risk and coverage requirements on all 15 holdouts, yet loses to this control on the identical accepted events. The complete original protocol also tests the source against history and correctly returns a negative

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.

decision. Thus our counterexample concerns inferring baseline-relative skill from an absolute-risk component; it is not a failure of the original joint protocol or of selective-classification theory. The temporal comparison further finds that even hindsight scalar rescaling cannot supply the fixed material gain along these observed source residuals. These empirical limits motivate the focus on what a reference adds, beyond how closely it agrees.

Our contributions are:

1. An explicit separation of exact coordinate replay from a conditional aggregate reference proxy, with controlled timing tests and finite-trace counterexamples that expose the missing pathwise assumptions.
2. Prospective evidence for this selector that low error with high coverage can coexist with worse event predictions than a constant label, and that a fixed unit-gain source correction can worsen a historical temporal forecast. We report both original and amended decisions, including all negative targets and dependent template variants. The empirical diagnosis is weak source-residual alignment: even target-wise hindsight scalar adjustment explains at most 6.65% of residual squared energy, insufficient for the fixed 10% RMSE gain. The projection identity used to measure this is elementary.
3. Matched write-back interventions that demonstrate large realized policy contrasts across fixed decoder designs while keeping forecasting value, causal mediation, and loss importance as separate questions.

The rounding condition, stochastic rounding, class-imbalance concern, and use of temporal baselines are established ideas. Our contribution is their empirical separation in these reference audits, not a new optimizer or a claim that every reference predictor fails. The early decoder experiments use analytic grids with fp32 compute/state; Pythia uses BF16 compute with fp32 containers for decoded E4M3 weights. Separate native-hybrid checks use persistent E4M3 matrices and native GEMMs. Each result is limited to its named configuration. A reference remains a paid pilot trajectory; "before the target run" does not mean before obtaining the reference.

Appendix Table 2 contrasts this diagnostic with training regimes that already bypass, compensate for, or exploit invisible updates.

2 Invisible updates and the reference-trace audit

We call a coordinate *frozen* on one update when its stored code is unchanged, and f_t is the fraction frozen on step t. This differs from the never-moved fraction and from held-out loss. A majority frozen fraction need not identify a fixed majority: two coordinates can alternate which one moves while $f_t = 1/2$ on every step. Appendix Table 1 fixes the complete terminology.

2.1 Exact coordinate event

Let Q_m denote round-to-nearest-even on the target grid, q_i an already stored grid value, and d_i the optimizer's signed proposal. The executable exact event is decoded-grid value equality

$$Q_m(q_i + d_i) = q_i. \tag{1}$$

For an interior finite code with lower and upper adjacent codes q_i^- and q_i^+, and away from midpoint ties, Eq. (1) is equivalent to

$$-\tfrac{1}{2}(q_i - q_i^-) < d_i < \tfrac{1}{2}(q_i^+ - q_i). \tag{2}$$

Equality here treats signed zeros as equivalent; it is not bitwise encoding identity. At a midpoint, ties-to-even decides. For an interior normal binary code with m explicit mantissa bits, both gaps are usually $2^{\lfloor \log_2 |q_i| \rfloor - m}$; they differ at binade boundaries. The code test also handles zero, subnormals, ties, and finite-range boundaries according to the target quantizer. The event applies to any optimizer because d_i is the emitted proposal, not a particular gradient rule. We use round-to-nearest-even, midpoint, subnormal, and binade terminology in the conventional sense (IEEE, 2019; Higham,

2002); the analytic target grids below are explicitly specified models, not a claim of IEEE-format conformance.

For homogeneous coordinate populations under gradient descent, the event can be summarized by

$$\rho = \frac{\eta\|g\|_2}{\varepsilon\|w\|_2}, \qquad \varepsilon = 2^{-(m+1)}. \tag{3}$$

Both Euclidean norms pool the same coordinates. If within-binade phases are approximately equidistributed, update-to-weight ratios concentrate, and boundary-code mass is negligible, a symmetric-spacing population approximation gives a median half-freeze boundary $\rho^\star = 1/\sqrt{2}$. This is a heuristic population summary, not a replacement for code equality. Heterogeneous tensors or boundary mass can make the pooled scalar uninformative.

2.2 Projected-reference proxy

From a high-precision trajectory $\{w_t^r, d_t^r\}$, we project each coordinate to a proxy stored code $\widehat{q}_{t,i} = Q_m(w_{t,i}^r)$ and compute

$$\widehat{f}_t(m) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[Q_m(\widehat{q}_{t,i} + d_{t,i}^r) = \widehat{q}_{t,i}\big]\,. \tag{4}$$

No low-precision observation enters Eq. (4). Its nontrivial assumption is that the high-precision path remains a useful counterfactual near the visibility boundary. In a predeclared 64-cell transfer test, the measured and projected paths diverge by update 3 and every uncertainty criterion fails. We therefore treat Eq. (4) as an empirical aggregate proxy, not pathwise or calibrated confidence. Appendix B gives the posterior disagreement bound, evidence hierarchy, complete margin result, and failure cases. It also formalizes why finite source observations and a common reached target prefix do not identify the next RNE event without quantitative regularity, and gives a restricted Hessian-based certificate for full-batch gradient descent.

3 Controlled audit validation and boundaries

We first isolate the mechanism in frozen-feature regression, a mantissa-truncation emulator, trainable multilayer networks, and a CNN. These archived controlled sweeps use the symmetric reference-spacing proxy; exact directional replay is reported where the stored references permit it. In the principal two-layer gradient-descent squared-loss grid, 55 of 72 design cells have both predicted and measured first-majority crossings after step 1. Conditional on this timing-eligible subset, measured crossings span a $384\times$ range (11–4222 steps), with median predicted/measured ratio 1.000 and 52/55 (94.5%) within 15%; the predicted range is 11–4575 steps. Across all 72 cells, four disagree on crossing category (at initialization, later, or absent). Within these controlled testbeds, the proxy remains accurate across the tested optimizer, loss, depth, and architecture changes even when the pooled ρ is loose. Finite-size scaling is consistent with a finite, nonsharpening arithmetic crossover rather than a critical transition. Figure 1 summarizes the exact mechanism and the controlled timing comparison.

A four-layer, 0.8M-parameter character GPT supplies one nontrivial mid-training crossing: measured onsets are 1644/1666 and reference predictions are 1640/1669. Appendix B gives the protocol and persistence boundary.

A predeclared same-testbed interpolation tests a heuristic functional endpoint over 90 held-out cells. Directional code equality trails the symmetric and scalar-ρ baselines (94.4%, 95.6%, and 100% cell accuracy), so its promotion criterion fails. The five directional false negatives are majority-frozen from step 1 yet learn 52.4–59.6% of reference progress, showing that a majority crossing is not a functional stopping time. Appendix D gives the complete rule, windows, and historical replay; none is a calibrated final-loss predictor or a vendor-fp4 claim.

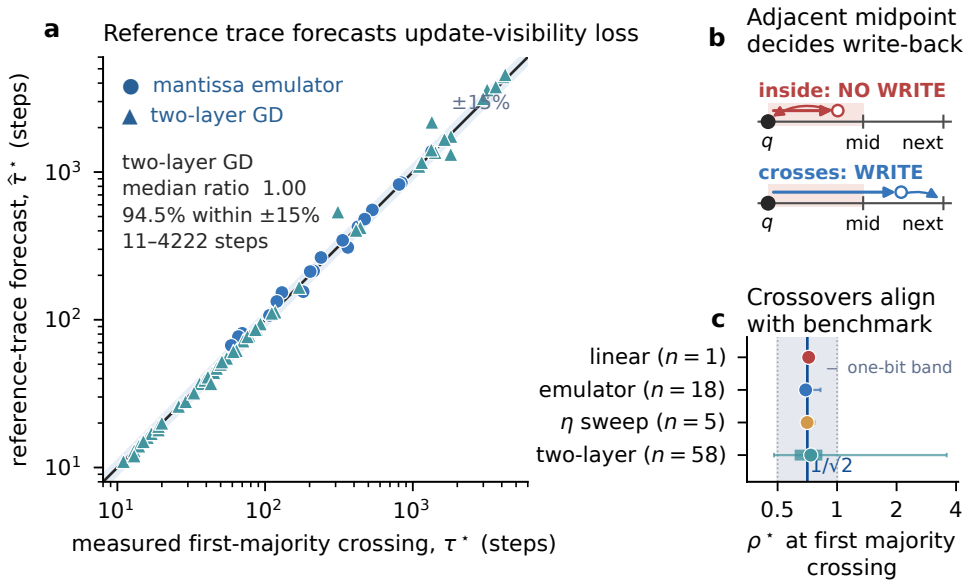

Figure 1: Across the archived controlled testbeds, the timing proxy is accurate in eligible cells but does not eliminate crossing-category errors. **(a)** The panel contains 18 emulator cells and 55 timing-eligible precision–rate–seed coverage cells among 72 fixed two-layer cells: 52/55 lie within the descriptive $\pm 15\%$ band, while four of all 72 disagree on crossing category. Cells sharing one reference are paired coverage tests, not independent replicates. **(b)** The midpoint diagram is a schematic, so no sample size applies. **(c)** Within each stated testbed, points are medians, thick segments are interquartile ranges, and thin lines are full ranges; the testbed-specific n is labeled. Its two-layer $n = 58$ requires a measured post-step-1 crossing, whereas panel (a)'s $n = 55$ also requires a predicted crossing. The $[1/2, 1]$ shading and $1/\sqrt{2}$ line are descriptive guides. No inferential intervals or hypothesis tests are reported.

4 Does a reused reference add predictive value?

4.1 Aggregate agreement and historical temporal structure

A new analytic-grid E4M3 audit freezes one fp32 source trajectory before three unseen target seeds and compares it with each target's direct-RNE trajectory. All training, variation, absolute-accuracy, and skill gates pass: source directional RMSE is 0.00858, versus 0.06834 for fixed-one, and the registered classification is `source_amortized_reference_audit_supported`. Directional specificity fails in every target, however; the symmetric predictor is better in all three.

Without opening any target-RNE outcome, a separate authenticated consumer seals two historical temporal templates and fixed source contrasts. The candidate formulas predate the first direct-RNE completion, but the fixed contrasts and consumer code do not; this is therefore an authenticated outcome-blind secondary analysis, not a fully pre-outcome challenge. The archived direct-RNE template B_Y achieves equal-weight macro RMSE 0.00360, lower in this fixed comparison than the raw fresh-source predictor X_S (0.00858) and its predeclared bias-corrected hybrid H_S (0.00438). This ordering holds in every target. A privileged matched-reference hybrid H_M reaches 0.00283, so matched target-specific information is detectable even though the registered unit-gain reused-source predictors do not capture it. A post-outcome descriptive decomposition attributes 99.65% of measured variation to a common time effect; the fresh common temporal mean correlates 0.99815 with B_Y. These fixed-target results challenge a source-specific interpretation of the registered success. They do not change its gates or prove that the schedule is the physical cause. Appendix C.6 gives the sealed formulas, per-target result, and runtime-mismatch boundary.

4.2 Prospective event and temporal controls at Pythia-410M

The follow-up fixes released Pythia-410M checkpoints (Biderman et al., 2023) at steps 7000, 72000, and 129000, with a reset optimizer and 512 updates per branch. Each age has three historical reference/target pairs, one fresh source, three development targets, and five holdouts: 45 branches in total. Reference weights remain unrestricted fp32; targets apply analytic E4M3 write-back to weights stored in fp32 containers. Both use BF16 forward/backward compute. The threshold is the lowest of 17 candidate bins satisfying the absolute error/coverage limits on all nine development targets, sealed before holdouts. The source and historical data define one common temporal forecast per age, rather than target-specific predictions of temporal deviations.

The source predicts unchanged when $Q(Q(w^{\text{ref}}) + d^{\text{ref}}) = Q(w^{\text{ref}})$. It scores the proposed reference value by normalized distance to its destination-cell boundary and accepts scores at least $1/8$ (bin 32). This geometric margin is not a probability of source-to-target agreement: local robustness within a rounding cell need not survive trajectory drift. Event risk uses 2^{20} fixed sampled coordinate trajectories over 512 updates; the temporal observable Y_t instead averages actual unchanged events over all 405,012,480 eligible coordinates. Appendix C.10 specifies the score, sampling, and finite-population bounds.

All 15 holdouts pass integrity and the original selective component (error upper bound ≤ 0.059, coverage lower bound ≥ 0.50). Nevertheless, every selector has higher empirical error than always predicting unchanged on exactly the same accepted events (Fig. 2a). Accepted unchanged-event prevalence is 96.45–99.77%. More specifically, let C and W count correct and incorrect predictions of change among accepted events, and let A count all accepted events. Then the paired excess risk is exactly

$$R_{\text{selector}} - R_{\text{unchanged}} = \frac{W - C}{A}. \tag{5}$$

Only 14.2–22.7% of accepted change predictions are correct. Departures from the majority label therefore introduce more errors than they correct in every target; prevalence alone is not the full diagnosis. This is a descriptive identity in the retained confusion counts, not a new fitted predictor or test.

For the temporal question, B_Y and B_X average three historical target and reference trajectories. The registered unit-gain correction is $H_S = X_S + (B_Y - B_X)$, where X_S is the fresh source forecast. In every full/leave-one-out comparison, H_S has higher RMSE than B_Y, both raw and after subtracting each trajectory's time mean (Fig. 2b,c). Raw ratios are 1.703–5.475; centered ratios are 1.725–5.475.

A separate post-outcome analysis asks whether choosing a different scalar gain could reach the predeclared 10% RMSE improvement. For $H_\alpha = B_Y + \alpha(X_S - B_X)$, an exact projection bound permits a distinct hindsight-optimal real α for every target, template, and metric. Across all 60 comparisons, even the largest attainable reduction is only 3.38% raw and 3.31% centered. Thus scalar rescaling alone cannot meet the required gain on any observed comparison. The bound uses holdout outcomes; it is neither a fitted replacement predictor nor a new confirmatory test. Appendix C.11 gives the derivation and scope.

The original temporal component already requires lower raw and centered RMSE than the template in every variant; it passes no holdout. The prospective amendment adds constant-event skill and material-effect requirements, also passed by no holdout. Both complete admission decisions are `NO_GO`, closing the larger-model route under this design. These failures occur in the direction of the comparisons, not only at their chosen improvement floors. Finalization required a disclosed post-outcome repair to associate holdout barriers with their sealed development ancestors; thresholds, statistics, inputs, and decision rules were retained. Appendix C.10 gives the full design, repair, simulation scope, and original/amended results.

What the controls establish. The event baseline shares the reference selector's acceptance mask, so the comparison tests its labels on accepted events, not a complete source-free abstention system. The historical template tests the added value of the chosen fresh-source correction for a common temporal forecast. Passing either control would still leave target-specific deviation prediction, independent-source reliability, and downstream training utility unestablished. The separate NeoX matched-reference result above also shows why failure of the reused-source construction should not be read as absence of all predictive information.

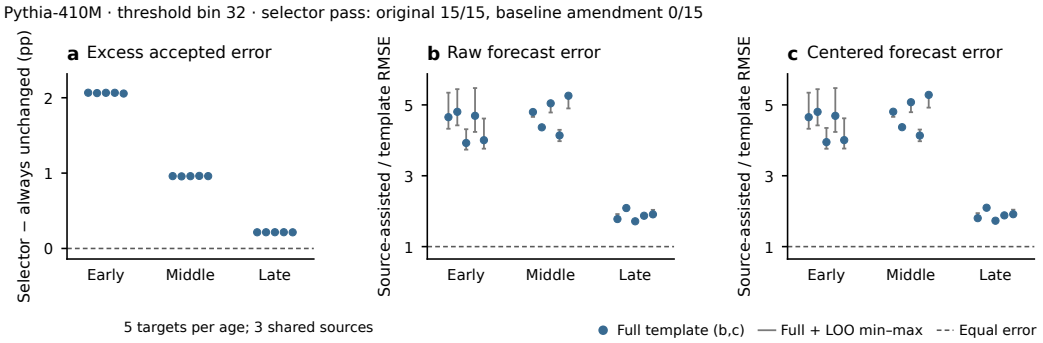

Figure 2: Original selective accuracy passes while matched controls win in the prospective Pythia-410M holdouts. **(a)** Paired excess accepted-event error over always predicting unchanged; zero means equal risk. **(b,c)** Source-hybrid/template RMSE ratios, raw and time-centered; one means equal RMSE. Each point is one target using the full template; segments in (b,c) span its full and three leave-one-out variants, not confidence intervals. Five targets at each checkpoint age share one fresh source and three historical pairs. All 15 targets and all 60 dependent target/variant comparisons are included; no population inference is made. The frozen selector bin is 32. All 15 original selector gates pass, while amended baseline skill and source-value pass counts are zero.

5 Write-back policy effects and persistent-native robustness

For clarity, the analytic-grid Experiments 1–3 are the controlled and decoder write-back studies above; the persistent-native E4M3 checks below are Studies 1–2, and the simulated signed-INT3 direct-state replication is Study 3. These labels identify protocol families, not additional independent training replicates.

5.1 Protocol and round-to-nearest persistence

We train the nanoGPT GPT-2-family 124M configuration (Radford et al., 2019; Karpathy, 2025) on OpenWebText (Gokaslan & Cohen, 2019) with AdamW (Loshchilov & Hutter, 2019), warmup, cosine decay, weight decay, and gradient clipping. After each optimizer step, all trainable 2-D decay-group weights are quantized onto an analytic E5M2 or E4M3 grid derived from the proposed FP8 encodings (Micikevicius et al., 2022), with no fp32 master copy; 1-D normalization parameters remain fp32. Forward and backward compute and optimizer moments remain fp32. This design isolates the destination-grid write-back; it is not a claim about fp8 GEMM.

At the default learning rate 6×10^{-4}, attention and feed-forward weights are majority-frozen from the first update. This at-initialization case tests frozen-fraction tracking but not nontrivial Transformer onset timing. The archived symmetric-gap reference proxy tracks pooled and attention/MLP group-mean frozen-fraction trajectories with RMSE 0.004–0.006. In the corrected matched 10,000-step runs, the final held-out RNE stored-grid-minus-fp32 gaps are 2.129 and 2.144 nats for E5M2, and 1.427 and 1.439 nats for E4M3. These closely reproduce the historical RNE extension on a separately validated split; we make no claim beyond the measured horizon.

