CCRT: A SINGLE-STATE OPTIMIZER WITH RETAINED GRADIENT HISTORY AND BOUNDED NONLINEAR UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern optimizers often retain multiple full-shape statistics, making optimizer state a substantial cost at scale. Current Carry Root Transport (CCRT) instead persists one BF16 gradient-history tensor per parameter, reinjects the fresh gradient, and applies a bounded root-rational readout without a second moment. Across three observed 134.8M training trajectories—two fresh paired runs and one earlier independent run—CCRT shows the same terminal validation-CE and validation-AUC ordering against calibrated AdamW and PowerStep. In a complementary single-seed 3B-scale reproduction, CCRT reaches validation perplexity 127.15, versus 246.22 for AdamW and 378.64 for PowerStep, while using 75% and 50% less persistent optimizer state, respectively. CCRT also completes a 1K-step single-GPU run at 6.98B parameters on an 80GB A800; validation CE decreases across all five recorded checkpoints, with a 13.00 GiB BF16 optimizer state. We also characterize the ideal full-precision update mathematically. Exact scalar quadratic dynamics exhibit a structured nonzero fixed-step two-cycle, while a deterministic smooth-nonconvex analysis proves a natural-stationarity residual of order $dL^3\eta^3$ and conventional $O(\epsilon^{-2})$ first-order stationarity complexity. The scalar cycle matches the step-size exponent of the general residual bound. Together, the empirical and theoretical results show that a low-state nonlinear optimizer can combine substantial state savings with effective large-model optimization and an explicit connection between fixed-step dynamics and general nonconvex stationarity complexity.

1 INTRODUCTION

Adam-style methods retain multiple gradient statistics (Kingma & Ba, 2014; Loshchilov & Hutter, 2019). Their effectiveness comes with persistent-state and execution costs, leaving open how much optimizer machinery sustained language-model optimization actually requires. State-compression methods retain adaptive-moment machinery in cheaper form. CCRT instead asks whether a single temporal state can support useful long-horizon behavior without approximating an Adam-style second moment.

CCRT keeps one gradient-history carry, injects the fresh gradient into the readout, and applies a bounded magnitude-sensitive transform. The carry transitions from a running-mean schedule to a fixed-timescale schedule. In implementation it is persisted in BF16 and reconstructed in FP32; no second-moment tensor is maintained.

A one-state optimizer still has three jobs. It must retain useful information from earlier gradients, respond when the current gradient differs from that history, and control coordinate update magnitude without second-moment normalization. CCRT assigns these roles to the scheduled carry, a direct path from the current gradient to the update, and a root-rational map that remains magnitude-sensitive for finite inputs while saturating at large magnitude. BF16 applies only to the state transported between updates; the current transient carry is formed and read in FP32.

We evaluate CCRT across a scale ladder spanning replicated 134.8M comparisons, a 3B optimizer comparison, and a 6.98B single-GPU feasibility run. The 134.8M Transformer is trained on FineWeb-Edu (Penedo et al., 2024). Two fresh paired 10K runs form the primary replication: CCRT has lower

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.

terminal CE and validation AUC than AdamW and PowerStep in both. An earlier independent 10K run shows the same terminal and AUC ordering, giving consistent ordering across all three observed 134.8M trajectories; we keep it separate from the primary replication. A separate single-seed 3B-scale reproduction then tests whether this ordering and the state advantage persist at a substantially larger model size. CCRT finishes below both comparators in validation CE while using 5.59 GiB of optimizer state, compared with 22.36 GiB for AdamW and 11.18 GiB for PowerStep. At 6.98B, a CCRT-only run completes 1K updates on one A800 80GB with decreasing recorded validation CE and a 13.00 GiB BF16 carry.

CCRT is useful not only as a low-state update: its nonlinear dynamics remain tractable enough to connect exact fixed-step behavior with general optimization complexity. For scalar quadratics, fixed steps generically approach a nonzero two-cycle whose gradient scale is $\Theta(\eta^2)$. For deterministic L-smooth objectives bounded below, we prove a finite-time bound in a natural CCRT stationarity measure, obtain $O(T^{-3/4})$ optimized decay in that measure, and recover $O(\epsilon^{-2})$ complexity for an ℓ_2 ϵ-stationary iterate. The scalar cycle shows that the fixed-step residual has the correct exponent.

Our contributions are:

- **Algorithmic.** A concrete optimizer with one BF16 gradient-history tensor, direct fresh-gradient readout, and no second-moment state.
- **Empirical and systems.** Three observed 134.8M trajectories—two fresh paired runs and one earlier independent run—a separate single-seed 3B optimizer comparison, and a 6.98B single-GPU feasibility run, with state, memory, runtime, and long-horizon measurements kept at distinct evidence levels.
- **Exact dynamics.** A scalar quadratic analysis and a finite-h full-carry classification into the stable set of the origin and the basin of a unique nonzero two-cycle.
- **Nonconvex theory.** Deterministic smooth-nonconvex stationarity complexity with an exponent-sharp $O(dL^3\eta^3)$ fixed-step residual and $O(\epsilon^{-2})$ conventional stationarity complexity.

2 Related Work

Adam and AdamW retain first- and second-moment statistics (Kingma & Ba, 2014; Loshchilov & Hutter, 2019); Adafactor factorizes second-moment state (Shazeer & Stern, 2018), and 8-bit optimizers quantize it (Dettmers et al., 2022). CCRT instead retains one historical tensor. Fresh-gradient and momentum mixing appears in QHM (Ma & Yarats, 2019), while Lion uses one momentum tensor with a sign readout (Chen et al., 2023).

Powerball provides signed-power precedent (Yuan et al., 2016). PowerStep applies a signed-power transform to momentum and is our close contemporary comparator (Lu et al., 2026); related bounded or smooth sign-like methods include SoftSignSGD and SoftSignum (Peng et al., 2025; Feoktistov et al., 2026). CCRT claims neither single-state momentum nor nonlinear readouts as individually new.

The ideal CCRT recurrence has an exact representation in generalized Lion-κ (Chen et al., 2024); Appendix A.4 gives the mapping. This is a structural representation within a broader formalism; the concrete CCRT studied here is defined by its scheduled carry, fresh-gradient reinjection, root-rational readout, and BF16 single-state persistence semantics. Matrix-structured methods instead exploit layer geometry, as in modular duality and Muon-related work (Bernstein & Newhouse, 2025).

Algorithm 1 CCRT update for one parameter tensor

Require: $\theta, G, \bar{C}, t, H, \eta, \lambda$
1: $t \leftarrow t + 1;\ \alpha \leftarrow 1/\min(t, H)$
2: $C^* \leftarrow (1-\alpha)\,\mathrm{float32}(\bar{C}) + \alpha G$
3: $\bar{C} \leftarrow Q_{\mathrm{BF16}}(C^*)$
4: $D \leftarrow G + C^*;\ U \leftarrow \mathrm{sign}(D)\sqrt{|D|}/(1+\sqrt{|D|})$
5: $\theta \leftarrow \theta(1-\eta\lambda) - \eta U$

3 Current Carry Root Transport

For a parameter receiving its t-th gradient G_t,

$$\alpha_t = 1/\min(t, H), \tag{1}$$

$$C_t^* = (1-\alpha_t)\,\mathrm{float32}(\bar{C}_{t-1}) + \alpha_t G_t, \tag{2}$$

$$\bar{C}_t = Q_{\mathrm{BF16}}(C_t^*), \qquad D_t = G_t + C_t^*, \tag{3}$$

$$U_t = \mathrm{sign}(D_t)\frac{\sqrt{|D_t|}}{1+\sqrt{|D_t|}}, \tag{4}$$

$$\theta_{t+1} = \theta_t(1-\eta_t\lambda) - \eta_t U_t. \tag{5}$$

We use $H = 32$ and $\lambda = 0$ in Eq. 5; t counts parameter updates for which a gradient is present.

Scheduled carry. The interpolation coefficient depends on the number of observed updates rather than starting with a fixed momentum coefficient. Early in training, the schedule incorporates all available history under the ideal recurrence; after the information horizon, it becomes a fixed-timescale summary without adding another tensor.

Fresh-gradient reinjection. The retained carry does not replace the current gradient. Forming $D_t = G_t + C_t^*$ gives the fresh gradient a direct path to the transport even when the carry reflects substantial accumulated history. This differs from reading out the carry alone, without presupposing its causal contribution.

Root-rational transport. The nonlinear readout controls large coordinate signals without discarding all finite-magnitude information. Unlike a pure sign update, its output varies continuously with input magnitude, but unlike an unbounded signed-power map, it approaches unit magnitude for large inputs, providing magnitude control without a second-moment normalizer.

The persisted $\bar{C}_t$ is BF16; C_t^* is the fresh FP32 transient carry used by the current update. Thus single-state means one full-shape optimizer tensor per parameter, with integer step metadata also present. The stored carry is not the exact arithmetic mean of the ideal recurrence. Appendix A gives exact implementation, clipping, and scheduling semantics.

This precision split places reduced precision on historical information transported between updates, while constructing the current carry and readout in FP32. Algorithm 1 summarizes this compact persistent representation, which is not by itself a robustness guarantee for BF16 or lower-bit formats.

4 Properties and Optimization Theory

4.1 Root-Rational Geometry

For finite real z, let

$$\phi(z) = \mathrm{sign}(z)\frac{\sqrt{|z|}}{1+\sqrt{|z|}}.$$

Then $|\phi(z)| < 1$, $\phi(z) \sim \mathrm{sign}(z)\sqrt{|z|}$ near zero, and $\phi(z) \to \mathrm{sign}(z)$ at large magnitude. Thus the readout preserves finite-magnitude information while bounding each zero-decay coordinate update by η_t. It is non-Lipschitz at zero; the bound alone implies neither convergence nor activation or function-space stability. Appendix A.3 gives the elementary derivations.