5.2 Stochastic rounding intervention

To test the cost of the RNE storage policy relative to SR, the corrected run holds the fp8 grid, initial projected codes, model, data order, optimizer, schedule, and held-out evaluation protocol fixed. Only post-optimizer write-back changes between the RNE and SR arms. For a nonzero in-range proposal between adjacent finite codes, stochastic rounding assigns nonzero probability to changing the code, removing the interior deterministic dead zone while introducing unbiased in-range rounding noise (Gupta et al., 2015; Connolly et al., 2021; Ozkara et al., 2025; Liu et al., 2025). Finite-range endpoints retain the clamping behavior specified in Appendix C.

For same-seed final losses and an intervention arm A, we define gap recovery as

$$R_A = \frac{L_{\text{RNE}} - L_A}{L_{\text{RNE}} - L_{\text{fp32}}}. \tag{6}$$

Here and in the fixed-design modern-decoder transfer below, the RNE and SR arms share projected target-grid codes, whereas the same-seed fp32 anchor starts from the raw fp32 initialization. The numerator therefore isolates the realized write-back-policy contrast, but R_A is a descriptive normalization against that raw-fp32 anchor, not a causal fraction attributable solely to write-back (paired outcomes in Appendix Table 6).

At 10,000 steps, E5M2 recovers 89.24% and 89.43% of the RNE gap, while E4M3 recovers 92.16% and 92.07%. The predeclared point-estimate rule required at least 90% in all four cells: it places E4M3 above and E5M2 narrowly below the threshold in both seeds. This is a fixed protocol decision, not a population threshold. Each training seed has one SR realization, and unrecorded Python hash iteration prevents seed-exact replay of its random stream. These runs cannot distinguish a format-specific sub-90% ceiling from rounding-stream variation. The coordinate-weighted attention/MLP never-moved fraction is about 0.599 under RNE E5M2 and 0.413 under RNE E4M3, but at most 1.18×10^{-8} in every SR cell. Recovery declines from 93.4–93.6% to 89.2–89.4% for E5M2 and from 95.6–97.7% to 92.1–92.2% for E4M3 over steps 3001–10,000. These fixed runs therefore show a large persistent policy contrast, but eliminating deterministic never-moved coordinates does not remove the full long-horizon loss penalty.

Figure 3 shows the paired loss gaps, horizon-dependent recovery, and never-moved fractions. Because the intervention changes the full post-optimizer write-back policy, it identifies the realized fixed-run contrast for that full policy change. Stochastic rounding also changes the rounding-error distribution, so the experiment does not identify deterministic persistence as the exclusive mediator.

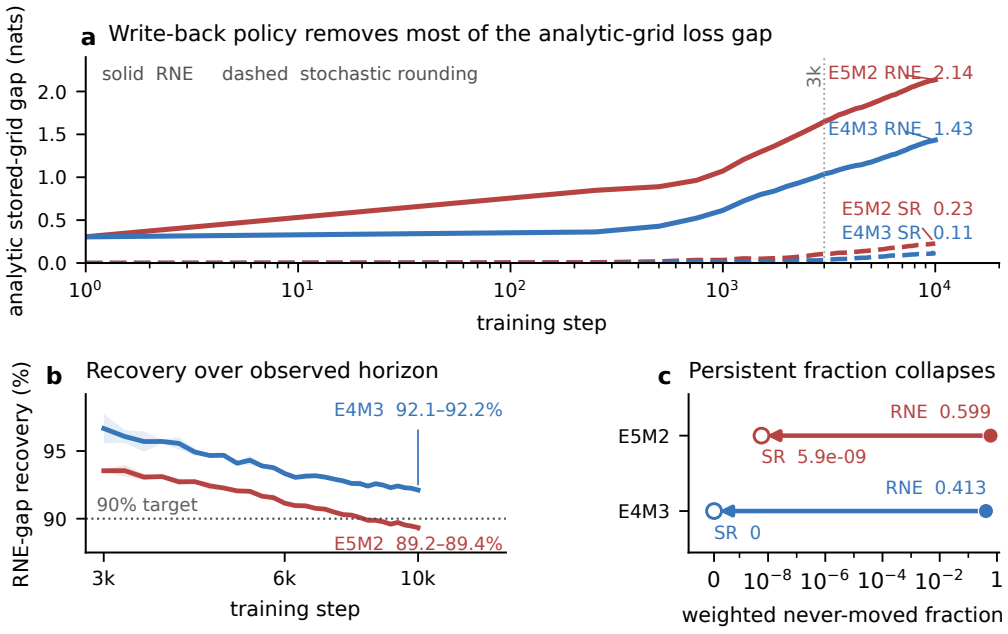

Figure 3: Across two fixed seeds per format, stochastic write-back removes most but not all of the analytic stored-grid loss gap and nearly eliminates the persistent never-moved fraction. The scientific units are fixed training seeds ($n = 2$ per format); red denotes E5M2 and blue E4M3. **(a)** Same-seed analytic stored-grid-minus-raw-fp32 loss gaps in nats under RNE (solid) and SR (dashed). **(b)** RNE-gap recovery in percent from step 3001; the dotted line is the predeclared 90% point target. In (a,b), lines are two-seed means and bands are the full two-seed ranges. **(c)** Filled and open endpoints are two-seed means of the coordinate-weighted attention/MLP never-moved fraction under RNE and SR, respectively; no range is drawn. No inferential intervals or hypothesis tests are reported.

A separate predeclared 162.2M-parameter RMSNorm/RoPE/SwiGLU decoder test (Zhang & Sennrich, 2019; Su et al., 2024; Shazeer, 2020) recovers 92.94–93.61% across two fixed seeds when only write-back changes. The policy effect repeats, but the visibility trajectories fail the registered informativeness and directional-added-value gates. The later allocator does not rerun or restore diagnostic transfer; Appendix Tables 7–8 give the full result.

A separate fixed NeoX/Pythia-style bundle pairs three target seeds across E5M2 and E4M3. Against a projected-start fp32 anchor, mean recovery is 0.958/0.978 for stochastic write-back and 0.819/0.957 for ECO error feedback, respectively. The historical registered policy-and-reference-audit-transfer label is retained, but we claim only policy transfer plus descriptive absolute-RMSE agreement: fixed-one also passes the registered RMSE gate, E5M2 fails the earlier informativeness rule, E4M3 passes, and directional added value is untestable. Appendix C.5 gives the per-format ranges and provenance limits.

5.3 Study 1: Persistent-native E4M3 schedule-shift robustness

Study 1 pairs three targets across constant-mid and cosine-restart schedules, using persistent E4M3 matrices and native hybrid-FP8 GEMMs. Its registered classification is inconclusive: one manual recovery remains producer-marked `scientific_result=false` and lacks canonical process identity. A separate validated retrospective analysis is descriptively negative: the constant-mid centered-RMSE ratio is 1.0525, above its 0.9 gate. The recovered cosine-restart cell cannot alter this disjoint cohort's failure. Appendix C.7 preserves both classifications and the complete protocol/recovery limitations.

5.4 Native write-back contrast and simulated-INT3 boundary

Study 2 pairs three fixed targets across fp32, persistent-native RNE, full software SR, and dead-zone-only SR. All 12 arms pass the registered training/replay gates. Mean full-SR and dead-zone-SR recoveries are 0.9766 and 0.9777; their mean improvement ratio is 1.0012. This fixed-run contrast does not establish causal mediation or hardware stochastic rounding. Appendix C.8 preserves per-target values and the registered protocol label.

The one-seed simulated-INT3 DQT wrapper and its manual continuation are reported in Appendix C.9. They support a fixed-run convergence result, not native INT3 performance or a reference-forecast test.

5.5 Visibility and loss importance

A separate exact-budget case study tests a source mask ranked by absolute proposal contribution $|g_{ti}u_{ti}|$, where g_{ti} is the clipped minibatch-loss gradient and u_{ti} the full signed AdamW proposal, including decay. Visibility records whether a proposal changes a code; this score instead provides a clipped-gradient proxy for first-order loss-contribution magnitude, without guaranteeing descent. A frequently frozen coordinate need not have a large contribution. At 10% fp32-master payload in a fixed 162.2M decoder, it recovers 0.604/0.610/0.607 of the same-target RNE–fp32 gap across three unseen targets; freeze-only selection recovers 0.009/0.007/0.007. This separates visibility from loss importance without making the established saliency score novel or optimal. A fresh GPT-2 budget study then finds no source mask satisfying the pre-fixed deployment tolerance rule after reference-cost amortization. Appendix Sections E.6–E.8 retain all budgets, controls, systems contrasts, and rule dependence.

6 Related work

Master weights accumulate sub-grid updates in a finer destination (Micikevicius et al., 2018); stochastic rounding removes the deterministic dead zone (Gupta et al., 2015; Liu et al., 2025); and unit scaling changes representational scale (Blake et al., 2023). Prior work reports lost bf16/fp16 updates and remedies (Zamirai et al., 2020; Hayford et al., 2024), stochastic accumulation and bf16-Adam analyses (Wang et al., 2018; Ozkara et al., 2025), and roundoff or half-ULP interventions (Xia et al., 2024; Yu, 2026). PULSESync tests visibility online; M+Adam combines additive and multiplicative updates; FORGE and ECO use stochastic write-back or optimizer error feedback (Miahi & Belilovsky,

2026; Liang et al., 2026; Kukreja et al., 2026; Nikdan et al., 2026). We do not claim that these methods miss the dead zone. Our question is whether offline reference forecasts add value on unseen targets beyond constant-event and historical temporal controls, while preserving their original accuracy decisions. Selective classification studies the coverage/error trade-off (El-Yaniv & Wiener, 2010); an absolute-risk requirement does not assert superiority to a constant label. Our cutoffs are study-specific design choices, not community standards. The contribution is the observed behavior and weak residual alignment of this arithmetic-margin reference construction, including the limit of scalar-gain correction. The constant-label identity, temporal controls, and projection formula are established tools. QuRL instead re-quantizes a full-precision RL actor (Li et al., 2026).

Allocation and precision scores are established (Molchanov et al., 2017; Dong et al., 2019; Fu et al., 2020; Pavlova et al., 2026; Mooppil et al., 2026; Zhou et al., 2026; Zhao et al., 2026; Datta et al., 2026). Other methods adapt optimizer configuration/state or optimize frozen-model assignments (Liu et al., 2026a;b; Helcig & Alistarh, 2026). Our supporting allocation study instead freezes an exact-byte fp32-master mask before unseen targets; it claims neither generic reference-guided novelty nor an optimal score.

GaLore projects gradients into a low-rank subspace and maintains optimizer states there (Zhao et al., 2024); reconstructed updates still face the destination's write-back rule. Differentiable FP4 quantization (Wang et al., 2025) addresses another arithmetic channel. Appendix F.2 distinguishes these memory-efficient methods from this stored-weight audit.

7 Limitations and outlook

Pythia's 15 targets share three fresh sources, one per age, and three historical pairs per age. Targets within a source block and leave-one-out variants are not independent replications. Optimizer restarts may alter transient alignment; uninterrupted pretraining is untested. Corpus changes may alter gradients, rounding margins, and temporal correlations; transfer to other curated datasets is untested. The scalar ceiling concerns observed residual directions and permits smaller improvements. Better forecasts need information more aligned with target residuals and validation against history on independent sources. Nonlinear or time-varying corrections are candidates, not guarantees. The post-outcome binding repair preserves rules and metrics but makes the finalized analysis distinct from the sealed executable. The earlier NeoX template comparison has a batching/attention mismatch and a descriptive variance decomposition; neither identifies the schedule as a physical cause.

The analytic decoder and native-hybrid interventions use two or three fixed targets and retain fp32 optimizer state. Stochastic streams are not additional replicates; the historical analytic SR stream is not seed-exactly reconstructable. Study 1 remains registered-inconclusive, with its independent retrospective negative reported separately; Study 2 tests software stochastic rounding and does not identify causal mediation. Study 3 is a one-seed simulated-INT3 wrapper with a manual continuation, not native INT3 storage or arithmetic. These experiments establish neither population generalization nor production efficiency.

The allocator studies use fixed controls, selected budgets, and a privileged target-specific comparator; they yield no registered practical deployment win. The exact event is local, while the aggregate proxy follows a diverging counterfactual. Pythia's finite-population selective bounds do not establish pathwise prediction, loss importance, or independent-source reliability. Appendix F.1 retains the full per-study scope, provenance, and systems limitations.

8 Conclusion

Pythia-410M selectors pass the original error/coverage gates while losing to a constant label, and every fixed source correction loses to its template. A post-outcome geometric bound excludes the required improvement by scalar rescaling of these observed residuals. The original joint protocol correctly rejects the forecast. This separates absolute-risk compliance from predictive skill and diagnoses weak residual alignment in this construction, without ruling out better reference methods.

Reproducibility statement

The appendix fixes model, quantizer, data, statistical units, and divergence conventions; bundles retain trajectories, frozen protocols, terminal records, validation outputs, manifests, and regeneration commands. Integrity receipts and exact source identifiers remain machine-readable release metadata rather than a paper-visible hash ledger. The anonymous package omits identity-linked metadata.

AI use statement

Generative AI tools assisted with hypothesis and conceptual refinement, mathematical derivations and claim review, experimental design including controlled synthetic testbeds, method, data-preparation, and analysis-code implementation, result interpretation, literature search, language translation, manuscript drafting and editing, figure preparation, and consistency checks. The work contains no human-subject or qualitative analysis. We checked AI-assisted code with tests, checked reported quantities against retained artifacts, reviewed citations against source material, and take responsibility for the final text, claims, and artifacts.

References

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.

Charlie Blake, Douglas Orr, and Carlo Luschi. Unit scaling: Out-of-the-box low-precision training. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2548–2576, 2023.

Roberto L. Castro, Andrei Panferov, Soroush Tabesh, Oliver Sieberling, Jiale Chen, Mahdi Nikdan, Saleh Ashkboos, and Dan Alistarh. Quartet: Native FP4 training can be optimal for large language models. *arXiv preprint arXiv:2505.14669*, 2025. Version 4.

Brian Chmiel, Maxim Fishman, Ron Banner, and Daniel Soudry. FP4 all the way: Fully quantized training of large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 101214–101235, 2025. doi: 10.52202/085713-3049. arXiv:2505.19115v2.

Musa Cim, Sarthak Arora, Poovaiah Palangappa, Miro Hodak, Ravi Dwivedula, Meena Arunachalam, and Mahmut Taylan Kandemir. Pretraining large language models with MXFP4 on native FP4 hardware. *arXiv preprint arXiv:2605.09825*, 2026. URL https://arxiv.org/abs/2605.09825v4. Version 4.

Michael P. Connolly, Nicholas J. Higham, and Theo Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing*, 43(1):A566–A585, 2021. doi: 10.1137/20M1334796.

Shrestha Datta, Hongfu Liu, and Anshuman Chhabra. Weight-adjusted gradients reveal parameter importance and failure modes in LLMs. *arXiv preprint arXiv:2607.10803*, 2026. URL https://arxiv.org/abs/2607.10803v1. Version 1; contemporaneous work.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Siyu Ding, Mingchuan Ma, Jiabo Tong, Xingrun Xing, Ziming Wang, and Guoqi Li. Full-stack FP4: Stable LLM pretraining with quantized projections, optimizers, and attention. *arXiv preprint arXiv:2607.04422*, 2026. URL https://arxiv.org/abs/2607.04422v2. Version 2.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian AWare quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019. doi: 10.1109/ICCV.2019.00038.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Giuseppe Franco, Ian Colbert, Pablo Monteagudo-Lago, Felix Marty, and Nicholas Fraser. dMX: Differentiable mixed-precision assignment for low-precision floating-point formats. *arXiv preprint arXiv:2606.04115*, 2026.

Yonggan Fu, Haoran You, Yang Zhao, Yue Wang, Chaojian Li, Kailash Gopalakrishnan, Zhangyang Wang, and Yingyan Lin. FracTrain: Fractionally squeezing bit savings both temporally and spatially for efficient DNN training. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Aaron Gokaslan and Vanya Cohen. OpenWebText Corpus. Online corpus, 2019. URL https://skylion007.github.io/OpenWebTextCorpus/.

Jose Javier Gonzalez Ortiz, Abhay Gupta, Christopher Rinard, and Davis Blalock. FlashOptim: Memory efficient optimizers for large-scale training. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63497. Spotlight; arXiv:2602.23349v2.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1737–1746, 2015.

Joel Hayford, Jacob Goldman-Wetzler, Eric Wang, and Lu Lu. Speeding up and reducing memory usage for scientific machine learning via mixed precision. *Computer Methods in Applied Mechanics and Engineering*, 428:117093, 2024. doi: 10.1016/j.cma.2024.117093. URL https://doi.org/10.1016/j.cma.2024.117093. arXiv:2401.16645.

Michael Helcig and Dan Alistarh. Model compression with exact budget constraints via riemannian manifolds. *arXiv preprint arXiv:2605.00649*, 2026. URL https://arxiv.org/abs/2605.00649v2. Version 2.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, second edition, 2002. doi: 10.1137/1.9780898718027.

IEEE. IEEE Standard for Floating-Point Arithmetic. IEEE Standard 754-2019, IEEE, 2019.

Andrej Karpathy. nanoGPT. GitHub repository, 2025. URL https://github.com/karpathy/nanoGPT/commit/0000000000000000000000000000000000000000. Commit 3adf61e.

Dikshant Kukreja, Kritarth Prasad, Avinash Anand, Zhengkui Wang, Erik Cambria, Timothy Liu, Aik Beng Ng, Simon See, and Bapi Chatterjee. FORGE: Fused on-register gradient elimination for memory-efficient LLM training. *arXiv preprint arXiv:2606.22932*, 2026.

Yuhang Li, Reena Elangovan, Xin Dong, Priyadarshini Panda, and Brucek Khailany. QuRL: Low-precision reinforcement learning for efficient reasoning. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10008335. Poster; matching arXiv:2602.13953v1 manuscript titled "QuRL: Efficient Reinforcement Learning with Quantized Rollout".

Xiaoyuan Liang, Sebastian Loeschcke, Mads Toftrup, and Anima Anandkumar. M+Adam: Low-precision training via additive–multiplicative optimization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XgVv56Y829. ICML 2026; arXiv:2607.10611v2.

Kang Liu, Wei Peng, and Jianchen Hu. Budget-aware auto optimizer configurator. *arXiv preprint arXiv:2605.04711*, 2026a. URL https://arxiv.org/abs/2605.04711v1. Version 1.

Minglu Liu, Cunchen Hu, Liangliang Xu, Fengming Tang, Ruijia Wang, and Fu Yu. STQuant: Spatio-temporal adaptive framework for optimizer quantization in large multimodal model training. *arXiv preprint arXiv:2604.06836*, 2026b. URL https://arxiv.org/abs/2604.06836v2. Version 2.

Taowen Liu, Marta Andronic, Deniz Gündüz, and George Anthony Constantinides. Training with fewer bits: Unlocking edge LLMs training with stochastic rounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14531–14546, 2025. doi: 10.18653/v1/2025.findings-emnlp.784.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. arXiv:1711.05101.

Erfan Miahi and Eugene Belilovsky. Understanding and exploiting weight update sparsity for communication-efficient distributed RL. *arXiv preprint arXiv:2602.03839*, 2026.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. FP8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022. URL https://arxiv.org/abs/2209.05433.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017.

Gautam Krishna Mooppil, Abhijai Sasikumar, Abeer Mathur, and Subramaniyaswamy Vairavasundaram. Sensitivity-guided mixed-precision quantization-aware training for large language models. *IEEE Access*, 14:84117–84145, 2026. doi: 10.1109/ACCESS.2026.3698449.

Mahdi Nikdan, Amir Zandieh, Dan Alistarh, and Vahab Mirrokni. ECO: Quantized training without full-precision master weights. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=dfVm5yZZ6G. ICML 2026; arXiv:2601.22101v1.

Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for LLM training: Theory and practice. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4402–4410. PMLR, 2025. URL https://proceedings.mlr.press/v258/ozkara25b.html.

Andrei Panferov, Erik Schultheis, Soroush Tabesh, and Dan Alistarh. Quartet II: Accurate LLM pre-training in NVFP4 by improved unbiased gradient estimation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65532. Poster; arXiv:2601.22813v2.

Mariya Pavlova, Harrison Bo Hua Zhu, Lidia Vitanova, Elizaveta Semenova, and Yingzhen Li. Quantizing time-series models as dynamical systems: Trajectory-based quantization sensitivity score. *arXiv preprint arXiv:2606.13300*, 2026. URL https://arxiv.org/abs/2606.13300. Version 3.

Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, Yuxiang Yang, Ze Liu, Yifan Xiong, Ziyue Yang, Bolin Ni, Jingcheng Hu, Ruihang Li, Miaosen Zhang, Chen Li, Jia Ning, Ruizhe Wang, Zheng Zhang, Shuguang Liu, Joe Chau, Han Hu, and Peng Cheng. FP8-LM: Training FP8 large language models. *arXiv preprint arXiv:2310.18313*, 2023. Version 2.

Haiquan Qiu and Quanming Yao. Why low-precision transformer training fails: An analysis on flash attention. In *International Conference on Learning Representations*, 2026. arXiv:2510.04212.

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.

Mehdi Rahimifar, Amin Darabi, Mehran Taghian Jazi, Xing Huang, Yao Wang, Zhijun Tu, Yufei Cui, Yunke Peng, and Hongliang Li. Stable FP4 training via transposition-invariant block quantization. *arXiv preprint arXiv:2607.24953*, 2026. Version 1.

R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1), 1974. doi: 10.1214/aos/1176342611.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. URL https://arxiv.org/abs/2002.05202.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063. arXiv:2104.09864.

Kristi Topollai and Anna Choromanska. Understanding quantization of optimizer states in LLM pre-training: Dynamics of state staleness and effectiveness of state resets. *arXiv preprint arXiv:2603.16731*, 2026.

Albert Tseng, Tao Yu, and Youngsuk Park. Training LLMs with MXFP4. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1630–1638. PMLR, 2025.

Naigang Wang, Jungwook Choi, Daniel Brand, Chia-Yu Chen, and Kailash Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1812.08011. arXiv:1812.08011.

Ruizhe Wang, Yeyun Gong, Xiao Liu, Guoshuai Zhao, Ziyue Yang, Baining Guo, Zheng-Jun Zha, and Peng Cheng. Optimizing large language model training using FP4 quantization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62937–62957. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25ae.html.

Haocheng Xi, Han Cai, Ligeng Zhu, Yao Lu, Kurt Keutzer, Jianfei Chen, and Song Han. COAT: Compressing optimizer states and activation for memory-efficient FP8 training. In *International Conference on Learning Representations*, 2025. arXiv:2410.19313v3.

Lu Xia, Stefano Massei, Michiel E. Hochstenbach, and Barry Koren. On stochastic roundoff errors in gradient descent with low-precision computation. *Journal of Optimization Theory and Applications*, 200(2):634–668, 2024. doi: 10.1007/s10957-023-02345-7.

Zhenyu Yu. Escaping the swamping zone: Fully low-precision training via relative ULP noise injection. In *Proceedings of the 2026 International Conference on Artificial Intelligence and Control*, pp. 568–572. ACM, 2026. doi: 10.1145/3807246.3807335.

Pedram Zamirai, Jian Zhang, Christopher R. Aberger, and Christopher De Sa. Revisiting BFloat16 training. *arXiv preprint arXiv:2010.06192*, 2020. URL https://arxiv.org/abs/2010.06192.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12360–12371, 2019. URL https://arxiv.org/abs/1910.07467. arXiv:1910.07467.

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.

Junchen Zhao, Ali Derakhshan, Jayden Hyman, Junhao Dong, Sangeetha Abdu Jyothi, and Ian Harris. CoopQ: Cooperative game inspired layerwise mixed precision quantization for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7566–7578, 2026. doi: 10.18653/v1/2026.findings-acl.373.

Kaiyan Zhao, Tsuguchika Tabaru, Kenichi Kobayashi, Takumi Honda, Masafumi Yamazaki, and Yoshimasa Tsuruoka. Direct quantized training of language models with stochastic rounding. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pp. 1150–1165. PMLR, 2025. URL `https://proceedings.mlr.press/v304/zhao26b.html`. Protocol snapshot: arXiv:2412.04787v3.

Changhai Zhou, Shiyang Zhang, Yuhua Zhou, Jun Gao, Qian Qiao, Shichao Weng, Weizhong Zhang, and Cheng Jin. Balancing fidelity and plasticity: Aligning mixed-precision fine-tuning with linguistic hierarchies. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15885–15896, 2026. doi: 10.18653/v1/2026.findings-acl.779.

Appendix A fixes terminology and arithmetic conventions; Appendix B gives the projected-reference bound and transfer tests; Appendix C specifies the Transformer intervention; Appendix D reports precision–rate stress tests; Appendix E records allocation protocols and controls; and Appendix F documents artifacts and units.

A Arithmetic condition, training regimes, and controlled testbeds

Table 1: Observables used to distinguish arithmetic events from functional outcomes.

Observable	Definition
Coordinate freeze	A stored coordinate has unchanged decoded-grid value on one update; signed zeros are equivalent.
Frozen fraction f_t	Fraction of coordinates unchanged on step t.
First majority crossing	First step with $f_t \geq 1/2$; used in controlled systems without a warmup recrossing.
Persistent majority crossing	First step after which $f_t \geq 1/2$ for the remainder of the observed run.
Never-moved fraction	Fraction of coordinates unchanged over every update up to step t.
Loss plateau	Functional outcome measured on held-out data; not implied by any single frozen-fraction threshold.

A.1 Symmetric-spacing population approximation

Away from a binade-boundary code, a nonzero normal binary value with m explicit mantissa bits has symmetric adjacent gap $u_m(w_i) = 2^{\lfloor \log_2 |w_i| \rfloor - m}$. Writing $\varepsilon = 2^{-(m+1)}$ and $\phi_i = \mathrm{frac}(\log_2 |w_i|)$, the symmetric-spacing approximation can be written

$$\rho_i \equiv \frac{|\Delta w_i|}{\varepsilon |w_i|} < 2^{-\phi_i}, \qquad \phi_i \in [0, 1). \tag{7}$$

The right-hand side spans $(1/2, 1]$, one bit, as a reference coordinate moves through a binade. If the phases are uniform, its median and geometric mean are $2^{-1/2}$ and its arithmetic mean is $1/(2 \ln 2)$. The main-text $\rho^\star = 1/\sqrt{2}$ is therefore an approximate median boundary only when update-to-weight ratios also concentrate, so that the weighted root-mean-square ρ represents a typical coordinate, and when projected boundary-code mass is negligible. At a power-of-two code the toward-zero gap is half the away-from-zero gap, so Eq. (7) is not the exact event. The corrected implementation instead evaluates Eq. (1), including zeros, subnormals, ties, and format boundaries according to the target quantizer.

The uniform-phase check uses $f_{\text{pred}} = \langle \mathrm{clip}(-\log_2 \rho_i, 0, 1) \rangle$. Across all widths, precisions, and seeds in the finite-size experiment, its cell-mean absolute error against the bitwise frozen fraction at the first majority crossing is 0.008 (coordinate-count-weighted 0.004); the worst cell is 0.090 at $H = 16$, $m = 8$, seed 43003. The measured pooled locations are 0.707 in the emulator, 0.72 on bf16 hardware, and approximately 0.74 in the trainable two-layer network.

Table 2: Low-precision update regimes and the remaining diagnostic question. Existing systems bypass, compensate for, or exploit invisible updates; our run-specific forecast is intended to inform future tests of where those mechanisms are useful. "Assignment" means choosing where to spend precision, not predicting a training-time visibility boundary.

Work/regime	Low-precision destination	Response or objective	Forecast requirement and reuse
Mixed precision (Micikevicius et al., 2018)	Low-precision compute; fp32 stored master	Accumulate every update in a finer copy	No matched reference; bypasses the stored-weight dead zone
Yu (Yu, 2026)	Direct low-precision weights	Detect current sub-ULP updates; recover ResNet/T5 tasks with ULP injection	Online per-update gate; no full matched reference
PULSESync (Miahi & Belilovsky, 2026)	bf16 synchronization cast	Exploit invisible deltas to reduce communication	Online cast sparsity; no pre-run full reference
BF16/8-bit SR (Zamirai et al., 2020; Wang et al., 2018; Ozkara et al., 2025)	Direct low-precision training	Preserve small updates with SR, Kahan, or chunked accumulation	No matched reference; mitigation is applied broadly
FORGE / ECO (Kukreja et al., 2026; Nikdan et al., 2026)	bf16 weights with optional INT8 optimizer states / directly quantized weights	Stochastic write-back or optimizer error feedback without a full master copy	No matched reference; mitigation is applied broadly
Quartet II / dMX (Panferov et al., 2026; Franco et al., 2026)	NVFP4 training / MXFP deployment quantization	Unbiased estimation / learned layer-format assignment	No matched reference; assignment objective differs
This work	Direct rounded 2-D stored weights	Forecast aggregate frozen fraction/onset; test write-back cost; use absolute proposal contribution for exact-payload tile protection	Audit: candidate-matched reference per seed/schedule, reusable across grids. Allocation: one regime-matched source trace reused across targets; target-P is a privileged comparator

A.2 Mantissa emulator and controlled systems

The emulator carries arithmetic in fp64 and rounds each stored weight to m mantissa bits after every optimizer step. Its wide exponent isolates destination mantissa resolution from exponent underflow. At $m = 7$, its frozen-fraction curve matches a direct `torch.bfloat16` weight update to RMSE 1.8×10^{-3} in frozen-feature regression, 6.7×10^{-3} in the two-layer network, and 2.2×10^{-2} in the CNN.

The frozen-feature system maps short contexts through a fixed tanh random feature map with eight outputs and trains only a linear readout. The trainable bridge is a two-layer tanh teacher–student regression with all weights and biases updated. The real-data test is a roughly 106k-parameter CNN on MNIST: two 3×3 convolution–pool blocks followed by $1568 \to 64 \to 10$ fully connected layers. These systems test the same destination-grid event under frozen versus trained features, squared versus cross-entropy loss, and linear, multilayer, and convolutional architectures.

To distinguish a finite arithmetic crossover from a sharpening transition, we reran the input-layer sweep at coordinate counts $N = 128$–4096 and $m \in \{6, 7, 8\}$ over six fixed seeds using the phase-correct coordinate $z_i = \log_2(g_i^{\text{dir}}/(2|\Delta w_i|))$, where g_i^{dir} is the adjacent-code gap in the proposal direction. At the first majority crossing, the primary width $Q_{0.84}(z) - Q_{0.16}(z)$ rises from 2.811 to 4.490, 3.267 to 4.734, and 3.426 to 4.566 for $m = 6, 7, 8$, respectively. Descriptive fits $W(N) \propto N^{-a}$ give $a = -0.100, -0.075, -0.069$ ($R^2 = 0.67, 0.50, 0.75$). Thus the exact widths do not narrow over the tested range, which is consistent with a finite, nonsharpening arithmetic crossover. This is neither evidence for a thermodynamic phase transition nor a universality or scaling-law claim.

B Projected-reference proxy and transfer tests

For every target grid, the corrected predictor projects a separate high-precision trajectory to a proxy stored code and evaluates Eq. (1) before the high-precision update is applied. The measured run independently rounds its stored weights after each step. For one trajectory, a nondegenerate crossing is a first majority crossing after step 1. A paired timing cell is eligible only when both the measured and predicted crossings exist and are nondegenerate. At-init and no-crossing cells remain categorical outcomes but do not carry a finite post-init timing ratio.

For finite real proposals, let $q^a_{t,i}$ and $d^a_{t,i}$ be the actual low-precision code and proposal, and let $f^a_t = N^{-1}\sum_i \mathbf{1}[Q_m(q^a_{t,i} + d^a_{t,i}) = q^a_{t,i}]$. Let $\gamma_{t,i}$ be the projected reference proposal's distance to its nearest rounding boundary, set to zero when no finite interior destination cell defines that distance. Define

$$c_t = \frac{1}{N}\sum_i \mathbf{1}[q^a_{t,i} \neq \widehat{q}_{t,i}], \qquad v_t = \frac{1}{N}\sum_i \mathbf{1}\left[q^a_{t,i} = \widehat{q}_{t,i}, |d^a_{t,i} - d^r_{t,i}| \geq \gamma_{t,i}\right]. \tag{8}$$

The local statement gives the aggregate decomposition

$$|f^a_t - \widehat{f}_t| \leq c_t + v_t. \tag{9}$$

Because its right-hand side uses the realized low-precision state and proposal, Eq. (9) is a posterior diagnostic, not an a priori guarantee. A pathwise certificate additionally needs common initial codes, projected-reference transition consistency, and a pre-observation proposal-drift bound below every claimed rounding margin.

Table 3: Claim and evidence hierarchy. Failure of a higher layer does not alter the exact lower-layer arithmetic event.

Object	Status in this work
Realized $Q_m(q^a + d^a) = q^a$	Exact local arithmetic event.
Projected-reference code equality	Counterfactual proxy conditional on path agreement.
Frozen fraction and onset	Empirical aggregate testbed forecast; archived sweeps partly use the symmetric proxy.
Pathwise/calibrated uncertainty	Unsupported after divergence; all uncertainty gates in the locked 64-cell test fail.
Crossing-to-loss map	Heuristic binary classifier; not calibrated final-loss prediction.
Proposal-derived importance scores	Separate first-order allocation proxies; visibility does not imply importance.

Deterministic disagreement bound. For finite real proposals, with equality denoting target-code identity, consider one projected-reference proposal $z^r_{t,i} = \widehat{q}_{t,i} + d^r_{t,i}$ whose destination has finite adjacent rounding boundaries $b_- < z^r_{t,i} < b_+$, define $\gamma_{t,i} = \min(z^r_{t,i} - b_-, b_+ - z^r_{t,i})$. Let

$$\widehat{I}_{t,i} = \mathbf{1}[Q_m(z^r_{t,i}) = \widehat{q}_{t,i}], \qquad I^a_{t,i} = \mathbf{1}[Q_m(q^a_{t,i} + d^a_{t,i}) = q^a_{t,i}]. \tag{10}$$

If $q^a_{t,i} = \widehat{q}_{t,i}$ and $|d^a_{t,i} - d^r_{t,i}| < \gamma_{t,i}$, then the actual proposal remains in the same open destination cell as $z^r_{t,i}$. Its next code therefore equals the projected-reference next code, and the common current code makes $I^a_{t,i} = \widehat{I}_{t,i}$. Setting $\gamma_{t,i} = 0$ for unsupported or boundary cases makes the corresponding margin condition conservatively false, so for every coordinate

$$\mathbf{1}[I^a_{t,i} \neq \widehat{I}_{t,i}] \leq \mathbf{1}[q^a_{t,i} \neq \widehat{q}_{t,i}] + \mathbf{1}\left[q^a_{t,i} = \widehat{q}_{t,i}, |d^a_{t,i} - d^r_{t,i}| \geq \gamma_{t,i}\right]. \tag{11}$$

Because $f^a_t = N^{-1}\sum_i I^a_{t,i}$ and $\widehat{f}_t = N^{-1}\sum_i \widehat{I}_{t,i}$, summing Eq. (11) and applying the triangle inequality gives Eq. (9). Strict inequality excludes midpoint ties. The bound is deterministic but posterior when it uses d^a. Induction also requires $q^a_{0,i} = \widehat{q}_{0,i}$ and projected-reference transition consistency $Q_m(\widehat{q}_{t,i} + d^r_{t,i}) = \widehat{q}_{t+1,i}$, in addition to a separately justified pre-observation bound $B_{t,i} \geq |d^a_{t,i} - d^r_{t,i}|$ with $B_{t,i} < \gamma_{t,i}$ at every claimed step. Nonfinite events are excluded from the theorem and treated as unit violations in empirical summaries.

Reachable-prefix finite-trace non-identifiability. Fix any finite collection of source full-batch-gradient-descent executions and one deterministic quantized full-batch-gradient-descent target execution. Fix their initial states, finite horizons, step sizes, and precision/rounding schedules independently of the objective. Let $F \in C^\infty$ realize all of the source traces and, for some $T \geq 1$, the target prefix

$$q_{t+1} = Q_t(q_t - \eta_t \nabla F(q_t)), \qquad 0 \le t < T, \qquad q_T = q. \tag{12}$$

Let S be the finite union of recorded source states. Assume $q \notin S$, $q_t \neq q$ for $t < T$, and that the fixed next target step has $\eta_T > 0$. Thus q is actually reached for the first time by the target and has not been queried by a source trace. Suppose the next coordinatewise RNE map $Q_{T,i}$ represents the finite code q_i and has a finite next larger code q_i^+. In the class of C^∞ objectives with no common global bound on gradients, Hessians, or higher derivatives, there are two objectives that reproduce every source trace and the complete reached target prefix through arrival at q, but have opposite next-code events at coordinate i.