4.2 Exact Fixed-Step Dynamics

The square-root local geometry prevents generic exact convergence under a fixed positive step. For the full-precision scalar quadratic $f(x) = \lambda x^2/2$, where here $\lambda > 0$ denotes curvature, trajectories outside the stable set of the origin approach a nonzero symmetric two-cycle. More strongly, for the fixed-beta full-carry map with $\beta = 31/32$, normalized step $h = \eta\lambda$, and $0 < h \le 2^{-17}$, every orbit belongs either to the full stable set of the origin or to the basin of the unique nonzero symmetric two-cycle; there are no other nonzero periodic or omega-limit behaviors in this regime. This is a scalar, deterministic, full-precision result proved over the complete finite-h interval, not a parameter scan (Appendix N).

4.3 Natural Stationarity and Nonconvex Complexity

For $g \in \mathbb{R}^d$, define the stationarity measure induced by the no-lag alignment term $g_i\phi(2g_i)$:

$$\mathcal{S}(g) = \sum_{i=1}^{d} \frac{\sqrt{2}|g_i|^{3/2}}{1 + \sqrt{2|g_i|}}. \tag{6}$$

It interpolates between an $\ell_{3/2}^{3/2}$-type quantity near the origin and an ℓ_1-type quantity for large gradients. Precisely,

$$(2 - \sqrt{2}) \sum_i \min\{|g_i|^{3/2}, |g_i|\} \le \mathcal{S}(g) \le \sqrt{2} \sum_i \min\{|g_i|^{3/2}, |g_i|\}. \tag{7}$$

Assume only that f is L-smooth and bounded below. For ideal full-precision CCRT with the actual $H = 32$ prefix and a constant-step fixed-beta tail, the proved finite-time bound has, after explicit finite prefix and carry transients,

$$\min_{t<T} \mathcal{S}(\nabla f(x_t)) \le \frac{1}{T} \sum_{t<T} \mathcal{S}(\nabla f(x_t)) = O\left(\frac{1}{\eta T}\right) + O(dL^3\eta^3). \tag{8}$$

No convexity, PL inequality, bounded-gradient, bounded-iterate, or compact-sublevel assumption is used. Optimizing η gives $\mathcal{S} = O(T^{-3/4})$. The proved norm conversion then yields an iterate with $\|\nabla f(x_t)\|_2 \le \epsilon$ in $T = O(\epsilon^{-2})$ deterministic first-order complexity. At a fixed step, Eq. 8 corresponds to gradient scale $O(d^{2/3}L^2\eta^2)$, not exact stationarity.

The carry is not bounded as a crude $O(\eta)$ perturbation. Writing $Q = \phi^{-1}$ and $p_t = \phi(D_t)$, the fixed-beta identity

$$Q(p_t) - \beta Q(p_{t-1}) = 2(1 - \beta)g_t + \beta(g_t - g_{t-1})$$

combines with the convex dual geometry of ϕ to telescope the carry interaction; the remaining mismatch is a stable geometric convolution. This preserves the η^3 residual exponent. On the exact scalar cycle, $|\nabla f| = \Theta(\eta^2\lambda^2)$ and $\mathcal{S}(\nabla f) = \Theta(\eta^3\lambda^3)$, matching the general upper-bound exponents. Appendix O states the tail indexing, proof mechanism, decaying-step corollary, and stochastic boundary.

5 Experimental Setup

We train a 134,764,032-parameter decoder-only Transformer on FineWeb-Edu (Penedo et al., 2024). Each update processes 1,024 tokens at sequence length 256. Model computation and gradients are FP32; the CCRT carry is BF16. All methods use clipping at norm 1.0, zero weight decay, 10 warmup steps, and cosine decay through 10K. The fixed validation cache contains 131,072 tokens. Architecture, data, tokenizer, batching, and validation details are in Appendix B.

The primary evaluation contains two fresh paired seeds with shared initialization, data order, and fixed validation data. CCRT uses learning rate 0.01 and $H = 32$; fused AdamW uses 6×10^{-4}, $(0.9, 0.999)$, and $\epsilon = 10^{-8}$; PowerStep uses 6×10^{-4}, momentum 0.9, and power 0.1. Separate two-seed 200-step runs calibrated AdamW and PowerStep; CCRT was not recalibrated (Appendix C). Metrics and systems definitions are in Appendix D; timings are same-hardware and implementation-specific.

Pairing controls two important sources of comparison noise: within a seed, the optimizers begin from identical parameters and consume the same examples in the same order. Calibration seeds are distinct from the official comparison seeds, and no optimizer is retuned after the fresh paired results are observed. The main comparison horizon is 10K updates; the separate 100K experiment evaluates CCRT alone.

The complementary scale comparison trains a 3,000,674,304-parameter model for 10,000 updates (81,920,000 tokens per run) on one NVIDIA A800 80GB. AdamW, CCRT, and PowerStep use frozen learning rates 3×10^{-4}, 5×10^{-3}, and 3×10^{-4}, respectively. This is one official seed per optimizer under the same seed definition, source configuration, validation schedule, and budget. The recorded serialized initialization hashes differ, and the original files are unavailable for tensor-level comparison; we therefore describe it as a single-seed scale reproduction rather than a strictly tensor-identical paired trial (Appendix L).

6 Main Results

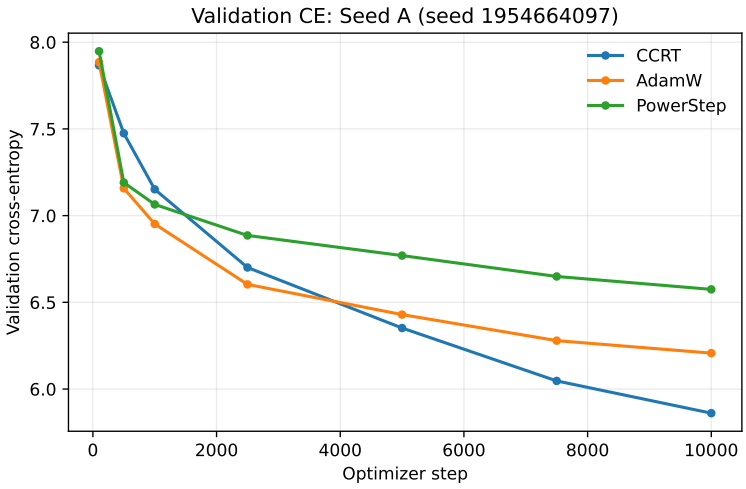
6.1 Paired 10K Optimization

As summarized in Table 1, both fresh seeds place CCRT below AdamW and PowerStep in terminal CE and validation AUC. Mean terminal CE is 5.8629, compared with 6.1993 for AdamW and 6.6000 for PowerStep, a mean difference of 0.3365 CE relative to AdamW. The mid- and late-window slopes are also more negative for CCRT in both runs. These two seeds provide directional replication, not a significance claim; Appendix E reports the per-seed values, and Appendix F separates the auxiliary historical run. Including that earlier independent run, all three observed 134.8M seeds have the same terminal-CE and validation-AUC ordering against both comparators, although we do not treat them as one preregistered three-seed sample.

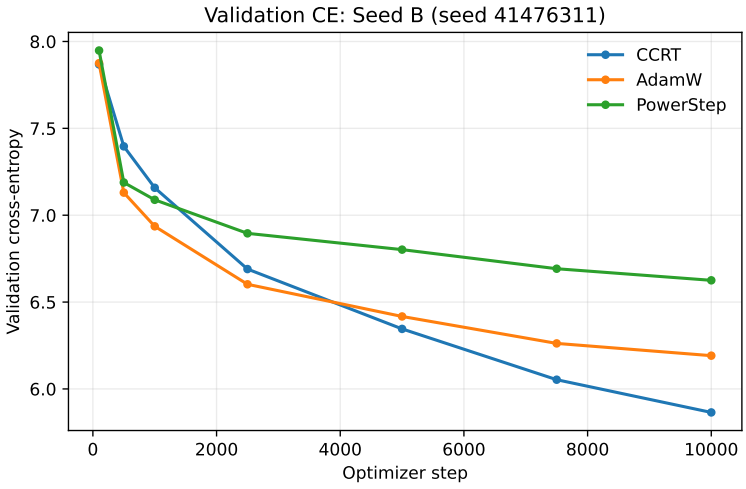
Figure 1 shows that CCRT is not uniformly best early in training. The curves separate more clearly in the middle and later windows, when CCRT descends faster and finishes below both comparators. Its lower AUC on each fresh seed shows that the difference extends beyond the terminal checkpoint to the measured trajectory.

Table 1: Mean quality metrics over the two fresh paired runs.

Optimizer	CE@10K	$AUC_{100:10K}$	$AUC_{1K:10K}$	Mid slope	Late slope
CCRT	**5.8629**	**6.4359**	**6.3278**	**-1.798e-4**	**-1.007e-4**
AdamW	6.1993	6.5094	6.4288	-1.139e-4	-5.031e-5
PowerStep	6.6000	6.8318	6.7799	-5.818e-5	-4.274e-5



(a) Fresh paired Seed A.



(b) Fresh paired Seed B.

Figure 1: Validation CE for the two fresh paired 10K-step runs.

6.2 Temporal Behavior

The optimizer ranking changes over the measured horizon. On both fresh seeds, CCRT crosses AdamW and PowerStep persistently, in line with the steeper mid- and late-window descent. Short-horizon ordering does not fully predict the 10K result.

Across the two fresh paired runs and the earlier independent run, CCRT eventually crosses both comparators persistently in all three observed trajectories. The crossing time varies substantially,

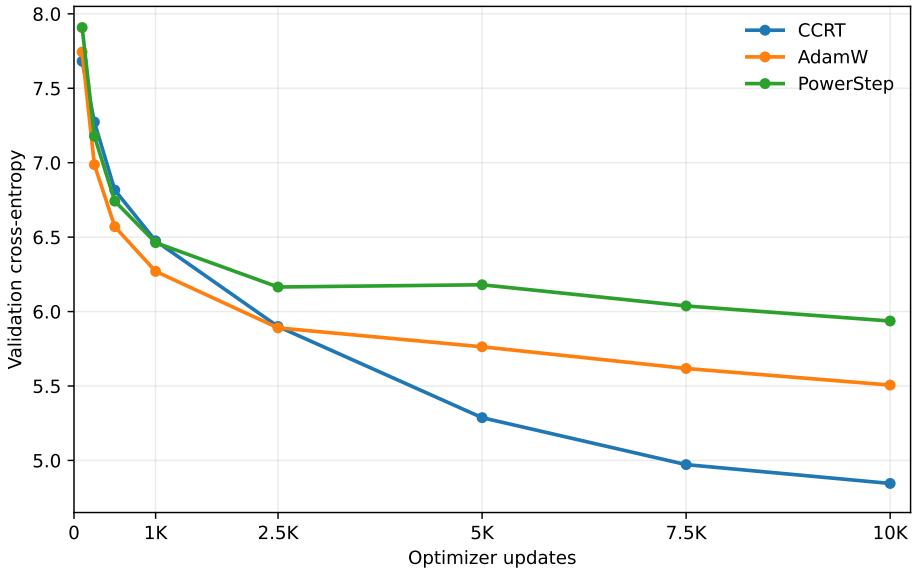


Figure 2: Validation cross-entropy in the 3,000,674,304-parameter, single-seed scale reproduction over 10K updates. Points are the recorded validation checkpoints; connecting segments are visual guides. AdamW is lower from 250 through 2.5K updates. CCRT overtakes between the recorded 2.5K and 5K checkpoints and continues improving thereafter; the exact crossing step is not observed.