Indeed, set $A = S \cup \{q_0, \dots, q_{T-1}\}$. Finiteness and $q \notin A$ permit a cutoff $\chi \in C_c^\infty$ that equals one near q and whose support misses A. For a desired proposal d, define

$$F_d(x) = F(x) + v_d^\top (x - q)\chi(x), \qquad v_d = -d/\eta_T - \nabla F(q). \tag{13}$$

The perturbation vanishes on neighborhoods of all states in A, so F_d agrees there with F in value and in every derivative. Induction under the fixed schedules therefore preserves every source execution and every target transition in Eq. (12); both constructed executions really reach q. Also $F_d(q) = F(q)$ and $-\eta_T \nabla F_d(q) = d$, so even the objective value on arrival is common and the gradient used to leave q is the first differing query. Choose $d^{\text{stay}} = 0$ and $d^{\text{cross}} = 3(q_i^+ - q_i)e_i/4$. Then

$$Q_{T,i}(q_i + d_i^{\text{stay}}) = q_i, \qquad Q_{T,i}(q_i + d_i^{\text{cross}}) = q_i^+, \tag{14}$$

with no midpoint tie. Reachability thus does not make a finite trace identify the next RNE event. The perturbations are compactly supported, so they preserve lower boundedness or coercivity when F has the corresponding property. This result is limited to deterministic full-batch gradient descent with the stated fixed schedules; objective-dependent schedules, optimizer state dynamics, or a function class with shared quantitative derivative bounds require separate assumptions.

A restricted prospective certificate. For full-batch gradient descent on $F \in C^2$, let $d_t^r = -\eta \nabla F(w_t^r)$, let $\widehat{q}_t$ be the projected-reference code, and set $e_t = \widehat{q}_t - w_t^r$. Suppose known row-wise envelopes satisfy

$$\overline{H}_{t,ij} \ge \sup_{s \in [0,1]} \left| \partial_{ij}^2 F(w_t^r + s e_t) \right|, \qquad B_{t,i} = \eta \sum_j \overline{H}_{t,ij} |e_{t,j}|. \tag{15}$$

For the actual recursion $q_{t+1}^a = Q_m(q_t^a - \eta \nabla F(q_t^a))$, suppose $q_0^a = \widehat{q}_0$, projected-reference transition consistency holds, every reference proposal has an open supported destination cell, and $B_{t,i} < \gamma_{t,i}$ for every coordinate and claimed update, then $q_t^a = \widehat{q}_t$ and $I_{t,i}^a = \widehat{I}_{t,i}$ throughout that horizon. Indeed, conditional on code equality at update t, the fundamental theorem of calculus bounds the proposal difference by Eq. (15); the strict margin condition keeps both proposals in the same destination cell, and induction gives the claim. All-coordinate certification is required because one uncertified coordinate can change later gradients. For a quadratic with $H = H^\top$, $F(w) = \frac{1}{2} w^\top H w + b^\top w + c$, the proposal difference is exactly $-\eta H e_t$, so the sharper computable choice is $B_{t,i} = \eta |[H e_t]_i|$. This is a restricted arithmetic certificate, not a decoder or AdamW guarantee.

Reference-validity prerequisites. A matched forecast fixes initialization, data order, optimizer, schedule, and seed. A new schedule or rate requires a new reference, although one reference can be reused across destination grids and allocation scores. Mapping a predicted crossing to reference loss additionally requires positive reference progress; the stopped cross-entropy protocol demonstrates why this is an integrity condition. Reference cost must be reported. Finally, agreement across several reference precisions does not establish closeness between the reference and low-precision paths.

Across four seeds in the principal trainable-network grid, 55 of 72 cells meet the paired timing rule. Their measured freeze times range from 11 to 4222 steps (predicted: 11–4575). The median predicted/measured ratio is 1.000 (mean 1.019), with 52/55 within 15%; four of all 72 cells disagree on whether the crossing is at initialization, later, or absent. One seed that freezes at step 61 while sibling

Table 4: Transfer of predicted/measured first-majority-crossing time. Counts are precision–rate–seed cells for which both crossings exist and occur after step 1; at-init and no-crossing cells are excluded from these timing ratios. Cells sharing a reference trajectory are coverage tests, not independent statistical replicates.

System	n	Median ratio	Within 15%
Two-layer, GD, squared loss	55	1.000	94.5%
Mantissa emulator	18	1.03	83.3%
Two-layer, Adam, squared loss	60	1.000	90.0%
Two-layer, GD, cross-entropy	64	1.016	92.2%
Three-layer, GD, cross-entropy	25	0.999	96.0%
CNN, GD, cross-entropy	15	1.000	100.0%

seeds freeze near 1800 is also predicted at 61. The predictor therefore follows individual trajectories rather than merely fitting a common schedule. Recomputing the reference at fp64, fp32, and a 16-bit reference mantissa gives identical predicted steps in all 25 clean comparison cells; the GPT studies use an fp32 reference.

The character-GPT bridge uses four layers, four heads, width 128, context 128, and approximately 0.8M parameters on Shakespeare. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, peak rate 6×10^{-4}, 100-step warmup, cosine decay to 6×10^{-5}, weight decay 0.1, clipping at 1.0, batch size 32, and 3000 steps. Compute and moments are fp32; the stored decay-group weights are rounded after each step with no master copy. For the bf16-equivalent $m = 7$ grid, persistent majority freeze begins at steps 1644/1666 in two seeds, versus predictions 1640/1669. The claim is persistence for the remainder of these observed runs, not permanence beyond them.

B.1 Locked forecast-margin transfer

The fresh margin test removes the earlier final-loss progress anchor and tests coordinate-event uncertainty directly. It uses a depth-3 tanh student $[8, 48, 48, 4]$, cross-entropy loss, full-batch fp64 gradient descent, and common one-time-quantized reference/RNE initial codes. Four new seeds, four mantissas $m \in \{5, 7, 9, 11\}$, and four held-out rates give 64 cells of 4,000 updates. The canonical execution is single-thread CPU/fp64; all declared cells complete and none is replaced or survivor-filtered.

Before any RNE measurement, the committed reference-only artifact fixes the exact midpoint-distance margin, the symmetric upward-ULP comparator, nine abstention thresholds, the primary threshold $1/8$, and all decision gates. The resulting exact-margin AUROC is 0.722 versus 0.656 for the comparator. At $1/8$, coverage is 0.932 and risk is 0.105 versus 0.118 at full coverage; the locked coordinate gate instead requires risk at most 0.059. Frozen-fraction interval containment is 0.368, and measured persistent onset falls inside its interval in 33/64 cells. The registered coordinate-margin, aggregate-interval, and full-path-substitution support decisions are therefore all negative without post-outcome retuning.

In all 64 cells, the measured and projected pre-update current-code vectors first differ at update 3. Across 762,880,000 coordinate-events, the same-code premise holds in 0.797%, and both premises of the one-step posterior certificate hold in 0.715%. Of all exact-predictor errors, 99.863% occur in diverged-code events; the error rate is 2.02% when current codes agree and 11.86% when they differ. Aggregate misses are directional rather than uniformly conservative: measured frozen fractions lie above the locked upper bound in 62,722/571, 48,294/7,518, 7,794/20,640, and 0/14,338 above/below step-events for $m = 5, 7, 9, 11$, respectively. The measured persistent onset is inside/earlier/later than the locked interval in 33/15/16 cells. These are fixed-design, post-outcome failure attributions, not a replacement predictor or an inferential sample over coordinates.

A strict independent deterministic replay covers all 64 reference and RNE cells and reproduces their transcripts, aggregate statistics, and decisions. The validation binds two disclosed ordering-only compatibility corrections: one to protocol versus lexical transcript aggregation and one to the order of five diagnostic onset-width lists. Neither correction changes a scalar, cell, multiset, gate, or scientific decision. An earlier cross-entropy transfer that depended on a positive reference-progress

anchor remains separately archived; it stopped before any RNE arm when 12/48 references failed that integrity condition and contributes no outcome to this result.

C Transformer protocol and matched policy measurements

C.1 Model, quantizer, and parameter groups

The primary transformer experiments use the nanoGPT GPT-2-family 124M configuration: 12 layers, 12 heads, width 768, context 1024, vocabulary 50304, no bias, and dropout 0. The 124M label follows nanoGPT; the original GPT-2 report names its smallest model 117M (Radford et al., 2019). The implementation and OpenWebText snapshot are identified in the accompanying release manifest (Karpathy, 2025; Gokaslan & Cohen, 2019). Training uses AdamW $(0.9, 0.95)$, 100-step warmup, peak learning rate 6×10^{-4}, cosine decay toward a configured minimum of 6×10^{-5} in both the 3,000- and 10,000-step runs, weight decay 0.1, clipping at 1.0, batch size 12, and a fixed held-out evaluation set (eight batches per evaluation). The E5M2/E4M3 encodings are derived from the proposed FP8 formats (Micikevicius et al., 2022). E5M2 uses scale 1; E4M3 uses the exact power-of-two scale 128. At initialization 99.5% and 99.0% of weights, respectively, are normal. On six million initialization-scale values per format, the analytic grid agrees bitwise with the corresponding `torch.float8` cast with zero mismatches. We do not claim cast equivalence outside those tested finite in-range values: the analytic E5M2 and E4M3 quantizers saturate, whereas a `torch.float8` overflow can produce nonfinite values with format- and PyTorch-version-dependent semantics. Trainable 2-D decay-group parameters remain fp32 containers constrained to fp8 codes; 1-D normalization parameters, forward/backward compute, and AdamW state remain fp32, and no fp8 GEMM is enabled.

The archived "dense" statistic is an equal mean of the attention and feed-forward group fractions, not a coordinate-weighted fraction over their union. Corrected runs retain group sizes for coordinate-weighted aggregation. The 38.6M-parameter `transformer.wte.weight` is reported separately in the allocator studies and is one parameter tied to `lm_head.weight`; it therefore receives both input-embedding and dense output-softmax gradient contributions. We make no claim that its measured leverage belongs to either role alone. The tile screen enumerates and charges this physical parameter once while using its combined autograd gradient from both roles.

The SR quantizer samples between adjacent in-range codes with probabilities whose expectation equals the unclamped proposal. An outward proposal at the maximum finite code is clamped to that endpoint and counted as staying with probability one; thus unbiasedness is an in-range statement, not a claim beyond the representable finite interval. The matched 10k artifact records initial normal fractions (99.5% for E5M2 and 99.0% for E4M3) but not trajectory-level normal, subnormal, or endpoint occupancy, so it cannot quantify how often these boundary semantics are exercised later in training.

C.2 Archived 3000-step policy intervention

The 3000-step intervention uses the same fp32 initialization seed, data order, optimizer, schedule, fp8 grid, and evaluation set within each seed. In the archived runs, the SR arm also uses stochastic rounding for the initial fp8 projection, so its initial codes need not equal the RNE arm's codes. The result therefore identifies the complete rounding policy, not post-update persistence alone. The completed matched follow-up below instead uses a common RNE initial projection before changing post-optimizer write-back. Recovery follows Eq. (6).

The six stored SR arms are complete realized runs under that whole rounding policy, but the historical producer traverses eligible tensor names through a Python set without recording PYTHONHASHSEED. As in the matched 10k run, the recorded training and SR seeds therefore do not uniquely determine which tensor receives each random-number block, and sensitivity to the SR realization is unmeasured. The two retained 3k artifact files record the same absolute held-out-data path and `held_out=true`, but no data-file digest. This preserves the within-seed paired contrasts while leaving byte identity of the split across the separately stored seed-42 and other-seed artifacts independently unverified.

At 3000 steps, equal attention/MLP group-mean never-moved statistics under RNE are approximately 0.626 (E5M2) and 0.443 (E4M3); under SR they collapse essentially to zero. The archived

Table 5: Held-out loss and gap recovery at 3000 steps. Losses are nats; each row uses its same-seed fp32 anchor.

Seed	Format	fp32	RNE	SR	RNE gap	Recovery
1337	E5M2	5.055	6.701	5.097	1.646	97.44%
1338	E5M2	5.029	6.703	5.102	1.674	95.64%
42	E5M2	5.069	6.700	5.138	1.632	95.76%
1337	E4M3	5.055	6.219	5.076	1.164	98.14%
1338	E4M3	5.029	6.237	5.046	1.209	98.54%
42	E4M3	5.069	6.232	5.081	1.163	98.92%

reference stay-probability proxy is $\widehat{p}_{\text{never}}(t) = \langle \prod_{s \le t}(1 - r_{s,i}) \rangle$, where $r_{s,i}$ used the symmetric upward spacing of the proposed update. At a binade boundary the corrected probability uses the directional adjacent gap. This separates cumulative persistence from the per-step frozen fraction, which can remain high under SR even though almost every coordinate eventually moves.

C.3 Matched ten-thousand-step write-back intervention

The corrected 10k experiment jointly runs fp32, RNE, and SR arms for seeds 1337/1338 at the default rate 6×10^{-4}, with a 10k cosine horizon and evaluations every 250 steps. RNE and SR begin from the same target-grid projection, and the only policy difference is post-optimizer write-back. The global model/data RNG is reset to the training seed for every arm, the fixed evaluation batches use seed 4242, and SR decisions use a dedicated device generator seeded by the training seed plus one. Consequently the study has one SR realization per training seed, not repeated write-back randomness at a fixed training trajectory. Because the historical producer iterates eligible tensor names through a Python set without recording PYTHONHASHSEED, the seeded stream is assigned to tensors in process-dependent order; exact SR trajectories are therefore not reconstructable from the recorded seeds alone. The OpenWebText binaries, source overlay, A100 environment, arm inventory, trajectories, coordinate weights, and final losses pass the predeclared analyzer.

Table 6: Matched held-out losses and recovery at step 10,000. Every row uses a same-seed raw-initialization fp32 anchor; RNE and SR share common projected target-grid codes.

Seed	Format	fp32	RNE	SR	RNE gap	SR gap	Recovery
1337	E5M2	4.283	6.412	4.512	2.129	0.229	89.24%
1338	E5M2	4.275	6.419	4.501	2.144	0.226	89.43%
1337	E4M3	4.283	5.710	4.395	1.427	0.112	92.16%
1338	E4M3	4.275	5.714	4.389	1.439	0.114	92.07%

The predeclared point-estimate decision rule required recovery of at least 90% in both formats and both seeds. It classifies only the two E5M2 cells below threshold. This deterministic protocol rule is not a population-level hypothesis test. Within the same artifact, recovery changes from 93.65/93.44% at step 3001 to 89.24/89.43% at step 10,000 for E5M2, and from 95.59/97.74% to 92.16/92.07% for E4M3. The coordinate-weighted attention/MLP never-moved fractions at 10,000 are 0.599/0.599 under RNE E5M2 and 0.413/0.413 under RNE E4M3; every SR cell is at or below 1.18×10^{-8}. This establishes substantial but horizon-dependent recovery and separates elimination of deterministic never-moved persistence from complete loss-gap removal.

The historical RNE-only 10k extension independently reports final gaps of 2.127/2.155 nats for E5M2 and 1.425/1.450 for E4M3 on its legacy split. Its archived symmetric-gap proxy gives measured and predicted at-init crossing in all 24 level–target–seed combinations, with RMSE 0.004–0.006 for pooled and attention/MLP group-mean scopes and up to 0.030 across all reported groups. Those predictor values remain legacy-proxy evidence; the matched follow-up does not splice its losses to that file.

C.4 Fixed-design modern-decoder transfer

The separate locked test changes the architecture bundle, not one component: 12 layers, width 768, 12 heads, RMSNorm, rotary position encoding, SwiGLU with intermediate width 2048, and un-

tied 50,304-by-768 input and output matrices. The exact inventory is 162,220,800 parameters, of which 162,201,600 2-D coordinates use the analytic E5M2 grid. It retains the GPT study's fp32 compute/state, AdamW hyperparameters, batch size, context, and 3000-step schedule. Each arm observes 36,864,000 training token positions. RNE and SR start from common RNE-projected codes and differ only in post-optimizer write-back. "Core" is the coordinate-weighted union of the 84,934,656 attention and SwiGLU coordinates; "all 2-D" contains all 162,201,600 eligible coordinates.

Table 7: Fixed-design modern-decoder outcome at step 3000. Never-moved columns report RNE/SR percentages; recovery uses the same-seed fp32 and RNE anchors.

Seed	fp32	RNE	SR	RNE gap	Recovery	Core	All 2-D
271828	4.707	6.261	4.817	1.554	92.94%	65.87% / 1.18×10^{-6}%	71.17% / 0.181%
314159	4.729	6.263	4.827	1.534	93.61%	65.83% / 0%	71.15% / 0.180%

Both seeds pass reference learning, a material RNE anchor, clean finite visibility, recovery, and never-moved-contrast gates. Finite-grid endpoint occupancy is zero in every RNE/SR arm. The registered category is nevertheless `policy_only_transfer`. Writing E for RMSE and V for the measured post-warmup 95th-minus-5th percentile range, accuracy requires $E_{\text{dir}} \leq 0.05$, informativeness requires $V \geq 0.10$ and $E_{\text{dir}} \leq 0.8E_{\text{const}}$, and directional added value requires $E_{\text{dir}} \leq 0.9 \min(E_{\text{sym}}, E_\rho)$ in both seeds and scopes. Table 8 shows that the range never reaches 0.10 and all four directional-added-value criteria fail. Low absolute RMSE against a nearly constant trajectory is descriptive agreement, not substantive transfer of the diagnostic.

Table 8: Registered forecast diagnostics by seed and co-primary scope. "Range" is the measured post-warmup 95th-minus-5th percentile range; the four comparison columns are RMSEs. "Constant" is the hindsight-best constant fit to the measured trajectory. "Info/add." reports the informativeness and directional-added-value gates.

Seed	Scope	Range	Directional	Constant	Symmetric	Scalar ρ	Info/add.
271828	Core	0.05212	0.008770	0.018607	0.008806	0.028405	No / No
271828	All 2-D	0.03179	0.009346	0.011387	0.009374	0.017242	No / No
314159	Core	0.05290	0.008851	0.018604	0.008887	0.028383	No / No
314159	All 2-D	0.03232	0.009382	0.011393	0.009409	0.017234	No / No

The required arithmetic mean [range] across the two fixed seeds is 4.71785 [4.70719, 4.72852] for fp32 loss, 6.26195 [6.26122, 6.26268] for RNE loss, 4.82169 [4.81683, 4.82656] for SR loss, 1.54409 [1.53415, 1.55403] for the RNE gap, and 0.93277 [0.92944, 0.93610] for recovery. Core RNE/SR never-moved fractions are 0.658526 [0.658342, 0.658710] and 5.89×10^{-9} [0, 1.18×10^{-8}]; all-2-D values are 0.711613 [0.711521, 0.711705] and 0.001806 [0.001803, 0.001808]. For (measured range, directional, constant, symmetric, scalar-ρ RMSE), the corresponding means are (0.05251, 0.008810, 0.018605, 0.008847, 0.028394) for core and (0.03205, 0.009364, 0.011390, 0.009391, 0.017238) for all 2-D; the per-seed ranges are given by Table 8. Gate outcomes are identical across seeds: forecast accuracy passes, while informativeness and directional added value fail in both scopes.

C.5 NeoX/Pythia-style paired-format transfer

The final write-back bundle fixes a GPT-NeoX/Pythia-style decoder with 12 layers, width 768, 12 heads, LayerNorm, parallel residuals, GELU MLPs of width 3072, 25% partial RoPE, and untied embeddings/output weights. It has 162,220,800 parameters and 162,201,600 eligible 2-D coordinates. Three target training seeds are paired across analytic E5M2 scale 1 and E4M3 scale 128; the scientific units are therefore three targets under two fixed format conditions, not six independent seeds. Each seed-format cell contains a projected-start fp32 reference, direct RNE, SR, ECO, and blanket master, for 30 runs and 90,000 updates.

The fp32 reference begins at the same RNE-projected grid values as its four treatments and then evolves in fp32, so these recovery denominators use a projected fp32 anchor rather than the raw-fp32 anchor of the matched GPT-2 and RMSNorm-decoder studies. SR changes the stochastic write-back decision; ECO retains RNE write-back and feeds its error into AdamW's first moment. Both pass

Table 9: NeoX/Pythia-style recovery and directional frozen-fraction agreement at update 3000. Brackets span the same three target seeds; RMSE is computed on the core attention/MLP trajectory.

Format	RNE–fp32 gap (nat)	SR recovery	ECO recovery	Directional RMSE
E5M2	1.721–1.754	0.958 [0.948, 0.968]	0.819 [0.809, 0.833]	0.00646 [max 0.00654]
E4M3	1.167–1.204	0.978 [0.967, 0.984]	0.957 [0.955, 0.961]	0.00815 [max 0.00826]

the registered absolute recovery gates in every cell. They are policy contrasts, not interchangeable implementations of stochastic rounding.

The historical policy-and-reference-audit-transfer classification is retained because the directional RMSE passes the registered absolute 0.15-per-cell and 0.10-mean gates. Independent validation reproduces that label, but a harmonized post hoc audit shows why the paper uses narrower wording. A fixed-one predictor also passes the same absolute gate (mean/max RMSE 0.05134/0.06940). Under the earlier post-warmup rule, E5M2 core fails the 0.10 measured-range informativeness floor in every target, whereas E4M3 core passes it in every target. Directional added value is untestable because symmetric and scalar-ρ trajectories were not retained. We therefore report descriptive absolute-RMSE agreement, not validated diagnostic transfer.

The sealed optimizer declares peak/minimum rates 6×10^{-4} and 6×10^{-5}, a 100-update warmup, and cosine decay through update 3000. Its implemented zero-based indexing reaches 5.940594×10^{-4} on update 100 and the peak on update 101; update 3000 is 1.5843×10^{-10} above the minimum. This tiny schedule deviation is common to all cells. Independent validation authenticates all 30 records, but the required full-shape readiness receipt was not retained, the pre-run manifest does not bind the protocol or launcher bytes, and fresh-process and physical-GPU identity are not authenticated. The result is consequently scoped to this retained analytic write-back bundle and does not establish componentwise architecture effects, seed or SR populations, native FP8 compute, distributed training, throughput, or production cost.

C.6 Source-amortized E4M3 audit and sealed template challenge

The prospective audit uses a fresh source seed and three fresh target seeds, all excluded from the historical scientific inventory. The source fp32 trace is completed and receipt-sealed before any target run starts. Each target then has one direct-RNE outcome and one privileged target-specific fp32 reference, for seven runs and 21,000 updates. Training uses the fixed NeoX/Pythia-style architecture above, analytic E4M3 at scale 128, microbatch 1×12 accumulation, and math-only scaled-dot-product attention. All seven training validity checks pass and no finite endpoint is observed.

The registered primary scope is the coordinate-weighted attention/MLP core over updates 101–3000. The reused source directional predictor passes every absolute accuracy and skill gate in all targets: its mean/maximum RMSE is 0.00858/0.00882 and its minimum skill against the hindsight-best constant is 0.9568. The privileged matched-reference regret gates also pass. Directional specificity fails in all targets because the directional predictor is worse than the symmetric alternative. The canonical classification therefore remains `source_amortized_reference_audit_supported`, with directional specificity explicitly unsupported.

The secondary challenge was fixed without a target prefix, calibration, smoothing, lag, weighting, or candidate selection. Let X_S be the fresh source directional trajectory, and let B_Y and B_X be the pointwise means of three archived direct-RNE and matched-fp32 trajectories. Its primary source correction is

$$H_S = X_S + (B_Y - B_X). \tag{16}$$

For target j, the privileged comparator substitutes its matched reference X_{Mj} to form $H_{Mj} = X_{Mj} + (B_Y - B_X)$. The predictor formulas and candidate inventory were committed before the first direct-RNE outcome completed, but the fixed primary contrasts and authenticated consumer were committed later. After the canonical bundle was pulled, that consumer reconstructed and receipt-sealed every predictor before opening any target-RNE file, then evaluated exactly that seal. No secondary threshold modifies a registered gate.

Table 10: Sealed temporal-template challenge on three fresh E4M3 targets. Metrics are equal-weight means of per-target core trajectories over updates 101–3000. X_M and H_M use privileged target-specific fp32 references.

Predictor	RMSE	Centered RMSE	Signed error	ΔRMSE vs. B_Y
B_Y historical outcome template	0.003599	0.003585	$+0.000144$	0
X_S fresh reused source	0.008580	0.005207	-0.006818	$+0.004981$
H_S source hybrid	0.004379	0.004370	$+0.000028$	$+0.000780$
X_M matched fp32	0.007764	0.003914	-0.006699	$+0.004165$
H_M matched hybrid	0.002829	0.002806	$+0.000147$	-0.000770

The ordering $B_Y < H_S < X_S$ holds in RMSE for every target. The matched hybrid H_M improves sum-squared error over B_Y by 38.1%, whereas H_S worsens it by 48.1%. A separate post-outcome decomposition, not a new decision rule, attributes 99.650% of fresh direct-RNE sum of squares to a common time effect and 0.346% to target-by-time residuals. The fresh outcome common mean correlates 0.99815 with B_Y; matched-reference target-specific deviations correlate 0.669–0.712 with measured deviations. Thus aggregate temporal structure is highly reusable, and matched reference deviations contain target-specific signal, but the registered unit-gain fresh-source predictors do not add it.

The archived template used physical batch 12 without locked math-only attention, whereas the fresh audit used microbatch accumulation and math-only attention. Its lower RMSE in this fixed comparison, despite this mismatch, challenges the value of the registered raw and unit-gain fresh-source predictors for this fixed observable; a loss would not have excluded an exact-runtime template. The decomposition is descriptive, time points are not independent scientific units, three fixed targets support no seed-population inference, and a dominant common temporal effect does not identify the schedule as a physical cause.

C.7 Persistent-native Study 1 protocol and provenance

A distinct registered schedule-shift rule is applied to three fixed target seeds paired across two schedules. "constant-mid" linearly warms for 100 updates to 3×10^{-4} and then holds that rate; "cosine-restart" warms for 100 updates to 6×10^{-4} and runs two cosine cycles between 6×10^{-4} and 6×10^{-5} over the 3,000-update horizon. Five target cells are canonical campaign records; cosine-restart target 1 is a manual recovery whose producer retains `scientific_result=false` and does not retain canonical process identity. The registered protocol explicitly forbids promoting that diagnostic record, so its classification remains `inconclusive_invalid_native_source_audit`. We neither rerun the cell nor describe the resulting six records as an uninterrupted canonical campaign.

A separate post-outcome retrospective protocol-deviation analysis uses an independent certificate of full occurrence transport, implementation identity, raw artifacts, and checkpoint-state replay; it does not make the manual record registered. Anti-selection follows from the cohort structure: cosine-restart target 1 is disjoint from the complete constant-mid cohort and cannot alter its failing macro and centered per-cell gates. Adding that cell therefore cannot rescue the joint rule to a positive result; it only permits a descriptive retrospective six-cell negative. In every cell, 50 two-dimensional matrices are stored persistently as E4M3 codes and consumed by native hybrid-FP8 dense GEMMs. The source-to-strongest-source-free centered-RMSE ratio is 1.0525 for constant-mid, failing its gate, and 0.8550 for cosine-restart, passing its gate. The joint rule requires, in both cohorts and all six cells, every raw Kineto scan, raw and centered cell gate, positive residual-correlation gate, training-validity gate, and recovery/ checkpoint gate. The constant-mid ratio misses the 0.9 macro gate and at least one centered per-cell gate fails. To keep this protocol deviation distinct from the frozen protocol's formal negative label, the retrospective analysis uses the independent label `post_outcome_retrospective_native_source_increment_not_supported_across_schedule_shift`. The manual record lacks training start-ticks/nonce evidence and an independent update-1000 replay. These are provenance limitations for the post-outcome retrospective protocol-deviation analysis; its existing quality gate uses the independently replayed update-0 and update-3000 endpoints. The two classifications therefore answer different

questions rather than promoting the diagnostic record. Both are limited to the three fixed targets and the stated persistent-matrix/native-GEMM hybrid; neither is seed-population evidence or a full-native-FP8 training claim.

C.8 Native E4M3 software-rounding intervention

A fixed 12-arm Study 2 pairs three training targets across fp32, persistent-native RNE, full SR, and dead-zone-only SR for 3,000 updates. All arms pass the training, common-start, coupled-stream, checkpoint, ledger, replay, and raw-artifact gates; every native arm also passes the authenticated SM120 dispatch check. Per-target full-SR recoveries are 0.9862/0.9743/0.9694 (mean 0.9766), while dead-zone-SR recoveries are 0.9764/0.9949/0.9619 (mean 0.9777); the corresponding dead-zone/full-SR improvement ratios (the registered fixed-run protocol metric, not a formal causal mediation estimand) are 0.9901/1.0212/0.9923 (mean 1.0012). The registered classification is therefore `native_deadzone_mediates_full_sr_recovery`, a protocol label rather than a causal mediation proof. The scientific units are the three fixed targets, and the stochastic streams are software-generated rather than hardware stochastic rounding.

C.9 DQT Study 3 wrapper and recovery lineage

Study 3 fixes one 134.1M-parameter DQT Llama configuration (Zhao et al., 2025), the Wikipedia 20231101.en corpus, seed 42, and 55,670 optimizer updates. Model values are signed-INT3 grid values held in fp32 containers; forward and backward use mixed fp16 compute with the fixed signed-INT8 activation quantizer, while Adam moments remain fp32. Globally token-weighted validation cross-entropy is evaluated at updates 0/11,134/22,268/33,402/44,536/55,670. The frozen positive rule requires an initial-to-final decrease of at least 1.0 nat and final cross-entropy at most 6.0. Independent replay of the six actual checkpoints gives mean-token cross-entropies 10.4825/3.3788/3.3156/3.2521/3.2018/3.1754 nats. The initial-to-final decrease is 7.3071 nats, or 69.71% descriptively, and both gates pass. The validated classification is `dqt_direct_state_int3_convergence_supported`. These values come from checkpoint replay, not losses copied from the training ledger.

The fixed training wrapper uses AdamW with learning rate 5×10^{-4}, betas $(0.9, 0.999)$, epsilon 10^{-6}, weight decay 0.01, global batch size 256, and a 2,000-update warmup followed by cosine decay to zero at the 55,670-update horizon. The scheduler advances after each optimizer update, so zero is its terminal state, not the rate applied to the final update. Training is cache-free with eager attention and non-reentrant activation checkpointing. These wrapper choices, together with the fixed split, horizon, and evaluation replay, define the reported protocol and are not claims about the upstream training script.

The original fixed-scaler process checkpointed update 11,134, completed updates through 11,850, and overflowed while attempting update 11,851. Its manual continuation restores the model, Adam moments and steps, schedule, data cursor, stochastic-rounding stream, and loss-scaler state. It retains parent updates 1–11,134, exactly reproduces every preserved per-rank optimizer-evidence payload for updates 11,135–11,850, and retries the same update-11,851 batch after loss-scale backoff (65,536 to 32,768) without advancing the data cursor, scheduler, or stochastic-rounding stream. This continuation is part of the same fixed-seed run, not another scientific unit. During the continuation, the same retry rule also encountered update 46,003 twice: its first retry backed off 32,768 to 16,384, and its second retry backed off 16,384 to 8,192. Both retries reused the same batch and left the data cursor, scheduler, and DQT stochastic stream unchanged. The recovered suffix is therefore an engineering recovery continuation of the registered run, not an uninterrupted fresh execution or a second independent seed.

C.10 Prospective Pythia-410M selective-forecast gate

This separate stage uses released Pythia-410M-deduped checkpoints at steps 7000, 72000, and 129000. Each age has three historical reference/target pairs, one fresh reference source, three development targets, and five untouched holdout targets: 45 branches of 512 updates each. AdamW state is reset at the released checkpoint; this is an optimizer restart, not continuation of the original pretraining optimizer. The fixed batch is 66 contexts of 2048 tokens per update. References retain

unrestricted fp32 weights, with BF16 forward/backward compute and fp32 optimizer state. Targets apply analytic E4M3FN nearest-even write-back at scale 128 after each update, storing decoded grid values in fp32 containers without an unconstrained master. This does not test native FP8 storage or throughput.

Training contexts are drawn from a fixed 2^{30}-token Pile training stream; the separate validation stream contains 2^{26} tokens. Historical pairs share their training schedule; fresh sources and targets use separately drawn schedules. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, and global gradient clipping at 1.0. Moments and optimizer step counters reset, but the learning-rate schedule does not: audit update u at checkpoint s uses virtual step $s + u$, with maximum/minimum rates $3 \times 10^{-4}/3 \times 10^{-5}$ and the original 1,430-step warmup and 143,000-step cosine horizon. No fresh local warmup is introduced. This design uses released weights with a common fresh optimizer state, preserving the checkpoint-age learning rate without restoring the original pretraining optimizer state. Bias correction remains enabled. Resetting moments at that rate can alter early proposals and source–target alignment; applying the same reset to all branches does not establish that these effects cancel. The comparisons therefore concern this restart regime. We do not measure the reset's causal effect or extrapolate these failures to uninterrupted pretraining or a restart with fresh warmup.

Event labels and geometric selection. For a reference coordinate, set $q = Q(w^{\text{ref}})$, $z = q + d^{\text{ref}}$, and $y = Q(z)$, where d^{ref} includes the full AdamW proposal, including decay. Predict unchanged when $y = q$; the actual label is the target's own code equality before and after its update. For an interior destination code y with adjacent codes y^-, y^+, let $\ell = (y^- + y)/2$ and $u = (y + y^+)/2$. The fixed score is

$$c = \text{clip}_{[0,1]} \frac{2 \min(z - \ell, u - z)}{u - \ell}, \qquad b = \min(255, \lfloor 256c \rfloor). \tag{17}$$

At either finite endpoint, use the clipped distance from its only midpoint boundary toward and beyond the endpoint, normalized by half the adjacent-code gap; the score is zero at the boundary and one at and beyond the endpoint. Zero uses both adjacent subnormal codes. Exact midpoint ties have score zero. This measures a local rounding-cell margin, not a predictive probability or a bound on path drift. The candidates are $\{0, 16, \ldots, 240, 255\}$; acceptance is $b \geq \tau$. The selected bin $\tau = 32$ is equivalent to $c \geq 1/8$.

Coordinate sampling and numerical bounds. Before target outcomes, one protocol-keyed uniform sample without replacement selects $k = 2^{20}$ coordinates from the $N = 405{,}012{,}480$ eligible trainable 2-D coordinates per checkpoint age. The same sample is reused across the source and same-age branches and across all $H = 512$ updates. For sampled coordinate i, A_i is its accepted-update count divided by H, and E_i its accepted-and-incorrect count divided by H. Let bars denote sample means. The implemented Serfling radius (Serfling, 1974) and ratio are

$$\begin{aligned} r_\alpha &= \sqrt{\frac{(1 - (k-1)/N) \log(1/\alpha)}{2k}}, \quad A_L = \max(0, \bar{A} - r_\alpha), \\ E_U &= \min(1, \bar{E} + r_\alpha), \quad R_U = E_U / A_L. \end{aligned} \tag{18}$$

Nonpositive A_L fails. Time points are dependent measurements within a sampled coordinate trajectory; they are not separate samples or training replications. Temporal X_t, Y_t instead use the full eligible-coordinate population at each update without confidence filtering.

The original numerical component uses $\alpha = 0.025$ for each of the two bounds. Those fixed-threshold formulas do not adjust for all 17 candidates. Because development chooses the threshold using the same coordinate sample, we report its recorded component-gate outcomes without asserting a selection-valid simultaneous 95% guarantee. The prospective amendment uses $\alpha = 0.05/(6 \cdot 17)$ for coverage, risk, and four paired baseline contrasts, uniformly over the grid for each fixed target. This does not give simultaneous inference across targets or independent-source generalization.

The lowest candidate passing on all nine development targets is frozen before holdouts; the selected bin is 32. The original selective gates require an error upper bound at most 0.059 and a coverage lower bound at least 0.50. These are study-specific cutoffs, not universal accuracy requirements. The original temporal gate already requires strictly lower raw and centered RMSE for H_S than B_Y in all four variants; it fails on every holdout. A prospective amendment additionally requires 10%

relative and 0.001 absolute error improvement over both constant event labels on the same accepted events. These baselines share the selector's acceptance mask. The temporal forecast compares the fixed hybrid $H_S = X_S + (B_Y - B_X)$ with B_Y, using all three historical pairs and each leave-one-out template. The amendment requires 10% relative and 10^{-4} absolute RMSE improvement in both raw and time-centered trajectories for every variant. Original decisions are retained separately from these stricter requirements.

For completeness, each event-baseline requirement uses both paired contrasts $\overline{E - B + 0.001A} + 2r_\alpha \leq 0$ and $\overline{E - 0.90B} + 1.9r_\alpha \leq 0$, where B_i is the baseline's accepted-error count divided by H. The range widths are 2 and 1.9. Both constant labels use the original selector's acceptance mask. The amended temporal requirement is a fixed-inventory effect check; it does not estimate a population effect of at least 10%.