Table 2: 3B single-seed scale reproduction. Peak allocation is PyTorch allocated GPU memory; state is persistent optimizer tensor state.

Metric	AdamW	CCRT	PowerStep
Final CE	5.506206	**4.845351**	5.936574
Final PPL	246.2151	**127.1479**	378.6355
Peak alloc. (GiB)	53.05	**36.29**	41.87
State (GiB)	22.36	**5.59**	11.18
Wall-clock (h)	6.738	**6.691**	6.895

particularly against AdamW, so the repeatable observation is the longer-horizon ordering change rather than a characteristic crossover step (Appendix F).

6.3 Single-Seed Scale Reproduction at 3B

Figure 2 and Table 2 report the 3B-scale comparison. CCRT is not uniformly ahead: AdamW has lower validation CE at the recorded checkpoints from 250 through 2.5K, where the values are 5.890414 for AdamW and 5.899434 for CCRT. CCRT overtakes AdamW between the recorded 2.5K and 5K checkpoints and continues to improve through 10K. Its final CE is 4.845351, compared with 5.506206 for AdamW and 5.936574 for PowerStep. Final perplexity is 127.1479, 48.36% below AdamW and 66.42% below PowerStep in this run. This is complementary single-seed scale evidence rather than a replicated comparison. All three runs completed without hard failure.

6.4 Systems Efficiency

Table 3 reports that CCRT uses approximately 75% less persistent optimizer tensor state than AdamW. Its measured optimizer path is $3.46\times$ faster, while complete-step time is 6.9% lower. The gap narrows at the step level because forward and backward computation, data handling, and gradient clipping are largely shared.

Against PowerStep, CCRT stores 50% less persistent optimizer tensor state and has lower measured optimizer and complete-step time in this harness. At 3B, the same state ratios correspond to 5.59 GiB

for CCRT, 22.36 GiB for AdamW, and 11.18 GiB for PowerStep. CCRT also uses 31.61% less peak PyTorch allocated memory than AdamW and 13.35% less than PowerStep, while throughput remains similar (Appendix L). Persistent state is not total GPU memory, and runtime measurements remain implementation- and hardware-specific.

Table 3: Systems measurements over the two fresh 10K-step runs.

Optimizer	Optimizer ms	Complete ms	Tokens/s	State MB
CCRT	**6.658**	**224.959**	**4552.1**	**269.5**
AdamW	23.066	241.597	4238.5	1078.1
PowerStep	33.802	252.278	4059.0	539.1

6.5 Long-Horizon Viability

The separate frozen CCRT trajectory reaches 100K steps and 102.4M tokens without a hard numerical failure. Validation CE falls from 5.3346 at 50K to 5.1542 at 75K and 5.0199 at 100K; the terminal checkpoint is the best recorded value. No matched comparator is present, so the experiment addresses viability only. Full diagnostics are in Appendix H.

Validation continues to improve without progressive growth in the sampled relative-update upper envelope through the final checkpoint. We use that telemetry descriptively, not as a convergence test. The appendix contains the full trajectory, perplexities, health classifications, and sampled-update values.

6.6 Single-GPU 7B Scale Feasibility

To test whether the one-state design remains operational at larger scale, we train a 6,981,064,704-parameter dense model with CCRT for 1,000 updates on one NVIDIA A800 80GB. The run completes without non-finite losses, optimizer fallback, or non-finite carry state. Training CE falls from 11.4165 to 6.4309, while validation CE decreases at each of five recorded checkpoints from 11.4419 to 6.7630. The BF16 carry occupies 13.00 GiB, exactly 2 bytes per parameter, and peak PyTorch allocated memory is 67.01 GiB. All 1,000 steps trigger the shared global gradient-norm clip at 1.0, and the fixed validation set contains 4,096 tokens. Because this is a single-seed CCRT-only run, we use it as scale-feasibility evidence rather than a same-scale optimizer comparison. Full diagnostics are in Appendix M.

7 Component and Sensitivity Analysis

7.1 Component and Interaction Diagnostics

Appendix Table 7 summarizes the frozen single-seed diagnostics. At 2.5K, No Carry trails Full CCRT clearly, supporting a material role for historical information in the run. The pure sign readout fares worse still: bounded direction alone does not match the finite-magnitude root-rational readout at the frozen operating point.

Carry Only behaves differently. It is slightly ahead at 2.5K, falls behind by 5K, and loses more ground through 7.5K and 10K. Full CCRT also has lower 10K AUC and more negative slopes. In this diagnostic, the fresh-gradient path matters mainly later in training (Appendix J.2).

Selected 2.5K contrasts are also non-additive. Retained history has a larger measured benefit under the root-rational readout than under sign, while fresh reinjection has a larger short-horizon benefit under sign. These are single-seed endpoint contrasts, not significance tests or general causal claims (Appendix J.1). A separate 10K replay finds no advantage for the early running-mean schedule: an uncorrected fixed $\beta = 31/32$ carry remains slightly better in terminal CE and AUC as its lead narrows (Appendix J.4).

7.2 Local Hyperparameter Sensitivity

Appendix Table 13 reports a frozen single-seed 5K study of $H \in \{16, 32, 64, 128\}$ without retuning. All runs remain finite. At 5K, $H = 32$ is best, $H = 16$ remains close, $H = 64$ separates moderately, and $H = 128$ has the largest gap; at step 500, $H = 16$ is ahead. This probes a local region, not global optimality or general robustness; full checkpoint and slope results are in Appendix K.1.

With the same seed and $H = 32$, LR 0.005 lowers both AUCs ($6.8080/6.6457$ vs. $6.8263/6.6659$) but finishes 0.0778 CE above LR 0.01; LR 0.0025 ends 0.2118 higher with worse AUC. The new runs use a 5K cosine endpoint versus the reference's 10K scheduler, so the diagnostic is local and scheduler-confounded (Appendix K.2).

8 Discussion and Limitations

At 134.8M, CCRT shows the same terminal-CE and validation-AUC ordering in all three observed trajectories: two fresh paired runs that form the primary replication and one earlier independent run. We keep the historical run separate from the primary replicated sample. Across the evaluated language-model runs, CCRT achieves the reported optimization trajectories without storing a second moment, and the state reduction becomes substantial at 3B scale. The 3B reproduction has one official seed per optimizer, and the archived initialization files cannot retrospectively establish tensor identity. The 100K trajectory is CCRT-only viability evidence. Runtime remains implementation- and hardware-specific, and persistent optimizer state is distinct from total or peak training memory. The 6.98B run establishes single-GPU feasibility under the tested configuration, but it is single-seed and CCRT-only, uses a 4,096-token validation set, and all steps operate under the global clipping threshold. The component diagnostics support retained history and the finite-magnitude readout without constituting a causal decomposition. Fixed $\beta = 31/32$ is comparable to or slightly better in the tested 10K trajectory; the running-mean prefix therefore remains part of the frozen definition without an established independent benefit. The horizon study is local, and the learning-rate diagnostic also changes scheduler horizon.

The nonconvex theorem is deterministic and full precision. The global dynamical classification is confined to scalar quadratics, fixed beta, and $0 < h \leq 2^{-17}$. Standard conditional unbiasedness and bounded variance alone are insufficient for stochastic CCRT because nonlinear transformation need not preserve unbiased alignment.

Future work. The fixed-step analysis motivates theory-guided readouts with a locally linear core, root-like intermediate transport, and large-signal saturation. The one-state formulation also makes lower-precision, structured, and low-rank carries—and combinations of these representations—natural targets for further state reduction. More broadly, the results suggest studying how near-zero readout geometry, memory kernels, and transformed-alignment conditions organize the dynamics of nonlinear stateful optimizers, including in sparse-gradient and mixture-of-experts settings.

9 Conclusion

CCRT retains temporal information in one BF16 tensor and applies a bounded nonlinear readout without storing a second moment. Across three observed 134.8M trajectories, including two fresh paired runs and one earlier independent run, CCRT shows the same terminal-CE and validation-AUC ordering against AdamW and PowerStep. A separate single-seed 3B-scale reproduction preserves the final ordering with substantially less optimizer state; a 6.98B CCRT-only run completes on one 80GB GPU with improving recorded validation CE; and CCRT remains finite through a separate 100K trajectory. Exact scalar dynamics identify the fixed-step two-cycle, while the deterministic smooth-nonconvex analysis gives an exponent-sharp η^3 natural-stationarity floor and $O(\epsilon^{-2})$ ℓ_2-stationarity complexity. Together, these results show that retained temporal information, substantial state savings, and explicit nonlinear dynamical analysis can coexist in a concrete large-model optimizer.

AI USE STATEMENT

Generative AI tools assisted orchestration, code review and debugging, literature search, result cross-checking, experimental-design discussion, mathematical exploration and proof auditing, and manuscript drafting and editing. All experimental results were produced by the authors' code and checked against underlying artifacts. Mathematical statements, references, configurations, and claims were manually reviewed and independently audited where applicable. AI suggestions were not treated as evidence. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A specify the exact update and source identity. Section 5 and Appendices B–H document model, data, calibration, seed, validation, and systems protocols; Appendix L gives the 3B reproduction, and Appendix M documents the CCRT-only 7B feasibility run. Dataset and tokenizer revisions, cache hashes, and the optimizer-source hash are recorded in Appendix B. The theorem statements, tail indexing, and proof mechanisms are in Appendices N–O. Component and sensitivity evidence remains in Appendices J–K.2. Anonymized experiment code, frozen result and figure artifacts (including CCRT-only scale-feasibility artifacts for the 7B run), and proof-verification artifacts are provided in the supplement.