All 15 holdouts pass integrity and the original selective gates; none passes the original source-value or joint gate. Amended selector, source, and joint counts are all zero. The empirical accepted-event error ranges for the selector versus always predicting unchanged are 5.608–5.617% versus 3.542–3.555% at the early age, 2.383–2.390% versus 1.423–1.430% at the middle age, and 0.4407–0.4420% versus 0.2261–0.2271% at the late age. Every target favors the constant-unchanged baseline. Unchanged events occupy 96.45–99.77% of the accepted set, while error among actually changed accepted events is 76.00–81.09%. Low aggregate error therefore coexists with poor performance on the rarer changed class in this fixed inventory.

Across all 60 target/template-variant comparisons, H_S/B_Y RMSE ratios range from 1.703 to 5.475 raw and 1.725 to 5.475 after time centering. Every hybrid is worse than its template. Both the original and amended funding decisions are `NO_GO`. The fixed 243-scenario simulation also fails admission; its original-protocol scope supplies no amended-power guarantee or measured larger-model outcome. The larger-model route stops under this frozen design.

The 243 scenarios comprise $(2 \times 4 + 1) \times 3 \times 3 \times 3$ settings: Gaussian and Student-t copulas (four degrees of freedom), each with latent correlation $\{0, 0.25, 0.5, 0.75\}$, plus a Fréchet–Hoeffding bound; covariance multipliers $\{1, 1.5, 2\}$; and adverse mean shifts $\{0, 0.5, 1\}$ in pooled-sigma units, chosen independently for the hypothetical 1B and 2.8B models. Marginals use the five untouched 410M holdouts per age, pooled within-age residual scale (12 residual degrees of freedom), age-stratified residual support, and all 27 joint resamples of the three historical pairs. Residuals scale by the square root of the covariance multiplier, with metrics clipped to natural bounds. The fixed seed is 970410066 and the Monte Carlo replicate count is 100,000. The Fréchet cases bound joint probabilities from simulated marginals; they are not another copula family or additional training runs. These assumptions define a model-based extrapolation from 410M outcomes.

Training finished before an analysis error exposed a schema mismatch: a holdout barrier identifies its development barrier by digest, whereas the recomputation helper expected the template field directly. A disclosed post-outcome repair reconstructs and validates that ancestor before applying the original binding checks. Local finalization recomputed all 15 holdouts, preserved their validation reports, and retained the original method seal, threshold, statistics, and simulation settings. Thus the final report uses an explicitly repaired analysis, not the unmodified sealed executable. The descriptive ranges above summarize authenticated reports rather than new scientific tests. Five holdouts per age share one fresh source and three historical pairs; they are not independent source replications. Time centering removes each trajectory's mean, not its shared temporal shape. These results establish neither target-specific deviation prediction nor a population-wide failure of reference forecasts.

C.11 Post-outcome scalar-gain ceiling

The prospective Pythia comparison fixes the unclipped predictor $H_S = X_S + (B_Y - B_X)$. Its failure leaves open whether choosing the scalar gain differently would attain the prespecified material improvement. We answer that narrower question retrospectively, keeping the observed source, template, target, evaluation times, and original decisions fixed. The calculation uses all 15 targets and their full and three leave-one-out templates, separately for raw and centered RMSE. It introduces no additional training, fitted predictor, or confirmatory test.

Let $c = X_S - B_X$ and $r = Y - B_Y$, with inner product $\langle u, v\rangle = 512^{-1}\sum_{t=1}^{512} u_t v_t$. Write $D = \langle c, c\rangle$, $E = \langle r, r\rangle$, and $A = \langle c, r\rangle$. For $D, E > 0$, completing the square gives

$$\begin{aligned}\mathrm{MSE}(B_Y + \alpha c, Y) &= E - 2\alpha A + \alpha^2 D \\ &= D(\alpha - A/D)^2 + E - A^2/D. \end{aligned} \tag{19}$$

Consequently, the maximum possible relative RMSE reduction over all real scalar gains is

$$G = 1 - \sqrt{1 - \frac{A^2}{DE}}. \tag{20}$$

For centered RMSE, apply the same calculation after subtracting the time mean from both c and r. These are elementary finite-vector identities, not new statistical estimators. Allowing a separate oracle for every target, variant, and metric is more permissive than any one shared scalar gain. No oracle coefficient or alternative prediction is selected for deployment.

In the 60 observed target/template comparisons, maximal G is 3.3785% raw and 3.3070% centered. Every baseline and source-residual energy is positive. The corresponding maximal squared-error reductions are 6.6429% and 6.5047%; a 10% RMSE reduction would require 19% squared-error reduction. No individual comparison can therefore meet the relative requirement by changing scalar gain alone, and hence none can meet the combined relative-plus-absolute requirement that way. This conclusion does not exclude smaller improvements.

The same inputs explain the fixed unit-gain loss: $\mathrm{MSE}(H_S, Y) - \mathrm{MSE}(B_Y, Y) = D - 2A$. For each metric, 23 of 60 comparisons have nonpositive alignment A; in the other 37, positive alignment is insufficient to offset correction energy. Negative alignment can still help an unrestricted oracle through a negative gain, which Eq. (20) already allows. The export retains the residual vectors; reproduction reconstructs the original raw/centered RMSEs and recomputes the bound without treating time points or leave-one-out variants as independent replicates.

If $D = 0$ and $E > 0$, scalar gain changes nothing and the reduction is zero. If $E = 0$, relative reduction is undefined and no positive absolute improvement over zero error is possible. A free intercept is outside the raw bound; a constant intercept disappears under time centering, so it cannot evade the centered bound. Clipping, nonlinear or time-varying corrections, multiple residual directions, and different sources/templates are outside this family. The bound is a post-outcome geometric description of these observed traces, not a prospective selection rule, population estimate, or future-target guarantee. Original and amended admission remain `NO_GO`.

D Effective-mantissa phase-diagram protocol

D.1 Historical sweep and post-outcome correction

The effective-mantissa testbed is a two-layer tanh teacher–student regression with input dimension 8, teacher width 8, student width 64, 512 training examples, 512 test examples, and Gaussian training-label noise with standard deviation 0.1. Inputs receive no added noise and test labels remain teacher-generated without added noise. Every cell runs full-batch fp64 gradient descent for 3000 epochs. Stored arms round to $m \in \{1, 2, 3, 4, 5\}$ after each update using either RNE or unbiased SR. The sweep uses learning rates $\{0.0125, 0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6\}$. Three base seeds cover all rates; three additional seeds cover the four boundary rates $\{0.1, 0.2, 0.4, 0.8\}$, producing 36 fp64 references and 180 mantissa–rate–seed cells, with RNE and SR stored-weight arms in each cell. The historical SR arms use stochastic rounding for their initial projection as well as for subsequent write-back; their recovery is not a post-update-only intervention.

For test error E, the learned fraction is

$$\lambda_a = \frac{E_0 - E_{a,T}}{E_0 - E_{\mathrm{fp64},T}}, \qquad a \in \{\mathsf{RNE}, \mathsf{SR}\}. \tag{21}$$

The reference-only classifier substitutes fp64 error at the code-exact proxy's predicted persistent majority crossing for $E_{a,T}$; a predicted no-freeze cell receives 1. This is a heuristic binary classifier, not a calibrated prediction of final low-precision loss. A criterion-positive rate has median $\lambda \geq$

0.5. The robust convention scores a diverged reference or arm as zero before taking the median. Under this convention, measured and predicted RNE windows agree for $m = 2$–5 while the proxy conservatively misses one marginal $m = 1$ cell (Table 11); SR meets the criterion at every sampled stable rate, 0.0125–0.4. Exact agreement holds at threshold 0.7 but not universally at 0.3, so the paper claims window classification at the declared thresholds rather than accurate low-mantissa freeze times in every cell.

The survivor-only convention additionally admits 0.8 for each mantissa, but the extra fp64 references diverge there and the edge rests on a minority of survivors. At 1.6 every fp64 reference diverges; it is an optimizer ceiling, not a precision boundary. The original predeclared expectation of an empty $m = 1$ RNE window was falsified. The retained result is the measured upward compression of the RNE window. The post-outcome implementation correction replays all 36 archived fp64 references with the code-exact proxy, leaves the $m = 2$–5 robust windows unchanged, and predicts only $\{0.4\}$ at $m = 1$, conservatively missing the measured rate-0.2 cell whose median learned fraction is 0.5195.

Table 11: Historical measured and post-outcome code-exact fp64-proxy criterion-defined RNE learning-rate windows at threshold 0.5.

Mantissa bits	Measured RNE	Predicted RNE
1	$\{0.2, 0.4\}$	$\{0.4\}$
2	$\{0.2, 0.4\}$	$\{0.2, 0.4\}$
3	$\{0.1, 0.2, 0.4\}$	$\{0.1, 0.2, 0.4\}$
4	$\{0.05, 0.1, 0.2, 0.4\}$	$\{0.05, 0.1, 0.2, 0.4\}$
5	$\{0.025, 0.05, 0.1, 0.2, 0.4\}$	$\{0.025, 0.05, 0.1, 0.2, 0.4\}$

D.2 Locked midpoint interpolation

The locked follow-up fixes code and decision rules in a protocol snapshot, then runs 18 single-thread CPU fp64 references at six held-out geometric-midpoint rates and seeds 2027–2029 within the same testbed. Its prediction-only artifact is committed and hashed before any of the 90 declared RNE cells run. The measured-run phase requires the committed blob hash, unchanged source hashes, identical CPU/software/thread metadata, and exact data and initialization fingerprints. Recomputing aggregate analysis from the locked result cells gives window matches of 5/5 for directional, symmetric, and scalar-ρ predictors, versus 0/5 for the fixed monotone baseline (Table 12). Promotion requires at least 4/5 directional window matches, at least 80% of eligible onset ratios within 15%, and directional cell accuracy no worse than the symmetric and monotone baselines; scalar-ρ is reported but is not a gate, and including it would leave the negative decision unchanged. The rule fails because directional cell accuracy is 94.4% versus 95.6% symmetric. The onset statistic covers 55/90 cells; the remaining 35 measured arms cross at step 1, including all five directional functional false negatives. A deterministic post-outcome sidecar decomposes those misses without changing the locked result.

All five false negatives occur at seed 2028's lower-edge rates. Those arms are majority-frozen from step 1 yet learn 0.524–0.596 of the reference improvement. Across them, 54.9–56.5% of coordinates move at least once although only 37.6–42.6% move on any one step, and normalized progress of 0.166–0.593 accrues after the predicted cutoff. Three also contain reference-path onset error. The decomposition therefore does not identify one exclusive cause: reference-path error and continued progress by a changing minority can coexist.

Table 12: Locked interpolation results. Cells sharing one of 18 references are paired coverage tests, not independent replicates. "Onset" is the predeclared subset with both crossings defined and measured $3 \leq \tau < T$. Scalar ρ was reported but was not a promotion gate; the monotone baseline has no onset prediction.

Proxy	Window sets	Cell accuracy	Onset within 15%
Directional code equality	5/5	94.4%	51/55
Symmetric spacing	5/5	95.6%	51/55
Scalar ρ	5/5	100.0%	50/55
Fixed monotone baseline	0/5	83.3%	—

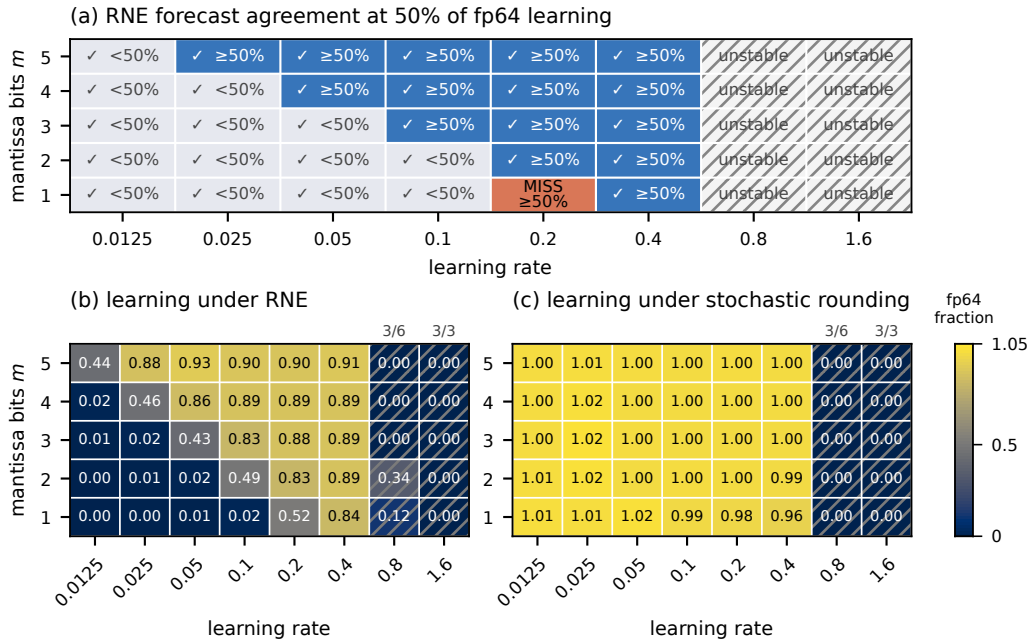

Figure 4: Within this fixed historical grid, the code-exact proxy misses one usable RNE cell, while SR retains near-fp64 learning across the stable low-mantissa cells. The scientific units are fixed seed-runs: $n = 3$ per cell at rates 0.0125–0.05 and 1.6, and $n = 6$ per cell at rates 0.1–0.8. **(a)** Check marks denote agreement above or below the 0.5 learned-fraction criterion; orange marks the sole forecast miss. **(b,c)** Cell values are medians of the fraction of fp64 learning retained under RNE and SR; 0 means none and 1 means matched, while values above 1 indicate lower final error than the fp64 anchor in this finite run. No range or interquartile range is drawn. Hatching marks cells containing one or more divergent fp64 references (the header gives diverged/total), and diverged runs enter the median as zeros; thus at $m = 2$, rate 0.8, three zeros and three finite RNE values give 0.34. No inferential intervals or hypothesis tests are reported.

E Precision-allocation protocols and controls

E.1 Historical whole-tensor study

The selective-protection study uses GPT-2-124M, OpenWebText, E5M2, seeds 1337/1338, and the same 3000-step recipe as the archived policy contrast. A separate 600-step fp32 reference ranks each decay tensor by the fraction of coordinates below the archived symmetric half-gap proxy at every ranking step. This costs 20% of one diagnostic training trajectory and contains no low-precision observations. It predates the directional correction; replaying the reference is required to know whether the selected sets change. The measured interventions, including the policy-independent tied-matrix-only arm, remain valid. Selection greedily adds whole tensors in descending risk order; anti-selection reverses the order. Because tensors are atomic, nominal budgets do not guarantee equal achieved memory.

For a protected tensor, the forward/backward parameter remains on the fp8 grid, but AdamW updates an fp32 master copy before requantizing it for the next step. That master is initialized from the raw preprojection fp32 parameter before the forward copy is projected; direct RNE retains only the projected code. Historical E1 recovery therefore combines preservation of this initialization residual with accumulation of later proposals, rather than isolating a common-projected-start write-back intervention. Memory is reported as protected parameter count divided by blanket protected parameter count. Recovery uses Eq. (6) with the corresponding same-seed frozen and fp32 anchors.

Table 13: Key selective-protection controls. Each entry is achieved master-weight memory / recovered loss gap for seeds 1337 and 1338.

Arm	Seed 1337	Seed 1338
Blanket	1.000 / 1.005	1.000 / 1.001
Risk select, nominal 0.25	0.258 / 0.274	0.262 / 0.273
Anti-select, nominal 0.25	0.266 / 0.082	0.266 / 0.078
Tied matrix only	0.311 / 0.559	0.311 / 0.552
Risk select, nominal 0.50	0.715 / 1.010	0.715 / 1.003
Tied matrix excluded, 0.50	0.504 / 0.309	0.504 / 0.291

The tied matrix's forecast risk is 0.558/0.560, placing it mid-ranking despite its high loss leverage. Its 55–56% recovery at 31.1% memory demonstrates sufficiency for roughly half the gap, not full recovery and not separation of its input and output roles. The drop in the 0.50 exclusion arm is not a clean ablation because achieved memory also falls by 21 percentage points.

Five random greedy draws per seed and budget were targeted to the selector's achieved parameter count. Only eight of 20 fall within 0.01 achieved-memory fraction; among those, select wins six and narrowly loses two. These are random masks nested within two training seeds, not eight independent training replicates. We therefore claim neither superiority nor noninferiority against random. The clean equal-memory comparison is risk selection versus anti-selection near 0.26 memory, where selection recovers 0.273–0.274 versus 0.078–0.082. A claimable allocator requires tied-weight-aware row/block granularity, exact byte budgets, and a predeclared statistical comparison. These limitations motivate, but do not validate, the registered tile screen below.

E.2 Registered exact-payload tile screen

The locked follow-up partitions the 124,354,560 unique physical 2-D decay coordinates into 161,920 contiguous 768-coordinate tiles without crossing a parameter boundary. Each matrix is flattened in logical row-major order via `.contiguous().view(-1)`: coordinate (r, c) maps to $r\, n_{\text{cols}} + c$, regardless of its original storage strides. The tied input/output matrix is enumerated once. A 600-step fp32 reference for each deployment seed records the clipped minibatch-loss gradient g_{ti} and the full AdamW proposal u_{ti}, including decoupled weight decay. With q_{ti} the projected E5M2 code, it defines

$$b_{ti} = \mathbf{1}[Q(q_{ti} + u_{ti}) = q_{ti}]\,, \qquad d_{ti} = \max(0, -g_{ti}u_{ti}). \tag{22}$$

For each tile, freeze-only averages b_{ti}, leverage-only averages d_{ti}, and the registered primary combined score averages $b_{ti}d_{ti}$ over the 600 steps and 768 coordinates. The exact code-equality event includes directional binade gaps and ties. The leverage quantity is a clipped minibatch-loss descent proxy induced by the AdamW proposal, not the exact first-order change of a fully specified regularized objective.

The 10% and 20% budgets contain exactly 16,192 and 32,384 tiles, or 49,741,824 and 99,483,648 bytes of fp32-master payload. Every deterministic policy and each of five fixed random masks selects exactly the same tile count at a budget. This accounting excludes selected-index metadata, dense temporary storage, allocator runtime, and device-level fragmentation. It therefore does not establish physical memory or throughput savings.