REFERENCES

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 3920–3930, 2025.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As Lyapunov predicts. In *International Conference on Learning Representations*, 2024.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2023.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022.

Dmitrii Feoktistov, Timofey Belinsky, Andrey Veprikov, Amir Zainullin, and Aleksandr Beznosikov. Softsign: Smooth sign in your optimizer for better parameter heterogeneity handling. *arXiv preprint arXiv:2605.31371*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yao Lu, Dengdong Fan, Shixun Zhang, and Yonghong Tian. PowerStep: Memory-efficient adaptive optimization via ℓ_p-norm steepest descent. *arXiv preprint arXiv:2605.10335*, 2026.

Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and Adam for deep learning. In *International Conference on Learning Representations*, 2019.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Hanyang Peng, Shuang Qin, Yue Yu, Fangqing Jiang, Hui Wang, and Wen Gao. SoftSignSGD (S3): An enhanced optimizer for practical DNN training and loss spikes minimization beyond Adam. *arXiv preprint arXiv:2507.06464*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604, 2018.

Ye Yuan, Mu Li, Jun Liu, and Claire J. Tomlin. On the powerball method for optimization. *arXiv preprint arXiv:1603.07421*, 2016.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Implementation Details

A.1 Exact CCRT State and Update Semantics

CCRT maintains one full-shape BF16 carry tensor for each parameter that receives a gradient, together with a non-tensor integer update counter. It does not maintain a second-moment tensor. The counter is incremented per parameter update for which that parameter receives a gradient; therefore, the index t in the optimizer equations is formally a per-parameter update count rather than an intrinsic global training-step index.

For a parameter with gradient G_t, the implementation computes

$$\alpha_t = \frac{1}{\min(t, H)}, \tag{9}$$

followed by the FP32 transient carry

$$C_t^* = (1 - \alpha_t)\,\mathrm{float32}(\bar{C}_{t-1}) + \alpha_t G_t. \tag{10}$$

The persistent state is represented as

$$\bar{C}_t = Q_{\mathrm{BF16}}(C_t^*). \tag{11}$$

Crucially, the current update is formed using the freshly constructed FP32 value C_t^* rather than by decoding the newly persisted BF16 state. The pre-transport signal is

$$D_t = G_t + C_t^*, \tag{12}$$

and the coordinate-wise nonlinear transport is

$$U_t = \mathrm{sign}(D_t)\frac{\sqrt{|D_t|}}{1 + \sqrt{|D_t|}}. \tag{13}$$

The general parameter update is

$$\theta_{t+1} = \theta_t(1 - \eta_t\lambda) - \eta_t U_t, \tag{14}$$

where λ denotes decoupled weight decay. All main experiments use $\lambda = 0$ and $H = 32$.

We use the term *single-state* to mean one full-shape optimizer tensor state per parameter. The implementation additionally persists the integer update counter described above.

A.2 External Gradient Clipping and Learning-Rate Scheduling

Gradient clipping and learning-rate scheduling are external to the CCRT transport itself. Gradients are accumulated over four sequential size-one chunks. Each chunk loss is divided by four before backward. After accumulation, a global ℓ_2 gradient-norm clip with maximum norm 1.0 is applied before `optimizer.step()`.

The learning rate η_t is supplied by the parameter group at the time of the optimizer update. The optimizer consumes the current rate, after which the external scheduler advances the schedule.

A.3 Additional Property Derivations

For the ideal recurrence, induction gives

$$C_t = \frac{1}{t}\sum_{k=1}^{t} G_k, \qquad 1 \le t \le H. \tag{15}$$

For $t > H$, with $\beta = 1 - 1/H$,

$$C_t = \beta^{t-H} C_H + \frac{1}{H}\sum_{j=0}^{t-H-1} \beta^j G_{t-j}. \tag{16}$$

These exact identities apply only to the ideal unquantized carry.

Writing $r = \sqrt{|x|}$ yields $|\phi(x)| = r/(1+r) < 1$. The expansions

$$\phi(x) = \operatorname{sign}(x)\sqrt{|x|} - x + \mathcal{O}(|x|^{3/2}), \tag{17}$$

$$\phi(x) = \operatorname{sign}(x)\left(1 - |x|^{-1/2} + \mathcal{O}(|x|^{-1})\right) \tag{18}$$

give the small- and large-signal limits. For $x \neq 0$,

$$\phi'(x) = \frac{1}{2\sqrt{|x|}(1+\sqrt{|x|})^2}, \tag{19}$$

which diverges at the origin; bounded output is not a global contraction.

A.4 Generalized Lion-κ Mapping

For the ideal carry,

$$D_t = (1+\alpha_t)G_t + (1-\alpha_t)C_{t-1}. \tag{20}$$

Set $\beta_{2,t} = 1-\alpha_t$, $\beta_{1,t} = (1-\alpha_t)/2$, and $z_t = (1-\beta_{1,t})G_t + \beta_{1,t}C_{t-1}$, so $D_t = 2z_t$. For

$$\kappa(z) = |z| - \sqrt{2|z|} + \log(1+\sqrt{2|z|}), \tag{21}$$

$\kappa'(z) = \phi(2z)$, hence $U_t = \kappa'(z_t)$. At $H = 32$ the post-horizon coefficients are $\beta_2 = 31/32$ and $\beta_1 = 31/64$. This is a positioning relationship within generalized Lion-κ.

A.5 Persistence-Precision Perturbation

For a persisted-carry perturbation e_{t-1} and unchanged current gradient,

$$\delta C_t^* = (1-\alpha_t)e_{t-1}. \tag{22}$$

After $H = 32$ the factor is $31/32$, so inherited perturbations are only weakly attenuated. Together with the non-Lipschitz transport near zero, this motivates empirical BF16 evaluation but not arbitrary low-bit claims.

B Reproducibility and Training Protocol

B.1 Model Architecture

All primary optimizer comparisons use the same 134,764,032-parameter causal decoder-only Transformer. The model contains 16 decoder layers, hidden width 768, 12 attention heads, one key-value head, a SwiGLU feed-forward block of width 2048 (Shazeer, 2020), RMSNorm (Zhang & Sennrich, 2019), learned absolute positional embeddings, tied token-embedding and output-head weights, and no dropout. The vocabulary size is 50,257.

Model parameters, forward computation, and gradients are FP32 in the reported experiments. The CCRT persistent carry is BF16.

B.2 Corpus and Tokenization

Training text is drawn from `HuggingFaceFW/fineweb-edu` (Penedo et al., 2024), configuration `sample-10BT`, using dataset revision `87f09149ef4734204d70ed1d046ddc9ca3f2b8f9`.

Documents are tokenized using `openai-community/gpt2` (Radford et al., 2019), revision

`607a30d783dfa663caf39e06633721c8d4cfcd7e`.

Exactly one end-of-text token is appended to each nonempty document. No BOS token or padding token is inserted into the training stream.

The streamed source is materialized into deterministic local `uint16` token caches before the reported comparison runs. The frozen training-cache SHA-256 is

`6ae1472d45bb63a01b006d7c4de3ae89470e490b5db754235bb2685199e7d18c`.

B.3 Batch and Optimization Protocol

Sequence length is 256 tokens. Each optimizer update processes four sequences, corresponding to 1,024 training tokens per update. Operationally, the four sequences are evaluated as four sequential size-one chunks in fixed index order, with each chunk loss divided by four before backward.

All primary 10K runs use zero weight decay and global gradient clipping at norm 1.0.

The learning-rate schedule consists of 10 linear warmup steps followed by cosine decay through step 10,000 to a minimum learning-rate ratio of 0.1.

B.4 Validation Protocol

Validation uses a fixed deterministic cache containing 131,072 tokens. For each evaluation, the model is placed in evaluation mode and the validation cache is processed without mutating training state.

Validation cross-entropy is computed as

$$\mathrm{CE} = \frac{1}{N}\sum_{i=1}^{N} \ell_i, \qquad N = 131072, \tag{23}$$

where ℓ_i is token-level cross-entropy. Perplexity is

$$\mathrm{PPL} = \exp(\mathrm{CE}). \tag{24}$$

Validation, checkpointing, and reload-adjacent work are excluded from the clean training-time measurements reported in the main systems comparison.

The frozen validation-cache SHA-256 is

`7fa4463490538b4c404477f4f3c900bbfe5136b9c63b75fd4519cc542ba969ae`.

B.5 Execution Environment

The frozen optimizer source `phase8_r3/optimizer.py` matches SHA-256

`33788111978c65c1059a57703b4dece01d086f0c6fcc1066ec3d6bfbd0037ede`.

Experiment-time software and systems provenance is taken from the frozen experiment artifacts rather than from the present development machine.

C Comparator Calibration

Comparator learning rates were selected in a calibration stage that was separate from the official fresh-seed evaluation. CCRT itself was not recalibrated in this stage, and CCRT metrics were not used to select comparator hyperparameters.

Two calibration seeds, 26081601 and 26081603, were used for 200-step screening runs. The frozen learning-rate candidate grids were

$$\text{AdamW}: \quad \{1.8\times10^{-4}, 6\times10^{-4}, 1.8\times10^{-3}\}, \tag{25}$$

$$\text{PowerStep}: \quad \{1.8\times10^{-4}, 6\times10^{-4}, 1.8\times10^{-3}\}, \tag{26}$$

$$\text{Lion}: \quad \{3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}\}. \tag{27}$$

Among stable candidates, selection first minimizes median training CE over the final 50 calibration steps. Candidates within 0.005 CE of the best value remain eligible. Remaining ties are resolved using the median final-100-step CE slope with tolerance 10^{-5}, followed by absolute log-distance from the predefined learning-rate anchor.