The E5M2 pilot uses deployment seed 2027 for 1000 steps under the fixed 3000-step schedule. Its fp32 and RNE anchors differ by 1.049 nats, and blanket master weights recover 0.993 of that gap. Table 14 reports the exact-budget arms. Combined exceeds the best registered comparator by 0.001 at 10% and is 0.002 worse at 20%; the primary gate requires a $+0.03$ advantage at one budget. The archived result is therefore an informative negative, and the locked follow-up phase was not run.

Leverage-only is above all five random masks at every recorded checkpoint from step 200 through 1000, but the unit remains one deployment seed. Moreover, combined and leverage-only select 96.1% and 97.4% common tiles at the two budgets, respectively. These diagnostics explain why

Table 14: Gap recovery in one registered 1000-step screening seed at exact fp32-master payload. Random values are the median [range] over five fixed allocation masks, not training replicates. The primary combined selector fails its predeclared promotion gate.

Policy	10% payload	20% payload
Freeze-only	0.024	0.024
Leverage-only	0.706	0.827
Combined	0.708	0.825
Random (five masks), median [range]	0.159 [0.152, 0.167]	0.265 [0.251, 0.270]

freeze weighting adds little here; they do not promote the comparator after observing the primary failure or reopen the prohibited combined-primary grid.

E.3 Separately locked leverage-primary fixed-design follow-up

The subsequent protocol treats leverage-only as a new hypothesis selected after the seed-2027 screen. Before any low-precision outcome for deployment seeds 2028–2030, it binds source and data hashes, initialization and training schedules, reference-derived selections, five budget-specific random masks, evaluation tensors, exact payloads, attempt order, validity checks, and scientific gates. Each mask identity is reused across the three training seeds, so training seed—not mask—is the scientific unit. The five 10% masks and five 20% masks are distinct pre-fixed controls rather than samples from a declared population.

Stage 1 contains 27 attempts: fp32, RNE, blanket master, leverage at 10%, and five 10% masks for each seed. It requires every leverage recovery to be positive, mean recovery at least 0.50, mean/minimum advantage over the within-seed mask median at least 0.25/0.10, and leverage to beat the best mask in at least two seeds. It also requires an anchor gap above 0.1 nats and blanket recovery above 0.8 in every seed. Independent validation passes all gates and commits before Stage 2 activation.

Stage 2 adds 21 attempts: leverage at 20%, five 20% masks, and a non-gating tied-matrix arm for each seed. Its corresponding thresholds are 0.65 mean leverage recovery, 0.25/0.10 mean/minimum advantage, two best-mask wins, and a mean 20%-minus-10% recovery gain of 0.05. The observed means are 0.812, 0.558, and 0.098, respectively, with best-mask wins in all three seeds. The 10% stage has mean recovery 0.714 and mean advantage 0.552, also with three best-mask wins. Table 15 reports every seed-level estimand and fixed-mask range.

Table 15: Separately locked leverage fixed-design follow-up. Payload is a fraction of blanket 2-D fp32-master bytes. Seed rows show median [min, max] over five pre-fixed random allocation controls; Mean rows average seed-level quantities. Masks are neither population samples nor training replicates. ΔR compares 20% with the same seed's 10% leverage arm.

Payload	Seed	R_{lev}	Fixed median [range]	A_s	ΔR
10%	2028	0.712	0.159 [0.141, 0.168]	+0.553	—
10%	2029	0.738	0.164 [0.146, 0.171]	+0.575	—
10%	2030	0.691	0.163 [0.153, 0.163]	+0.529	—
10%	Mean	0.714	0.162	+0.552	—
20%	2028	0.840	0.259 [0.236, 0.261]	+0.581	+0.128
20%	2029	0.812	0.249 [0.237, 0.259]	+0.563	+0.073
20%	2030	0.785	0.254 [0.246, 0.260]	+0.530	+0.093
20%	Mean	0.812	0.254	+0.558	+0.098

Both validators check receipts, immutable attempt inventories, source/data and selection bindings, recompute losses, recoveries, summaries, and decisions, and reject unregistered extras; they do not rerun training. The tied-matrix arm's mean recovery of 0.772 at 154,533,888 payload bytes is descriptive and cannot separate its embedding and output-projection roles. No selector metadata, temporary workspace, runtime, physical device memory, or throughput enters the payload accounting.

E.4 Sequential saliency-baseline challenge

The comparator follow-up was specified after the leverage-primary outcomes above were known, but before any new comparator outcome for seeds 2028–2030. It is therefore a prospective test of those new outcomes, not a blind or independent replication of leverage. At exact 10% payload, the six score baselines rank tiles by freeze frequency F, mean gradient magnitude G, mean AdamW-proposal magnitude U, mean parameter magnitude W, Taylor magnitude $T = |gw|$, or proposal contribution magnitude $P = |gu|$. Two exact-K structural controls sample within the tied matrix or preserve leverage's per-parameter tile counts while randomizing locations. Every policy uses the same tile count, initialization, schedule, and held-out evaluation tensors as the authenticated leverage arm.

Table 16: Final step-1000 gap recovery in the sequential exact-10% challenge. The three columns are fixed deployment seeds, not population samples. The leverage arm is authenticated reused evidence; all other rows are newly measured comparator outcomes.

Policy	2028	2029	2030	Mean
Leverage L	0.7123	0.7385	0.6914	0.7141
Freeze F	0.0310	0.0208	0.0123	0.0214
Gradient magnitude G	0.6073	0.6193	0.5904	0.6056
Update magnitude U	0.1703	0.1742	0.1864	0.1770
Parameter magnitude W	-0.0140	-0.0218	-0.0242	-0.0200
Taylor magnitude T	0.6512	0.6451	0.6440	0.6467
Abs. proposal contribution P	0.7165	0.6864	0.6779	0.6936
WTE-restricted random	0.1614	0.1444	0.1526	0.1528
Composition-matched random	0.1552	0.1406	0.1477	0.1478

All 24 new physical attempts complete, the anchor and blanket-master validity checks pass in every seed, and an independent validator reconstructs the artifact inventory, selections, recoveries, maxima, and decisions. Absolute proposal contribution is the maximum among the six score baselines in every seed. Leverage's margins over that maximum are -0.00418, 0.05203, and 0.01347, with mean 0.02044; this fails both the strictly positive margin in every seed and the mean-margin threshold of 0.05. The best structural control is WTE-restricted random in every seed, giving leverage margins 0.55091, 0.59411, and 0.53879, with mean 0.56127. Although the structural-family conditions pass, the predeclared decision is their intersection with the failed score-family conditions. The valid result is therefore negative. Individual policy differences are descriptive, and the conditional 20% stage is not authorized.

A deterministic audit performed after these outcomes compares the locked tile masks without changing any gate. At 10%, leverage and absolute proposal contribution share 96.39–96.76% of selected tiles (mean 15,640 of 16,192), whereas leverage and freeze-only share 0.62–0.82%. Mean selected-tile group fractions (tied matrix / position embedding / attention / MLP) are 0.479/0.063/0.223/0.235 for leverage, 0.445/0.063/0.229/0.263 for absolute proposal contribution, 0.457/0.063/0.243/0.237 for Taylor, and 0.276/0.063/0.254/0.406 for gradient magnitude. Freeze and update magnitude are more tied-matrix-heavy (0.772 and 0.990), while parameter magnitude selects it exclusively. These post-outcome summaries describe mask geometry only; they do not establish score equivalence or a new inferential comparison.

E.5 Held-out reuse of one fixed source mask

A target-only prospective check fixes the exact seed-2031 absolute proposal contribution (source-P), freeze, and composition-random memberships before any target computation. It then evaluates those memberships without modification on fixed target seeds 2035–2037. The source result was already observed when this target protocol was written, so the design is not an end-to-end source-to-target preregistration. Each target also receives a target-P mask ranked from its own matched fp32 reference over updates 1–600; that privileged same-target comparator is neither a deployable selector nor an oracle.

All targets use GPT-2-124M/OpenWebText and a 1000-step endpoint under the fixed 3000-step cosine schedule. Measured stored-weight arms use analytic E5M2 with fp32 compute and optimizer

state. The exact 16,192-tile budget selects 12,435,456 eligible 2-D coordinates, corresponding to 49,741,824 fp32-master payload bytes, exactly 10% of the eligible 2-D blanket-master payload; this accounting excludes metadata, optimizer state, temporaries, fragmentation, runtime storage, and throughput.

Table 17: Step-1000 validation cross-entropy (nats) in the held-out fixed-mask evaluation. "All" denotes blanket fp32-master accumulation; Src-P and Tgt-P denote the fixed seed-2031 and target-specific absolute proposal contribution masks.

Target	fp32	RNE	All	Src-P	Tgt-P	Src-F	Comp-R
2035	5.691	6.776	5.699	6.028	6.010	6.733	6.602
2036	5.710	6.780	5.704	6.046	6.049	6.741	6.614
2037	5.679	6.777	5.712	6.032	6.041	6.732	6.618

Table 18: Gap recoveries and locked margins for the same exact-10% eligible-2-D fp32-master payload. $M = R(\text{Src-}P) - R(\text{Tgt-}P)$, $D_F = R(\text{Src-}P) - R(\text{Src-F})$, and $D_C = R(\text{Src-}P) - R(\text{Comp-R})$. The three target seeds are the scientific units; the fixed masks and selected tiles are not replicates.

Target	All	Src-P	Tgt-P	Src-F	Comp-R	M	D_F	D_C
2035	0.993	0.690	0.706	0.040	0.161	-0.017	0.650	0.529
2036	1.006	0.686	0.683	0.036	0.155	0.003	0.650	0.531
2037	0.970	0.678	0.671	0.041	0.145	0.008	0.637	0.533
Mean	0.990	0.685	0.687	0.039	0.154	-0.002	0.646	0.531

Independent final validation recomputes every loss, recovery, and margin and passes all validity checks for all three targets: the RNE–fp32 anchor gaps are 1.070–1.098 nats and blanket-master recoveries are 0.970–1.006. Source-P recovery is 0.690/0.686/0.678, and the ordinary three-target mean is 0.685. The source-minus-target-P margins are $-0.0165/0.0032/0.0075$ (mean -0.00195). The source-minus-freeze margins are 0.650/0.650/0.637 (mean 0.646), while source-minus-composition margins are 0.529/0.531/0.533 (mean 0.531); the latter is the specificity margin S in every target. The locked per-target floors are $R(\text{Src-}P) \geq 0.50$, $M \geq -0.10$, and $S \geq 0.15$; their mean gates are 0.60, -0.05, and 0.25. All gates pass, so the locked intersection decision is positive. Composition-random is one fixed source mask, not a sample of random allocations. This result therefore supports only unchanged reuse of this one source mask in these three fixed targets under the stated regime.

E.6 Fully prospective modern-decoder source-to-target allocation

The final allocation study separates source selection from all target observations. Stage 0 freezes the 162.2M-parameter untied RMSNorm/RoPE/SwiGLU decoder, OpenWebText revision, analytic E5M2 grid, fp32 compute and optimizer state, 600-step source window, 1000-step target horizon, tile catalog, exact budget, score, tie-breaking, seeds, arm order, evaluation batches, retry rules, and numerical gates. Source seed 58709217 then produces the absolute-proposal-contribution ordering. Its exact mask and artifact hash are committed before target references or outcomes for seeds 1249306074, 1629484925, and 695340460 begin. Earlier v1–v3 protocols remain provenance only: v4 retires all of their role seeds and reuses no prior trace, reference, score, ordering, mask, endpoint, or role-specific schedule as scientific input.

The catalog contains 211,200 contiguous 768-coordinate eligible 2-D tiles. The source mask selects exactly 21,120 tiles and 64,880,640 fp32 payload bytes, 10% of the eligible blanket-master payload. For every target, measured arms include same-seed fp32 and RNE anchors, blanket master, the unchanged source-P mask, a target-specific P mask constructed from that target's own 600-step reference, the source freeze-only mask, and five fixed composition-matched controls. The latter are drawn once to match source-P's exact per-parameter selected-tile quotas and reused unchanged across targets. Target-specific P is a privileged, reference-expensive comparator rather than an or-

acle. Target training seed is the scientific unit; the five fixed controls are neither replicates nor a sample from a declared random-mask population.

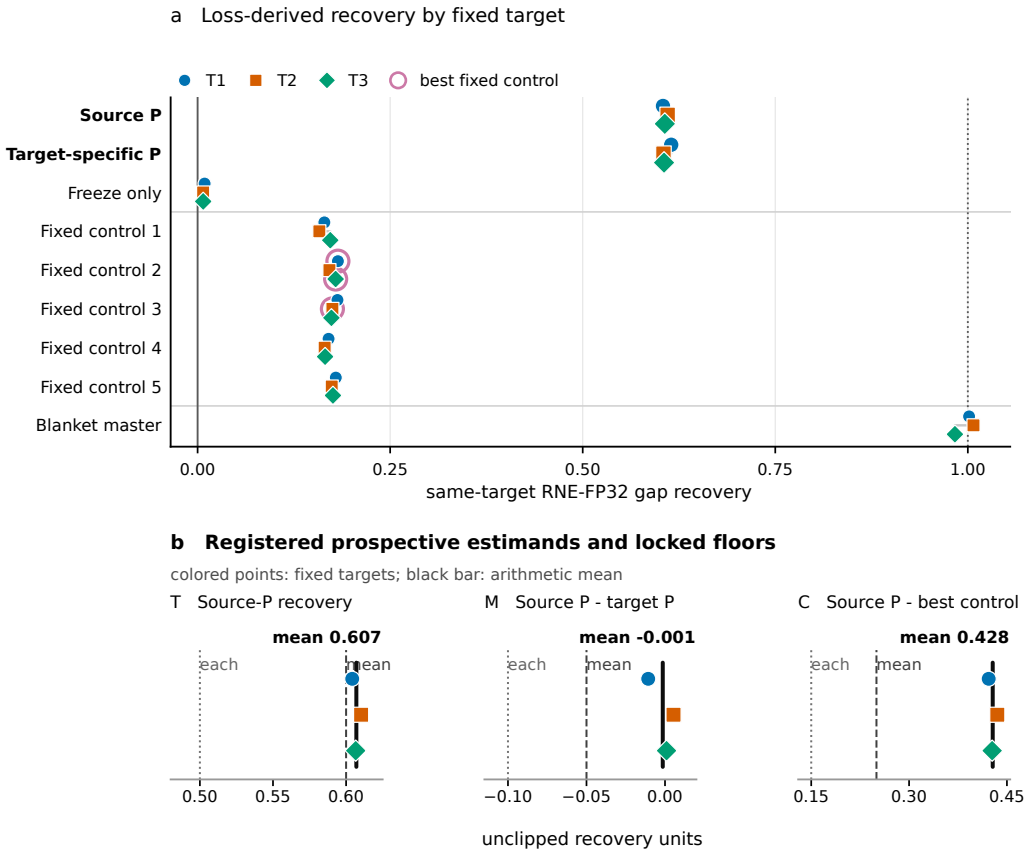

Figure 5: Prospective transfer of a fixed source-P allocation at exactly 10% of the eligible two-dimensional fp32-master payload in the tested 162.2M-parameter decoder. Thin segments span the three fixed target seeds; rings mark the best named fixed control. Target-specific P is a privileged comparator, and the controls are not random-population samples. No inferential interval or hypothesis test is reported.

Table 19: Raw update-1000 held-out validation cross-entropy (nats) for the fully prospective allocation anchors and importance arms. "All" is blanket fp32-master accumulation over eligible 2-D weights.

Target	fp32	RNE	All	Src-P	Tgt-P
1249306074	5.343	6.502	5.341	5.802	5.789
1629484925	5.351	6.488	5.343	5.794	5.800
695340460	5.338	6.524	5.358	5.805	5.806

The anchor gaps are 1.137–1.186 nats and blanket-master recoveries are 0.983–1.007, passing the frozen validity floors of 0.1 and 0.9 in every target. The positive rule requires source-P recovery at least 0.50 per target and 0.60 in the mean, source-minus-best-control at least 0.15 per target and 0.25 in the mean, and source-minus-target-P at least -0.10 per target and -0.05 in the mean. Table 20 passes every gate. The independently sealed validation re-extracts all 33 target endpoints, recomputes every estimand and the fixed intersection decision, and authenticates all three target bundles; it does not claim to replay training.

Table 20: Fully prospective exact-10% modern-decoder allocation at step 1000. Recoveries use same-target fp32 and RNE anchors. "Best named" is the maximum among source-freeze and five pre-fixed composition-matched controls for that target, not a random-population estimate. $\Delta_T = R(\text{Src-}P) - R(\text{Tgt-}P)$ and $\Delta_C = R(\text{Src-}P) - R(\text{Best named})$. Target training seeds are the three scientific units.

Target	Src-P	Tgt-P	Best named	Src-F	Δ_T	Δ_C
1249306074	0.604	0.615	0.182	0.009	-0.011	0.422
1629484925	0.610	0.605	0.175	0.007	0.005	0.435
695340460	0.607	0.606	0.179	0.007	0.001	0.427
Mean	0.607	0.608	0.179	0.008	-0.001	0.428

The packed-path P0 is a preliminary, post hoc systems measurement with 15 retained process records; it is not an execution of the cited prospective 44-process protocol. Versus blanket master, packed source-P saves 584,649,728 peak-allocated bytes and its paired accounted-wall median is 1.402 s lower. The three paired timed-window throughput changes are -0.556%, $+0.352\%$, and -0.259% (median -0.259%); comparing arm medians instead would give the misleading $+0.076\%$. Versus direct RNE, packed adds 65,846,272 peak-allocated bytes, 24.939 s accounted wall at the paired median, and has 7.196% lower arm-median timed throughput. Its persistent packed inventory is 65,049,600 bytes plus an estimated 3,848,396 bytes of Python metadata; other temporary and allocator costs remain incomplete. The full/incremental measured source-score costs are 310.305/153.547 s and cover a combined magnitude and exact-freeze accumulator, not a source-P-only implementation. No break-even rule was registered for this run; the retained null extrapolations are descriptive.

Table 21: Preliminary P0 packed-source systems contrasts. Every entry is packed source-P minus the named reference; negative values favor packed source-P. Peak allocation and accounted wall time are measured quantities, not complete production costs. The blanket comparison uses paired medians; the direct-RNE throughput comparison uses arm medians, as stated in the row label. These 15 post hoc process records support no inferential interval, hardware-generalization claim, or registered break-even decision.