The selected survivor configurations are:

- CCRT: learning rate 0.01, $H = 32$, BF16 persistent carry;

- AdamW: learning rate 0.0006, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, `amsgrad=False`, using `torch.optim.AdamW(fused=True)`;
- PowerStep: learning rate 0.0006, momentum coefficient $\gamma = 0.9$, signed-power exponent $\beta = 0.1$, using a PyTorch foreach implementation.

The official fresh paired seeds are distinct from the calibration seeds, and no post-hoc learning-rate retuning is performed after observing the official comparison results.

D METRIC DEFINITIONS

Validation AUC. Let (s_i, L_i) denote validation checkpoint step and validation CE. The normalized AUC over $[s_a, s_b]$ is computed by trapezoidal integration over the frozen checkpoint grid:

$$\mathrm{AUC} = \frac{1}{s_b - s_a} \sum_i \frac{L_i + L_{i+1}}{2}(s_{i+1} - s_i). \tag{28}$$

Temporal slopes. Early-, mid-, and late-window slopes are ordinary least-squares slopes of validation CE against optimizer step using all frozen checkpoints in the intervals

$$[100, 1000], \qquad [1000, 5000], \qquad [5000, 10000], \tag{29}$$

respectively.

Persistent crossover. For CCRT and comparator B, define the checkpoint difference

$$d_i = L_i^{\mathrm{CCRT}} - L_i^B. \tag{30}$$

We determine the initial nonzero sign and identify the first later checkpoint at which that sign reverses and remains reversed for that checkpoint and the following two frozen checkpoints. The first checkpoint of this three-point sequence is the reported persistent crossover.

Relative parameter update. For a deterministic sampled index set $\mathcal{I}$ from representative parameter families,

$$r = \frac{\mathrm{RMS}(\theta_{\mathrm{after},\mathcal{I}} - \theta_{\mathrm{before},\mathcal{I}})}{\mathrm{RMS}(\theta_{\mathrm{before},\mathcal{I}}) + 10^{-12}}. \tag{31}$$

Persistent optimizer state.

Optimizer-state bytes are computed by summing `numel` × `element_size` over tensor values stored in optimizer state after the run. Python integer metadata is not included.

Peak memory. Peak memory is measured using `torch.cuda.max_memory_allocated` after resetting the CUDA allocator statistic immediately before the first training step.

Complete-step timing. Complete-step timing begins immediately before batch fetch/forward and includes forward computation, backward computation, gradient clipping, `optimizer.step()`, and scheduler advancement. CUDA-event timings are materialized after synchronization.

The primary clean median is computed over steps 21–10,000 and excludes validation, checkpointing, startup, and reload-adjacent timing.

Optimizer timing. Optimizer timing brackets `optimizer.step()` only. Gradient clipping is outside this interval.

Throughput. Median throughput is computed from

$$\mathrm{tokens/s} = \frac{\text{tokens per step}}{\text{median complete step time}/1000}. \tag{32}$$

Table 4: Validation CE and trajectory summaries for the two primary fresh paired seeds. Lower CE and AUC are better. More negative mid and late slopes indicate faster validation-CE descent within the corresponding finite window.

Seed	Optimizer	100	500	1K	2.5K	5K	7.5K	10K	$\mathrm{AUC}_{100:10K}$	$\mathrm{AUC}_{1K:10K}$	Mid slope	Late slope
A	CCRT	7.8664	7.4746	7.1517	6.7009	6.3519	6.0469	5.8610	6.4380	6.3278	-1.786e-4	-1.009e-4
A	AdamW	7.8848	7.1576	6.9518	6.6035	6.4293	6.2792	6.2072	6.5154	6.4342	-1.134e-4	-4.918e-5
A	PowerStep	7.9478	7.1908	7.0645	6.8856	6.7696	6.6492	6.5747	6.8166	6.7630	-6.126e-5	-4.446e-5
B	CCRT	7.8682	7.3963	7.1579	6.6900	6.3456	6.0531	5.8648	6.4339	6.3278	-1.809e-4	-1.005e-4
B	AdamW	7.8752	7.1299	6.9360	6.6023	6.4172	6.2621	6.1915	6.5033	6.4233	-1.143e-4	-5.145e-5
B	PowerStep	7.9481	7.1883	7.0886	6.8956	6.8019	6.6919	6.6252	6.8469	6.7967	-5.509e-5	-4.102e-5

Table 5: Systems measurements in the frozen fresh-seed harness. Timing is implementation- and hardware-specific.

Optimizer	Complete ms	Optimizer ms	Tokens/s	State bytes
CCRT	224.9587	6.6583	4552.1093	269,528,064
AdamW	241.5968	23.0659	4238.4676	1,078,112,844
PowerStep	252.2784	33.8022	4059.0455	539,056,128

E Detailed Fresh-Seed Results

The primary replication contains exactly two fresh paired seeds. We make no statistical-significance claim from $n = 2$.

Table 4 reports the source-closed validation checkpoints and trajectory summaries for both seeds.

All six optimizer runs reach their best reported validation value at step 10,000. CCRT has lower terminal CE and lower validation-CE AUC than both comparators on each fresh paired seed. Persistent crossover is also observed against both comparators on both fresh seeds.

F Historical Third-Seed Sensitivity

Before the frozen two-fresh-seed replication, we collected an independent 10K run using seed 26081701. This run is not counted as a member of the primary replication and was not generated under the frozen two-seed replication protocol. It appears here only as historical independent confirmation.

At step 10K, validation CE is

$$5.8565 \text{ (CCRT)}, \qquad 6.2130 \text{ (AdamW)}, \qquad 6.6075 \text{ (PowerStep)}. \tag{33}$$

Normalized AUC over steps 100–10K follows the same ordering:

$$6.4267, \qquad 6.5176, \qquad 6.8369. \tag{34}$$

The persistent crossover against PowerStep occurs at step 250, whereas the crossover against AdamW occurs at step 4000. This differs from the two fresh replication seeds, for which both pairwise crossings occur at step 250.

Across the two primary fresh seeds and this historical run, the observed terminal-CE ordering, AUC ordering, and existence of persistent crossover are consistent in all $3/3$ observed seeds against both comparators. We do not treat these three runs as a preregistered three-seed statistical sample. Figure 3 summarizes the terminal values.

G Systems Measurement Details

Table 5 gives the pooled fresh-two-seed systems measurements.

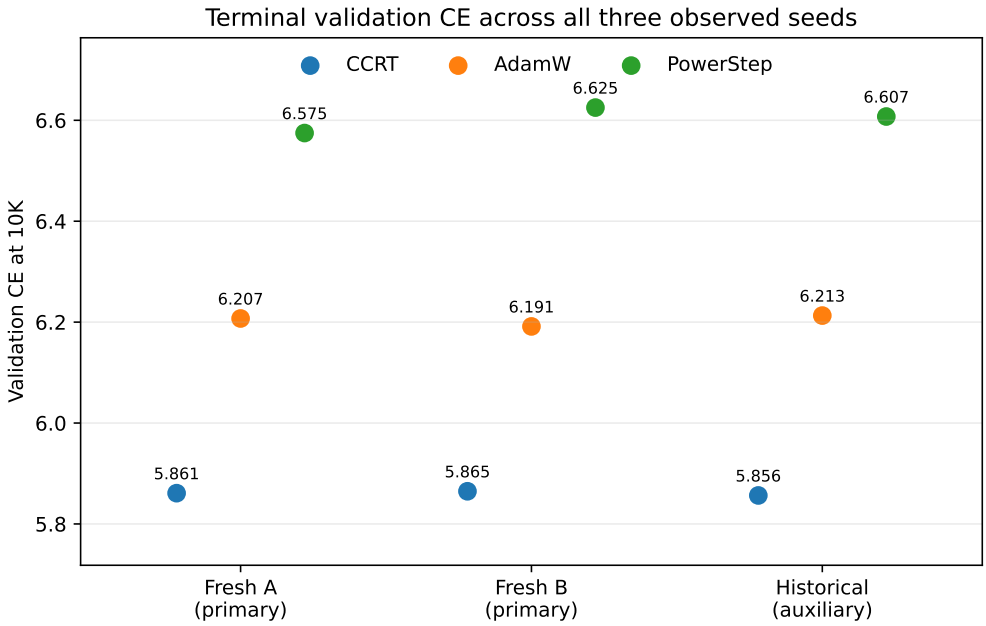


Figure 3: Terminal validation CE for the two primary fresh seeds and the auxiliary historical seed.

Table 6: Frozen long-horizon CCRT validation checkpoints.

Step	Validation CE ↓	PPL ↓
50,000	5.33465	207.400
75,000	5.15424	173.164
100,000	5.01995	151.403

Relative to AdamW, CCRT uses approximately 75% less persistent optimizer tensor state and its measured optimizer execution is approximately $3.46\times$ faster. Median complete-step time is approximately 6.9% lower, and median throughput is approximately 7.4% higher.

Relative to PowerStep, CCRT uses 50% less persistent optimizer tensor state, its measured optimizer execution is approximately $5.08\times$ faster, median complete-step time is approximately 10.8% lower, and median throughput is approximately 12.1% higher.

These values describe the measured implementations on the experimental hardware. Persistent optimizer tensor state should not be interpreted as total training memory or peak GPU allocation. Figure 4 provides the corresponding state–optimizer-time view.

H Long-Horizon CCRT Details

The long-horizon experiment follows one frozen CCRT trajectory. Because there is no matched AdamW or PowerStep run, it addresses long-horizon viability rather than comparator performance.

The run reaches 100,000 optimizer steps, corresponding to 102.4M training tokens. Validation results at the principal long-horizon checkpoints are shown in Table 6.

The best frozen validation checkpoint is the terminal 100K checkpoint. From 50K to 100K, validation CE decreases by 0.31470, corresponding to approximately 5.9% relative to the 50K value. Validation perplexity decreases by approximately 27.0% over the same interval.

The run remains numerically finite with no recorded hard failure. Carry health is reported as `HEALTHY`, scale telemetry as `STABLE`, and activation telemetry as bounded, finite, and non-runaway despite entering a warning regime.

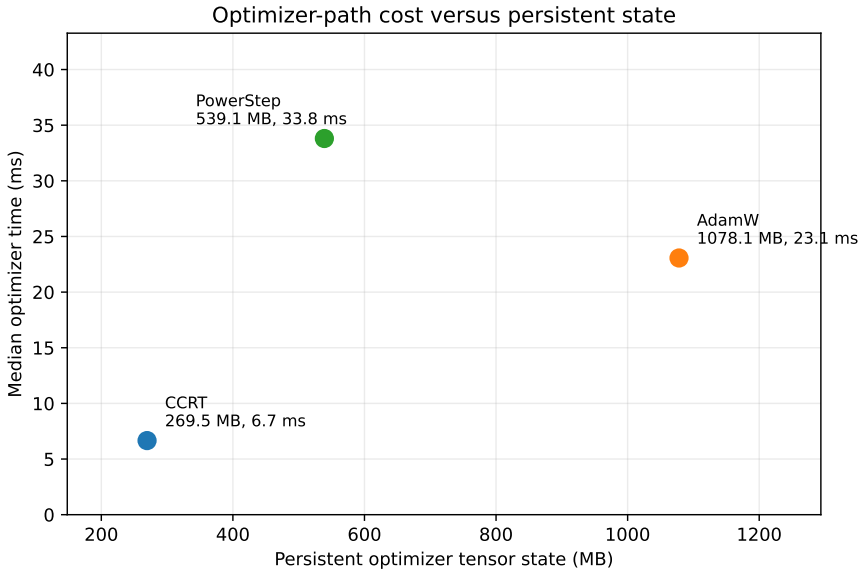


Figure 4: Persistent optimizer tensor state versus measured median optimizer execution time. Timing is implementation- and hardware-specific.

The authoritative long-horizon artifacts define no aggregate root-transport status class, so none is assigned here.

H.1 Relative-Update Telemetry

The maximum sampled relative update over monitored parameter families is approximately

$$1.121 \times 10^{-3}, \qquad 1.120 \times 10^{-3}, \qquad 1.130 \times 10^{-3} \tag{35}$$

at steps 50K, 75K, and 100K, respectively.

The approximately stable upper envelope coexists with continued validation improvement through the final checkpoint. It is a descriptive property of the trajectory, not a convergence criterion. Figure 5 shows the complete trajectory.

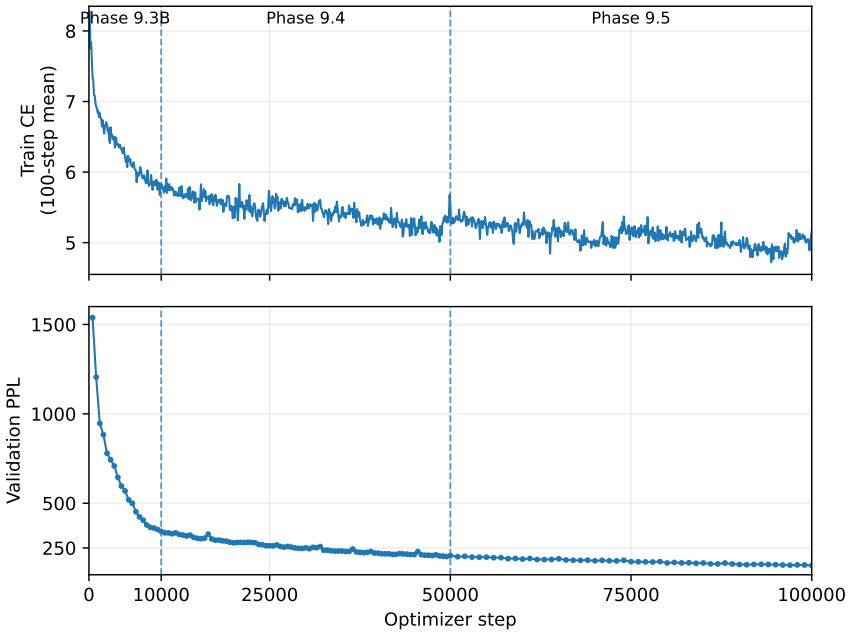


Figure 5: Full frozen CCRT trajectory through 100K optimizer steps. This is long-horizon viability evidence, not a paired optimizer comparison.

I Additional Optimizer Screening

The final 10K survivor comparison was preceded by a shorter 1K four-optimizer screen containing CCRT, AdamW, Lion, and PowerStep. Comparator learning rates were selected using the calibration procedure described in Appendix C.

The 1K screen was used as a filtering experiment rather than as the primary quality claim. AdamW had the strongest terminal and trajectory quality over this short horizon, followed by PowerStep and CCRT, while Lion was substantially weaker and was not retained in the frozen 10K survivor roster.

CCRT nevertheless showed the most negative late-window validation slope in this screen, motivating the longer-horizon survivor experiment rather than elimination based solely on the 1K terminal ordering.

The subsequent 10K experiments show that short-horizon ranking is not necessarily representative of the longer-horizon ordering observed in this setup.

J Component Ablation Details

These frozen single-seed diagnostics complement rather than replace the primary paired comparison. At step 2500, Full CCRT has CE 6.700906 and $\mathrm{AUC}_{100:2500}$ 7.126777.

No Carry uses $D_t = G_t$ with the root-rational readout unchanged. Sign Readout retains $D_t = G_t + C_t^*$ but uses $U_t = \mathrm{sign}(D_t)$. Both gaps favor retained history and finite-magnitude transport at the frozen operating point. They do not establish universal necessity.

Three additional variants—Fixed Beta Carry, Carry Only Sign, and No Carry Sign—completed exactly 2,500 steps without NaN/Inf, fallback, or hard failure. No second seed, old-reference rerun, or further variant was run. Fixed Beta Carry remained near Full CCRT and motivated the prespecified 10K extension; both sign combinations produced clear quality gaps.

Carry Only Sign retains the scheduled carry, removes explicit fresh-gradient reinjection, and uses $D_t = C_t^*$ with $U_t = \mathrm{sign}(D_t)$. No Carry Sign has no retained optimizer tensor state and uses

Table 7: Component and interaction diagnostics at step 2500. Deltas are relative to Full CCRT; lower quality metrics are better.

Variant	CE	Δ CE	$\text{AUC}_{100:2500}$	State bytes
Full CCRT	6.700906	—	7.126777	269,528,064
No Carry	7.129969	+0.429063	7.407246	0
Carry Only	6.686101	-0.014805	—	269,528,064
Sign Readout	7.873543	+1.172636	8.294845	269,528,064
Fixed Beta Carry	6.670558	-0.030349	7.071921	269,528,064
Carry Only Sign	8.268300	+1.567394	8.790119	269,528,064
No Carry Sign	8.111223	+1.410317	8.291916	0

Table 8: Selected single-seed interaction contrasts at step 2500.

Interaction	CE contrast	AUC contrast
History × Readout	+0.191382	+0.283397
Fresh × Readout	-0.409563	-0.464711

$D_t = G_t$ with $U_t = \text{sign}(G_t)$. The source-closed summary does not report Carry Only $\text{AUC}_{100:2500}$, so Table 7 does not reconstruct one from rounded interaction values.

J.1 Selected Interaction Diagnostics

The predefined scores are $Y = -\text{CE}$ and $Y_{\text{AUC}} = -\text{AUC}$. Table 8 reports the selected difference-in-differences contrasts.

Under the root-rational readout, adding retained history improves the 2.5K CE endpoint by approximately 0.429; under sign, the corresponding improvement is approximately 0.238. Retained history and readout are therefore not purely additive at this endpoint: the measured history benefit is larger under the root-rational readout. Fresh reinjection has little or slightly adverse endpoint effect with the root-rational readout at 2.5K, but a substantially larger positive effect under sign. These contrasts are endpoint-specific single-seed diagnostics, without significance or general causal interpretation.

J.2 Carry-Only 10K Replay

Table 9 reports the Carry Only replay with $D_t = C_t^*$. It exactly reproduced the earlier checkpoints through 2.5K and continued to 10K.

After 2.5K, the ordering reverses and the gap widens through 10K. The classification is `SIGN_REVERSAL`. In this run, fresh reinjection contributes at longer horizons; general necessity is not established.

J.3 Fixed-Beta Carry Diagnostic

Fixed Beta Carry replaces the scheduled coefficient by $\alpha = 1/32$ from the first step. With $\beta = 31/32$ and no bias correction, its carry and readout are

$$C_t^* = \beta\,\text{float32}(\bar{C}_{t-1}) + (1-\beta)G_t, \tag{36}$$

$$\bar{C}_t = Q_{\text{BF16}}(C_t^*), \qquad D_t = G_t + C_t^*. \tag{37}$$

The root-rational transport is unchanged. At 2.5K, Fixed Beta Carry has CE 6.670558, a delta of -0.030349 relative to Full CCRT, and $\text{AUC}_{100:2500}$ 7.071921. Its persistent optimizer state is $269{,}528{,}064$ bytes. This near-equivalence motivated the prespecified 10K extension; it did not motivate a change to the frozen CCRT method.

J.4 Fixed-Beta 10K Replay

Only Fixed Beta Carry was run in the extension. The original 2.5K diagnostic did not persist a resumable checkpoint, so steps 0–2500 are a deterministic integrity replay and steps 2501–10000

Table 9: Full CCRT and Carry-Only replay. Positive gaps favor Full CCRT.

Metric	Full CCRT	Carry Only	Carry Only − Full
CE@2500	6.700906	6.686101	-0.014805
CE@5000	6.351926	6.390803	+0.038877
CE@7500	6.046900	6.124124	+0.077224
CE@10000	5.860996	5.987265	+0.126270
$\text{AUC}_{100:10K}$	6.437977	6.500670	+0.062693
$\text{AUC}_{1K:10K}$	6.327826	6.387604	+0.059778
$\text{Slope}_{1K:5K}$	-0.000178646	-0.000172239	+0.000006407
$\text{Slope}_{5K:10K}$	-0.000100901	-0.000094871	+0.000006031

Table 10: Full CCRT and Fixed-Beta Carry through 10K. Deltas are Fixed Beta minus Full CCRT.