Reference (aggregation)	Δ peak alloc. (B)	Δ accounted wall (s)	Δ throughput
Blanket master (paired)	$-584{,}649{,}728$	-1.402	-0.259%
Direct RNE (arm medians)	$+65{,}846{,}272$	$+24.939$	-7.196%

The separate ECO systems artifact is likewise descriptive: four order-rotated repetitions compare the analytic E5M2/fp32 implementation with direct RNE. Their full-path wall-throughput ratios are 0.899775, 0.902332, 0.899015, and 0.901190 (paired median 0.900483), so two repetitions miss the nominal 0.90 reference. ECO minus packed source-P is $-65{,}846{,}272$ overall peak-allocated bytes and $-407{,}764{,}992$ post-backward temporary bytes; ECO and direct RNE have equal recorded persistent tensor storage. Accounted- wall and full-update wall ratios are 1.106431 and 1.110516. Because no external pre-run artifact sealed the producer and threshold, these descriptive values must not be converted into pass/fail or predeclared decisions. Both systems artifacts carry no training-outcome, native-FP8, distributed, production-cost, or hardware-generalization claim.

E.7 Controlled depth-only allocation migration

The scale-migration contract inherits the base model/training design with one field changed: the v4 decoder's layer count moves from 12 to 17. Width, tokenizer and corpus, analytic scale-1 E5M2 write-back, fp32 compute and optimizer state, schedule, score, budget, source seed, target seeds, and evaluation contexts remain fixed. The 17-layer model has 197,617,920 parameters and 257,280 eligible 768-coordinate tiles. A new 600-step depth-matched source trace selects 25,728 tiles, or 19,759,104 coordinates and 79,036,416 fp32 payload bytes. Within the reported canonical bundle, its regenerated source mask is sealed before that bundle's target references or outcomes; the 12-layer mask is not reused. The scientific inventory is narrower than v4: it retains the anchors, source-P, and target-P but runs only one composition-matched random control and no freeze-only arm.

Before the reported bundle, three technical launches were terminated while systems and validator faults were resolved; none enters the reported 19-run decision. One retains only a completed source arm, while two lack complete remote scientific records. The retained archive therefore cannot certify that no target process began in every technical launch. The canonical bundle regenerated its source trace and mask, and no pre-canonical artifact contributes to its estimands.

Table 22: Controlled depth-17 allocation at update 1000. D is the RNE–fp32 loss gap in nats, R_{All} is blanket-master recovery, $T = R(\text{Src-}P)$, $M = R(\text{Src-}P) - R(\text{Tgt-}P)$, and $C_1 = R(\text{Src-}P) - R(\texttt{random_01})$. The one fixed composition-matched `random_01` is not the best-of-six control in Table 20.

Target	D	R_{All}	T	M	C_1
1249306074	1.136	0.998	0.659	0.009	0.496
1629484925	1.145	0.988	0.661	0.009	0.491
695340460	1.128	1.004	0.672	0.008	0.511
Mean	1.136	0.997	0.664	0.009	0.499

The bundle completes one 600-update source arm and six 1000-update arms for each of the three fixed targets: 19 runs and 18,600 updates in total. Every target passes $D \geq 0.10$, $R_{\text{All}} \geq 0.90$, $T \geq 0.50$, $M \geq -0.10$, and $C_1 \geq 0.15$; the corresponding mean floors are 0.60, -0.05, and 0.25 for T, M, and C_1. A separate invocation of the frozen validator authenticates the source-mask barrier and all 19 terminal records and returns the same `GO` decision. Because this study reuses the fixed seeds and regenerates its mask, it supports a depth-controlled repetition of the procedure, not cross-depth mask transfer or a trend with scale.

E.8 Prospective GPT-2 budget and systems comparison

The final allocation test returns to GPT-2-124M with a fresh source seed and three fresh target seeds. The score and budgets are informed by the earlier studies, but a pre-run manifest freezes all new seeds, schedules, data and code hashes, the 40-run order, and the decision rule. One 600-update source trace seals nested top-P masks at exact 5/10/20% tile payload before any target computation. Each target then runs one 1000-update fp32 reference and 12 matched arms: direct RNE, blanket master, ECO, three source-P budgets, three privileged target-P budgets, and one fixed composition-matched control per budget. The complete inventory contains 39,600 updates. This is prospective for the new source-to-target outcomes, not an outcome-blind discovery of the score or budgets.

Table 23: Fresh GPT-2 source-to-target recovery at update 1000. Entries are means over three fixed targets; brackets give the minimum source-P recovery. Δ_{ST} is source-P minus privileged target-P. The quality gate also requires pre-fixed per-target, mean, and control-separation floors.

Payload	Source-P	Target-P	Control	Δ_{ST}	Quality gate
5%	0.492 [0.484]	0.491	0.050	$+0.001$	Fail
10%	0.698 [0.680]	0.702	0.169	-0.004	Pass
20%	0.800 [0.788]	0.797	0.278	$+0.003$	Pass

All three anchor gaps are 1.090–1.108 nats and blanket recovery is 0.977–0.991, so the validity gates pass. ECO mean recovery is 0.934 and blanket recovery is 0.984. Mean accounted wall time is 304.60 s for direct RNE, 342.95 s for ECO, 335.62 s for blanket master, and 330.94/331.82/333.73 s for the three source-P arms before source-cost amortization. The source trace costs 210.30 s and peaks at 25,057,738,240 allocated bytes; its peak is not amortized. The registered tolerance-dominance points combine mean recovery, maximum target-or-source peak allocation, maximum persistent tensor storage, and mean wall time, adding source cost at reuse counts $1, 3, 10, 30, 100$. At every count its surviving labels are exactly direct RNE, ECO, and blanket master. Under strict zero-tolerance dominance the same labels survive at reuse 1, 3, and 10, while all three source-P budgets additionally survive at reuse 30 and 100. Even at reuse 100, source-P wall time is 333.04/333.92/335.84 s after amortization. Thus 10% and 20% replicate ranking transfer as a positive mechanism result, but no tested budget passes the separate registered deployment tolerance rule; the fixed reuse-3 decision is `retain_allocator_headline=false`. The measured coordinates are instrumented PyTorch GPU allocation, tensor storage, and wall time; the records

mark selector, temporary-buffer, and runtime-overhead accounting as incomplete. The manifest requires one process per 40-item inventory entry and records an RTX 5090 device class with PyTorch 2.7.1+cu128 and CUDA 12.8. Individual records retain neither PIDs nor GPU UUIDs, so distinct operating-system processes and one physical GPU are not independently authenticated.

F Reproducibility and statistical units

Artifact ledger. The retained evidence contains the machine-readable manifests, raw run records, independent validation outputs, and regeneration scripts for every reported result. Historical release archives predate the Pythia-410M addition; its report summary is included in the current checkout, while full replay requires the retained local campaign. The archive verifier checks the integrity receipts and the validators recompute numerical decisions from the raw records; internal receipt values are omitted from this anonymous review archive. This keeps the paper focused on the estimands, statistical units, and claim boundaries while preserving full reproducibility in the accompanying release.

All decoder loss and recovery summaries are computed per training seed. The archived 3k policy contrast has three training seeds; the matched 10k intervention, historical RNE-only 10k replication, and historical whole-tensor allocator have two. The fixed-design modern-decoder transfer has two separately scheduled training seeds and one SR realization per seed. Its loss-anchor standard errors use 128 paired evaluation contexts as evaluation-noise units; those contexts are not training replicates. The NeoX bundle has three target training seeds paired across the two fixed formats; format cells, five policy arms, checkpoints, and the one SR stream per cell are not additional replicates. The native coupled intervention likewise has three fixed target seeds; its four paired treatments per target, checkpoints, evaluation contexts, and stochastic streams are not additional scientific units. The registered combined- primary tile screen has one deployment seed; the separately locked leverage- primary follow-up has three. Historical and locked-interpolation cells reuse a reference across five mantissas and are treated as paired coverage tests. Historical masks are nested within training seeds. Follow-up masks are fixed by budget and reused across the three seeds; neither set is a training replicate or a sample from a specified mask population. The held-out fixed-mask evaluation has three fixed target training seeds, 2035–2037, as its scientific units. Its one seed-2031 source mask and its fixed source-freeze and composition-random controls are reused across all targets; masks, tiles, arms, checkpoints, and evaluation contexts are not replicates. The fully prospective modern allocation has three fixed target training seeds as its scientific units. Its one source seed constructs a single reused mask; target-specific masks and five fixed controls per target are comparators, not additional scientific units. No confidence interval treats these dependent cells as independent replicates. The depth-17 migration uses the same three fixed target seeds as scientific units but constructs a new depth-matched source mask. Reusing seed roles across depths creates a controlled comparison, not six independent seed draws; its one fixed `random_01` mask and target-P masks are comparators rather than additional units. The post hoc ECO comparison adds one arm to each of the same three v4 targets and reuses their sealed anchors; it therefore contributes no new independent training seeds, and no interval or population claim is made from those paired fixed-target contrasts. The source-amortized E4M3 audit contributes three fresh fixed target seeds. Its single reused source, three privileged matched references, temporal checkpoints, and archived template trajectories are predictors or comparators, not additional scientific units. The GPT-2 Pareto bundle contributes three fresh fixed target seeds as its scientific units. Its one fresh source ranking, three nested budgets, three target-specific rankings, and fixed composition controls are dependent comparators, not additional units; the five reuse counts are deterministic cost scenarios rather than experiments. Stored files retain per-step trajectories, final metrics, achieved payload, selected-list hashes, the complete 20-arm pilot and 27+21-arm follow-up inventories, divergence flags, and both criterion-window divergence conventions. The fixed-mask release additionally retains the complete 18-arm target inventory and its start journals.

F.1 Scope and lineage details

The 10,000-step GPT-2 intervention and 3,000-step RMSNorm-decoder policy transfer each use two fixed seeds, and the latter changes several decoder choices; neither supports seed-population or componentwise claims. The NeoX bundle uses three target seeds paired across two fixed formats, not six independent units, and its missing full-shape readiness receipt limits its provenance. The archived 3k and matched 10k GPT-2 SR arms each retain only one process-dependent stochastic realization per

seed, so their sensitivity to rounding randomness is unresolved. The analytic decoder experiments retain fp32 compute and optimizer state. Study 1 uses persistent E4M3 matrix state and native hybrid-FP8 dense GEMMs, but retains fp32 non-GEMM arithmetic and optimizer state. Its sixth cell is a manual recovery rather than an uninterrupted canonical-campaign record; the original producer flag remains false and canonical process identity is absent, so the registered classification remains inconclusive. Independent validation of data, implementation, artifacts, and checkpoint replay supports only the post-outcome retrospective protocol-deviation analysis. Its descriptive negative is not selected by the recovery: the disjoint cosine-restart cell cannot alter the decisive constant-mid macro and centered gates. Missing training start-ticks/nonce evidence and independent update-1000 replay further limit its provenance; that analysis uses the independently replayed update-0/update-3000 endpoints. Study 2 tests software stochastic rounding, not hardware stochastic rounding, and its three fixed targets support no population inference; neither study quantizes optimizer state. The exact event is local, whereas the aggregate proxy follows a diverging counterfactual. Pythia's selective bounds concern sampled coordinate trajectories in the fixed target; they do not give pathwise prediction or independent-source generalization.

These fixed-design checks do not upgrade the analytic Experiment 1–3 results to native-precision population generalization, hardware stochastic rounding, formal causal mediation, or production efficiency. Study 3 uses simulated INT3 values in fp32 containers and covers one fixed model, corpus, tokenizer, and seed. Its direct-state arithmetic follows an archived DQT source snapshot, but the locally frozen split, horizon, evaluation, cache policy, and activation-checkpointing wrapper are not a literal reproduction of the upstream training script or a reported upstream-paper endpoint. Signed-INT3 values reside in fp32 parameter containers, compute is fp16, and optimizer moments remain fp32. This demonstrates neither native INT3 storage nor native INT3 arithmetic, and provides no evidence for INT3 memory, throughput, or production efficiency. The manual continuation remains part of the same seed and supports no seed-, corpus-, or architecture-population inference.

The source-amortized audit adds three fixed E4M3 target seeds, not a sampled seed population. Its historical template differs in batching and attention runtime, and the post-outcome variance decomposition is descriptive. The template win rules out neither an exact-runtime template nor every useful source trace; it shows that the registered raw and unit-gain source predictors add no value for this tested aggregate observable and that registered absolute accuracy does not establish source-specific information transfer.

Pythia's 15 targets share three fresh sources and three historical pairs per age; they are not 15 source replications. The negative concerns the frozen selector and unit-gain common-temporal correction, not every possible reference method. Time centering removes each trajectory's mean rather than the shared temporal shape. The post-outcome analysis repair limits an unmodified-executable claim, while preserving the pre-holdout statistical rules and stored metrics. The experiment establishes a counterexample to the sufficiency of those original accuracy criteria, not a calibration theorem, source-universal negative, or population estimate of failure frequency.

The allocator studies cover two fixed decoder configurations, one corpus, analytic E5M2, fixed seeds, and one optimizer schedule. The fresh GPT-2 bundle is prospective for its new source and target outcomes, but the score and 5/10/20% budgets were selected from prior evidence; its nested budgets are not independent experiments. Depth-17 regenerates rather than transfers the mask, so it is not a scaling law. Target-P is privileged, fixed controls are not population samples, and three targets support no seed-population inference. The score is neither novel nor optimal, native selective allocation remains untested, and no tested source-P budget survives the pre-fixed tolerance rule. Strict zero-tolerance dominance does retain source-P at reuse 30 and 100, so this is not a rule-free Pareto exclusion. Appendix E records historical initialization, systems, and comparator boundaries.

F.2 Broader precision-system context

GaLore reduces optimizer-state memory by projecting gradients into a low-rank subspace, then reconstructing full-parameter updates (Zhao et al., 2024). Whether those updates survive write-back still depends on the destination representation and accumulation policy. Wang et al.'s differentiable gradient estimator for FP4 addresses backpropagation through weight quantization in a mixed-precision training system (Wang et al., 2025). This changes gradient estimation rather than guaranteeing preser-

vation of sub-grid stored-weight updates. Neither method alone removes the need to specify the write-back channel; neither is evaluated as a forecast baseline here.

Forward-pass attention rounding and optimizer-state staleness are separate arithmetic channels (Qiu & Yao, 2026; Topollai & Choromanska, 2026). Quartet/Quartet II and native MXFP4 target hardware FP4 linear layers and gradient estimation (Castro et al., 2025; Panferov et al., 2026; Cim et al., 2026). FP4 All the Way and Full-Stack FP4 extend quantization through backward/update, optimizer, or attention paths, while Stable FP4 isolates transposition-consistent block scaling (Chmiel et al., 2025; Ding et al., 2026; Rahimifar et al., 2026). dMX learns deployment-format assignments, whereas FlashOptim compresses master weights and optimizer state (Franco et al., 2026; Gonzalez Ortiz et al., 2026). We therefore use blanket-master and direct-write-back baselines rather than one explanation for every low-precision failure.

Hardware equivalence. The archived study JSONs include seeds and protocols but not a complete GPU, driver, CUDA/PyTorch, TF32, and data-hash record. The completed guarded 10k launcher records those fields and verifies source and tokenized-data hashes before running on an NVIDIA A100-PCIE-40GB. The analytic grid is device-independent, but fp32 trajectories are not assumed bitwise identical. A non-A100 run remains a separately labelled hardware robustness result even when its matched fp32 trajectory is close; it is not pooled into or substituted for the A100 headline. The matched 10k result retains the analytic grid rather than native fp8 kernels. The locked criterion-window result instead uses single-thread CPU fp64 with exact phase-to-phase tensor and environment fingerprints. No GPU kernel enters that numerical path, so the installed GPU model is irrelevant to the canonical prospective result. GPU replicas are robustness checks and are not assumed bitwise equivalent. The allocator reference, pilot, and fixed-design follow-up use the canonical A100 contract with a constant device UUID, software and determinism settings, and bound source, data, evaluation, and schedule hashes. The later fixed-mask target protocol amended the operational contract to permit any A100 with compute capability 8.0 while retaining pinned software, deterministic settings, data, evaluation, schedule, and mask hashes; all target arms ran on one recorded A100-PCIE-40GB. Independent validation authenticates the complete attempt inventories and recomputes decisions without rerunning training. Exact fp32 payload accounting in the allocation outcome bundle does not itself measure full device memory, metadata overhead, runtime, or throughput. The separate systems-only P0 is a preliminary post hoc 15-process measurement on one recorded RTX 5090 stack, not execution of the later 44-process protocol. It is not pooled with training outcomes or interpreted as native-low-precision performance, and it has no registered break-even rule. The ECO target systems campaign uses four order-rotated repetitions of direct-RNE, ECO, packed source-P, and blanket-master in 16 distinct fresh processes on one RTX 5090. Its descriptive memory contrasts compare ECO with packed source-P, and persistent storage is compared with direct RNE. Two of four throughput ratios miss 0.90. No external pre-run artifact seals the producer and threshold, so the campaign has no pass/fail or predeclared decision; it is systems-only and contributes no training-outcome units. The modern-decoder run likewise used one admitted A100, exact source and data hashes, deterministic settings, and a pre-measurement readiness commit. After a post-completion packaging failure, hash-bound recovery preserved the failure record and commit lineage, recovered byte-identical scientific JSONs, and independently reproduced the registered analysis. No training arm or scientific result was affected. The fully prospective allocation bundle ran its source and every target arm on one NVIDIA GeForce RTX 5090 under driver 570.190, PyTorch 2.7.1+cu128, CUDA 12.8, and one frozen container/software stack. Its fp32 anchors and all compared arms are matched within that bundle; they are not pooled with earlier A100 experiments or treated as bitwise-equivalent cross-GPU replicates. After an instance restart, the ECO arms ran on another RTX 5090 UUID; all sealed compute-relevant runtime fields matched. The compatibility rule excludes GPU UUID and container hostname, although the observed hostname is unchanged. This is not bitwise identity on one physical device. The controlled depth-17 migration uses the same recorded RTX 5090 GPU/software stack under a separate P1 runtime seal and fresh-process arms. Its matched endpoints are interpreted within that bundle, not as an additional hardware replicate or a native-fp8 measurement. The GPT-2 Pareto manifest requires 40 separate processes and records an RTX 5090 device class under PyTorch 2.7.1+cu128 and CUDA 12.8, but the run records omit PIDs and GPU UUIDs. Process distinctness and one physical device are therefore requirements, not independently authenticated provenance. Its quality and systems coordinates are paired within the bundle; they are neither a cross-hardware replication nor a native-fp8 measurement.