Metric	Full CCRT	Fixed Beta	Fixed − Full
CE@2500	6.700906303	6.670557652	-0.030348651
CE@5000	6.351926353	6.312204976	-0.039721377
CE@7500	6.046899974	6.019653630	-0.027246345
CE@10000	5.860995505	5.834226627	-0.026768878
$\text{AUC}_{100:10K}$	6.437976851	6.401057350	-0.036919501
$\text{AUC}_{1K:10K}$	6.327825904	6.297302879	-0.030523024
$\text{Slope}_{1K:5K}$	-0.000178646	-0.000181147	-0.000002501
$\text{Slope}_{5K:10K}$	-0.000100901	-0.000098965	+0.000001936
Best CE	5.860995505	5.834226627	-0.026768878
Best step	10000	10000	0

provide the new extension evidence. The replay exactly matches the original 2.5K validation trajectory, and a resumable checkpoint was saved at step 10,000. Table 10 reports quality metrics; Tables 11 and 12 report systems measurements and checkpoint gaps.

Fixed Beta is ahead at 500 steps. The gap narrows later but remains modestly favorable at 10K and in both AUC windows. The scheduled carry continues to define Full CCRT, but this diagnostic does not distinguish the early running-mean schedule favorably from the uncorrected fixed EMA.

J.5 Mechanistic Hypothesis for the Schedule Diagnostic

The following is a hypothesis, not an empirical mechanism claim. For the ideal scheduled recurrence and $t \leq 32$,

$$C_t^{\text{sched}} = \frac{1}{t} \sum_{k=1}^{t} G_k, \tag{38}$$

so its carry coefficients sum to one from the first step. The uncorrected fixed recurrence initialized at zero instead gives

$$C_t^{\text{fix}} = \frac{1}{32} \sum_{k=1}^{t} \beta^{t-k} G_k, \qquad \beta = \frac{31}{32}, \tag{39}$$

with total coefficient mass $1 - \beta^t$. This mass is $1/32 \approx 0.03125$ at $t = 1$ and approximately $1 - (31/32)^{32} \approx 0.64$ at $t = 32$. At the first update,

$$C_1^{*,\text{sched}} = G_1, \quad D_1^{\text{sched}} = 2G_1, \qquad C_1^{*,\text{fix}} = \frac{G_1}{32}, \quad D_1^{\text{fix}} = \frac{33}{32} G_1. \tag{40}$$

We hypothesize that the uncorrected fixed-beta carry may act as an implicit soft start for the carry path. Because CCRT already contains a direct fresh-gradient path, the scheduled recurrence gives the carry full normalized weight immediately, whereas the fixed EMA introduces it gradually. The root-rational map is nonlinear, so changing the early scale of D_t is not generally equivalent to uniformly rescaling the learning rate; it may alter update RMS, relative coordinate scales, proximity to transport saturation, and the balance between fresh-gradient and carry information.

Table 11: Fixed-Beta replay state, systems, and failure measurements.

Metric	Full CCRT	Fixed Beta	Fixed − Full
State bytes	269528064	269528064	0
Optimizer ms	6.670688152	6.661503792	-0.009184361
Complete-step ms	226.301132202	224.379814148	-1.921318054
Tokens/s	4524.944219392	4563.690383150	+38.746163758
Hard failures	0	0	0

Table 12: Validation-CE gap, Fixed Beta minus Full CCRT, across the 10K replay.

Step	Gap	Step	Gap
100	+0.009119611	4500	-0.028167985
250	-0.097334750	5000	-0.039721377
500	-0.169812206	5500	-0.043310042
750	-0.103438701	6000	-0.024717137
1000	-0.029601093	6500	-0.036114812
1500	-0.036289450	7000	-0.018507577
2000	-0.015454073	7500	-0.027246345
2500	-0.030348651	8000	-0.031645492
3000	-0.024065137	8500	-0.032566447
3500	-0.031533893	9000	-0.029233072
4000	-0.044323944	9500	-0.027984019
—	—	10000	-0.026768878

If future gradients were held fixed, an early carry-state perturbation would be attenuated by the post-horizon β factor. In the actual optimization runs, however, parameter trajectories and subsequent gradients need not remain identical. The persistent validation gap is therefore consistent with an early path-dependence hypothesis, but does not establish it. Exponential recency weighting itself may instead be more favorable than the scheduled early uniform average. The present experiment confounds early carry amplitude with uniform-versus- exponential weighting; a bias-corrected fixed-EMA diagnostic could separate them in future work.

K Hyperparameter Sensitivity

K.1 History-Horizon Sensitivity

Table 13 reports Full CCRT for $H \in \{16, 32, 64, 128\}$ under one frozen seed, base learning rate 0.01, identical initialization and stream, and the frozen 10K scheduler observed through step 5K. The $H = 32$ reference was not rerun. All variants remained finite without NaN/Inf, fallback, or hard failure.

At 5K, the CE gaps relative to $H = 32$ are $+0.024593$, $+0.061032$, and $+0.110155$ for $H = 16, 64, 128$. Each non-reference variant is `MIXED_TEMPORAL`; the overall classification is `PREFERS_H32_REGION`. The experiment probes local sensitivity; it does not establish global optimality or general robustness.

Table 13: Full single-seed 5K history-horizon sensitivity results.

Metric	$H = 16$	$H = 32$	$H = 64$	$H = 128$
CE@500	7.381324	7.474621	7.562852	7.632000
CE@1000	7.163658	7.151689	7.176511	7.259438
CE@2500	6.708267	6.700906	6.708739	6.736638
CE@3500	6.596481	6.592614	6.614535	6.657501
CE@5000	6.376519	6.351926	6.412958	6.462082
$\text{AUC}_{100:5000}$	6.830985	6.826333	6.849768	6.892401
$\text{AUC}_{1000:5000}$	6.677406	6.665874	6.687180	6.728922
$\text{Slope}_{1000:5000}$	-0.000173696	-0.000178646	-0.000168362	-0.000171121
Best step	5000	5000	5000	5000

K.2 Learning-Rate Diagnostic

Two new Full CCRT runs test base learning rates 0.005 and 0.0025 with seed 1954664097, $H = 32$, BF16 persistent carry, zero weight decay, and global clip norm 1.0. Model initialization, training and validation caches, batching, and chunking match the frozen reference. Both new runs completed exactly 5,000 steps without NaN/Inf or hard failure. No second seed, horizon rerun, additional learning rate, retuning, or ablation was run. The 0.01 reference is reused from its source-closed trajectory rather than rerun.

The comparison is not scheduler matched. Both lower-rate runs use 10-step warmup followed by cosine decay to 0.1 times the base rate at step 5,000. The frozen 0.01 reference uses its original 10-step warmup and 10,000-step cosine horizon, reaching the same ratio only at step 10,000. Base-rate scale and scheduler horizon therefore change together. Table 14 reports the full metrics, and Table 15 gives the checkpoint gaps.

LR 0.005 is classified `MIXED_TEMPORAL`: it has lower AUC in both windows, but a less-negative 1K–5K slope and a terminal CE gap of $+0.077801093$. LR 0.0025 is classified `MODERATE_DEGRADATION`, with a terminal gap of $+0.211757697$, higher AUC, and the least-negative slope. The overall classification is `MIXED`. The late separation is consistent with the shorter 5K cosine trajectory reducing the effective learning rate earlier, but the design does not separate scheduler-horizon and base-rate effects.

Table 14: Full single-seed learning-rate diagnostic. The lower-rate runs use a 5K cosine endpoint; the 0.01 reference retains its 10K endpoint.

Metric	LR 0.01	LR 0.005	LR 0.0025
CE@500	7.474621154	7.521079917	7.515961446
CE@1000	7.151688728	7.133575380	7.193537083
CE@2500	6.700906303	6.662433792	6.736642815
CE@3500	6.592614323	6.544558708	6.640077114
CE@5000	6.351926353	6.429727446	6.563684050
$AUC_{100:5000}$	6.826333039	6.808012931	6.884719114
$AUC_{1000:5000}$	6.665873550	6.645706780	6.734662120
$Slope_{1000:5000}$	-0.000178646	-0.000159202	-0.000139023
Best CE	6.351926353	6.429727446	6.563684050
Best step	5000	5000	5000
State bytes	269528064	269528064	269528064
Optimizer ms	6.642720222	6.650799990	6.651215792
Complete ms	225.266113281	224.565856934	231.030143738
Tokens/s	4545.734753818	4559.909569436	4432.322048685
NaN	0	0	0
Inf	0	0	0
Hard failure	0	0	0

Table 15: Checkpoint CE gaps relative to the frozen LR 0.01 reference.

Step	LR 0.005 gap	LR 0.0025 gap
100	+0.026369989	+0.191699661
250	-0.064705916	-0.079329923
500	+0.046458762	+0.041340292
750	-0.029954154	-0.012560494
1000	-0.018113349	+0.041848354
1500	-0.027652074	+0.038766734
2000	-0.038927495	+0.019446354
2500	-0.038472511	+0.035736512
3000	-0.049461771	+0.031144314
3500	-0.048055615	+0.047462791
4000	-0.014265224	+0.100332256
4500	+0.025656655	+0.150616575
5000	+0.077801093	+0.211757697

L Approximately 3B Scale Reproduction

The scale reproduction uses a 3,000,674,304-parameter model, 10,000 optimizer updates, and 81,920,000 training tokens per run on an NVIDIA A800 80GB. AdamW, CCRT, and PowerStep use frozen learning rates 0.0003, 0.005, and 0.0003, respectively. All three runs completed without hard failure. The optimizer-specific learning rates were selected before the official 3B runs and then frozen; the corresponding selection record and exact configurations are included in the supplementary artifacts. The runs share the official seed definition, source configuration, budget, and validation schedule. Their recorded serialized initialization hashes differ; because the original files are no longer available for tensor-level comparison, this experiment is not described as strictly tensor-identical paired evidence.

CCRT uses 75% less persistent optimizer state than AdamW and 50% less than PowerStep. Its peak allocated memory is 31.61% below AdamW and 13.35% below PowerStep. These state ratios follow from the optimizer representations; wall-clock and throughput remain specific to the measured implementations and hardware.

Table 16: Validation CE checkpoints for the 3B single-seed scale reproduction. Values are recorded evaluations; no intermediate points are inferred.

Step	AdamW	CCRT	PowerStep
100	7.743309	7.681376	7.907777
250	6.986978	7.273258	7.178867
500	6.570243	6.815938	6.741767
1000	6.270088	6.474894	6.462756
2500	5.890414	5.899434	6.164954
5000	5.763298	5.287305	6.179944
7500	5.617660	4.971864	6.037755
10000	5.506206	4.845351	5.936574

Table 17: Quality and systems summary for the 3B reproduction. Peak allocation is PyTorch allocated memory, not total device memory.

Metric	AdamW	CCRT	PowerStep
Final validation CE	5.506206	4.845351	5.936574
Final validation PPL	246.2151	127.1479	378.6355
Wall-clock (h)	6.738	6.691	6.895
Throughput (tokens/s)	3377.1	3401.1	3300.3
Peak allocated (GiB)	53.05	36.29	41.87
Persistent state (GiB)	22.36	5.59	11.18
State precision	FP32	BF16	FP32

M SINGLE-GPU 7B SCALE FEASIBILITY

The CCRT-only scale-feasibility run uses a 6,981,064,704-parameter dense model for 1,000 updates and 8,192,000 training tokens on one NVIDIA A800 80GB. It completes all 1,000 steps in 95.16 minutes, with steady-state throughput of 1,439 tokens/s. Training CE decreases from 11.4165 to 6.4309, with a final-100-step mean of 6.5434. On the fixed 4,096-token validation set, CE decreases from 11.4419 to 6.7630 and perplexity from 93140.6 to 865.26; both metrics strictly decrease at all five recorded checkpoints. All 1,000 training steps trigger the global gradient-norm clip at 1.0, so the trajectory is observed under the shared clipping constraint.

Peak PyTorch allocated and reserved memory are 67.01 GiB and 68.16 GiB, respectively, with approximately 11.09 GiB of measured remaining reserved headroom. The BF16 carry occupies 13.00 GiB, exactly 2 bytes per parameter. All 651 carry tensors are finite, the CCRT fallback record is $\{\}$, FlashAttention-2 is enabled, and all 1,000 metric rows are continuous and complete with finite recorded losses. The run-manifest SHA256 matches the frozen manifest. No `.pt`, `.pth`, `.ckpt`, `.bin`, or `.safetensors` checkpoint files are present; checkpoint files written and retained are both zero. These diagnostics establish operational and single-GPU feasibility under the tested configuration, not a same-scale optimizer comparison or a strong generalization claim.

N SCALAR FIXED-STEP DYNAMICS

Consider the deterministic full-precision scalar quadratic $f(x) = \lambda x^2/2$ with curvature $\lambda > 0$, gradient $g_t = \lambda x_t$, fixed-beta carry

$$C_t = \beta C_{t-1} + (1-\beta)g_t, \qquad x_{t+1} = x_t - \eta\phi(g_t + C_t),$$

and normalized step $h = \eta\lambda$. The origin is stationary, but the square-root behavior of ϕ near zero makes generic fixed-step trajectories overshoot rather than converge exactly.

Finite-h global classification. For $\beta = 31/32$ and every $0 < h \le 2^{-17}$, the map has a unique nonzero symmetric two-cycle. Every initial condition belongs exactly to one of two classes: the full stable set of the origin, or the basin of that two-cycle. There are no other nonzero periodic or omega-limit behaviors in this interval. The result is scalar, quadratic, deterministic, full precision, and fixed beta.

The proof uses a computer-assisted interval argument over the complete h interval rather than sampling parameter values. Its structure is to establish local cycle uniqueness, certify capture and outward motion on complementary regions, and combine those certificates with an absorbing energy argument. The certificates use outward-rounded bounds, so the classification covers each interval cell rather than only floating-point representatives.

On the attracting cycle,

$$|\nabla f| = \Theta(\eta^2\lambda^2), \qquad \mathcal{S}(\nabla f) = \Theta(\eta^3\lambda^3).$$

Thus fixed positive steps have a structured residual even in the exact model; the theorem does not support exact fixed-step convergence to the stationary point.

O Deterministic Smooth-Nonconvex Complexity

Assume throughout this section that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and bounded below. No convexity, PL inequality, bounded-gradient, bounded-iterate, or compact-sublevel assumption is used. The analysis concerns ideal full-precision CCRT; BF16 rounding is not part of the theorem.

O.1 Natural Stationarity Measure

Equation 6 defines

$$\mathcal{S}(g) = \sum_i g_i\phi(2g_i) = \sum_i \frac{\sqrt{2}|g_i|^{3/2}}{1+\sqrt{2|g_i|}}.$$

Let $\vartheta(r) = \sqrt{2}r^{3/2}/(1+\sqrt{2r})$. In addition to Eq. 7, the proved norm comparison is

$$\mathcal{S}(g) \geq \vartheta(\|g\|_{3/2}) \geq \vartheta(\|g\|_2).$$

The measure therefore vanishes only at first-order stationarity and directly converts to a conventional Euclidean gradient guarantee.

O.2 Fixed-Beta Tail and the Scheduled Prefix

For the fixed-beta tail, write $p_t = \phi(D_t)$ and $Q = \phi^{-1}$. The exact identity

$$Q(p_t) - \beta Q(p_{t-1}) = 2(1-\beta)g_t + \beta(g_t - g_{t-1}) \tag{41}$$

separates the no-lag alignment term from a gradient difference. Convex dual identities for Q make the latter telescope, while the residual carry mismatch is bounded by a stable geometric convolution. Treating $C_t - g_t$ only as an $O(\eta)$ perturbation would lose the cubic fixed-step scale.

The actual optimizer uses $\alpha_t = 1/\min(t, 32)$, not fixed beta from the first update. The first 32 running-mean steps enter through finite prefix and carry transients. If $\eta_{32} = \eta_{33} = \cdots$, the tail index is $a = 33$; if constant steps begin only at update 33, it is $a = 34$. These cases are kept distinct in the finite-time statement.

Suppressing only explicit constants and finite transient formulas, the proved bound for a constant tail step is

$$\min_{a\leq t<a+T} \mathcal{S}(\nabla f(x_t)) \leq \frac{1}{T}\sum_{t=a}^{a+T-1} \mathcal{S}(\nabla f(x_t)) = O\left(\frac{1}{\eta T}\right) + O(dL^3\eta^3).$$

Balancing the two terms gives $\mathcal{S} = O(T^{-3/4})$. Combining this rate with the norm comparison gives $T = O(\epsilon^{-2})$ for the existence of an iterate with $\|\nabla f(x_t)\|_2 \leq \epsilon$. At a fixed step, the corresponding gradient scale is $O(d^{2/3}L^2\eta^2)$. The scalar cycle in Appendix N matches both step-size exponents, so no uniform $o(\eta^3)$ natural-stationarity floor is possible in general.

O.3 Decaying Steps and Stochastic Boundary

For positive nonincreasing tail steps satisfying $\eta_t \to 0$ and $\sum_t \eta_t = \infty$, the η-weighted average stationarity measure converges to zero and $\liminf_t \|\nabla f(x_t)\|_2 = 0$. If additionally $\sum_t \eta_t\eta_{t-1}^3 < \infty$,

then the full gradient sequence converges to zero. In particular, power-law steps $\eta_t \asymp t^{-r}$ satisfy the latter conclusion for $1/4 < r \leq 1$; square summability of η_t is not required.

The analogous stochastic conclusion does not follow from conditional unbiasedness and uniformly bounded variance alone. For the actual $H = 32$ schedule, a one-dimensional history-adaptive oracle for $f(x) = \sin x$ can remain conditionally unbiased with finite support and uniformly finite variance, with $\mathbb{E}[p_t \mid \mathcal{F}_{t-1}] = 0$, while, with probability at least $197/200$, keeping $|\nabla f(x_t)| \geq 1/2$ for all time. The obstruction is that nonlinear transformation need not preserve unbiased alignment: $\mathbb{E}[\phi(g + \xi)] \neq \phi(g)$ in general.

O.4 Further Boundaries

Scalar cycle uniqueness does not extend directly to general quadratic models. Diagonal positive-definite quadratics can have multiple nonzero cycles; coordinatewise ϕ is not invariant under orthogonal diagonalization; and flat directions can generate continuous cycle families. These facts do not affect the nonconvex complexity proof, which uses Euclidean smoothness and summed coordinatewise dissipation rather than scalar uniqueness or one-step positive alignment.

More generally, a near-zero readout $\phi_p(z) \sim \operatorname{sign}(z)|z|^p$ with $0 \leq p < 1$ has the natural scalar fixed-step scale $\eta^{1/(1-p)}$. Sign-like $p = 0$ therefore gives an $O(\eta)$ scale, whereas CCRT's $p = 1/2$ gives $O(\eta^2)$; locally linear readouts belong to a different stability regime.

P Scope of Empirical Claims

Primary replicated quality evidence is limited to the 134.8M setup and two fresh paired 10K seeds. The historical seed is auxiliary and is not retroactively part of the frozen replication. The 3B comparison supplies one-seed scale evidence under a shared source configuration and budget; it is not described as tensor-identical paired evidence because the recorded serialized initialization hashes differ and the original files cannot be compared. The 100K trajectory remains CCRT-only viability evidence. The 6.98B experiment is a single-seed CCRT-only scale-feasibility run; it is not a same-scale optimizer comparison, and its 4,096-token validation set and always-active global clipping restrict its interpretation to operational feasibility under the tested configuration.

Selected component interactions are measured, but the primarily single-seed diagnostics are not a complete factorial or causal decomposition. Horizon and learning-rate studies are local; the latter also changes scheduler horizon and cannot isolate a globally optimal base rate. The early running-mean schedule has no demonstrated independent advantage over the tested fixed-beta alternative.

Runtime measurements are implementation-, harness-, and hardware-specific; optimizer tensor state is not total or peak training memory. The running-mean interpretation applies to the ideal unquantized recurrence, whereas the implementation persists BF16-rounded carry and uses the fresh FP32 transient carry for the current readout.

The deterministic complexity theorem concerns ideal full-precision CCRT on smooth objectives bounded below. The global dynamical classification is scalar and quadratic, and standard conditional unbiasedness with bounded variance is insufficient for a general stochastic theorem. The coordinate-wise exact-arithmetic bound implies neither bounded activations or function-space change nor robustness to arbitrary low-precision state. The two fresh seeds provide directional replication only; no statistical-significance or population-level superiority claim is made from $n = 2$.