BRIEF TREATMENTS SELECT LONG-RUN FUNCTION IN POST-LN MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a 25-update treatment change a model's function after thousands of unmodified updates? In an unstable Post-LN decoder trained without warmup, we modify only updates 4–28 of paired runs sharing an exact parent and then resume native AdamW through update 5000. On six complete master seeds, a scaled ascending learning-rate ramp is healthy on 6/6, while an unscaled ascending ramp and a norm-preserving directional intervention are each healthy on 5/6 and fail on different seeds; untreated ACTUAL, an exact-dose constant profile, and a release-continuous V0 profile are healthy on 0/6. Exact-dose reverse and half-LR controls also fail on the original seeds. Together with a successful release-shock control, these comparisons show that the tested outcomes separate strongly by within-window temporal organization beyond cumulative dose, release continuity, and sustained release-step magnitude. Transfer is recipe-dependent: at eight layers, 25-, 50-, and 100-update ramps are healthy on 0/3, 2/3, and 3/3 seeds; at 107M parameters, tested 25-, 100-, and 200-update ramps are 0/3, while Warmup500 remains more robust. At two lower-LR release states, parameter state and complete AdamW state each suffice for healthy continuation, whereas isolated moments do not. Release diagnostics show distinct organizations for healthy directional and scalar routes. These are tested-point results, not a universal warmup law or an early-commitment period.

1 INTRODUCTION

Warmup protects unstable training, but it also changes the path by which an optimizer reaches its later state. Can a brief intervention change long-run function even after the intervention is completely removed? We test this question in an unstable Post-LN language model. Paired runs share an exact update-3 parent, data order, and optimizer state; only updates 4–28 differ. All branches then resume the same native AdamW schedule through update 5000. We assess function using validation NLL, centered-token diversity D_{16}, and context benefit C.

Without warmup, this recipe typically reaches NLL near 6.22 with nearly zero D_{16} and no context benefit; a 500-update warmup remains healthy. Two shorter routes also produce persistent healthy continuations: a norm-preserving directional pulse, and a scalar ascending LR ramp over the same 25 updates. The ramp succeeds where exact-dose constant and reverse profiles fail. Release-shock, release-continuity, and scaled-ramp controls narrow the explanation to early-window organization among the tested profiles. Tests at eight layers and 107M parameters show that the same short schedule does not transfer uniformly; duration and recipe matter.

Our contributions are threefold. **C1: Persistent state selection and temporal controls.** Directional and scalar interventions can change update-5000 function after complete release; matched-dose and release-transition controls distinguish tested early-window profiles. **C2: Configuration dependence and state carriage.** Duration and LR-transfer tests map where short ramps work, while two lower-LR cut experiments identify redundant parameter-state and full-optimizer-state routes to healthy continuation. **C3: Distinct release-state organizations.** Pre-continuation diagnostics show that healthy directional and scalar routes need not share the same representation or optimizer-state geometry.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

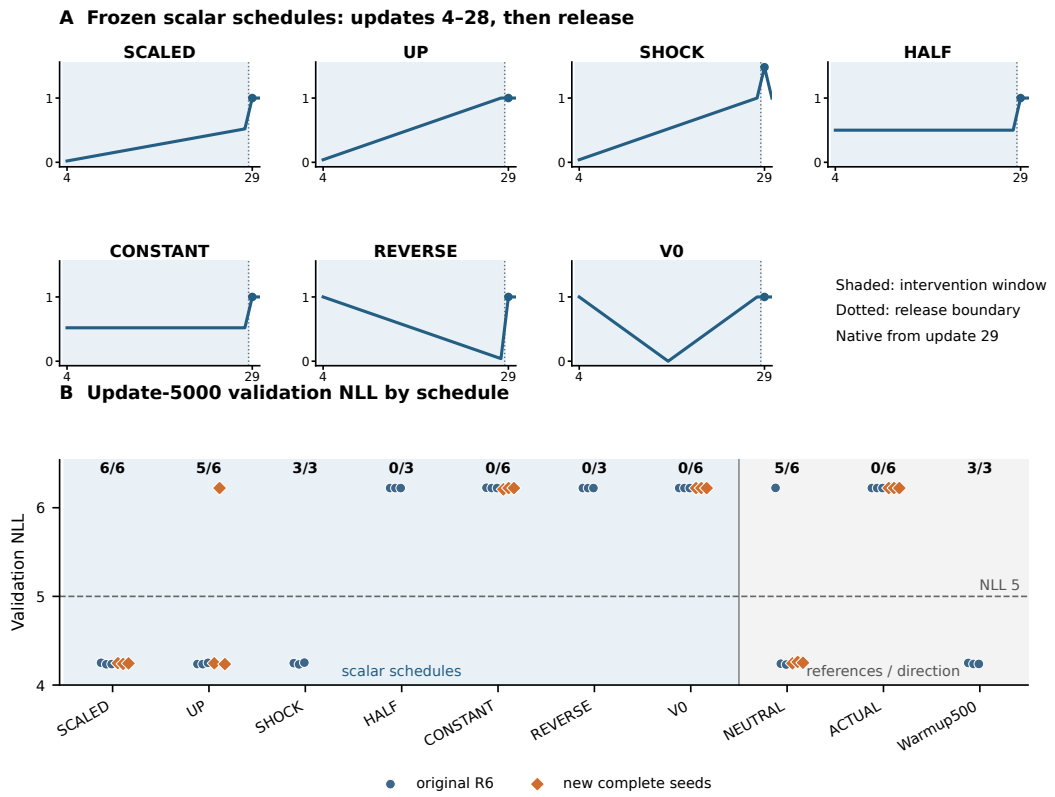

Figure 1: **Early-window scalar profiles and persistent outcomes.** (A) The seven tested scalar multipliers begin at their actual update-4 values, span updates 4–28, and show the release point at update 29; no full-LR update occurs at the left edge. (B) Update-5000 validation NLL for three original R6 seeds (blue circles) and complete new confirmation seeds (orange diamonds) where the same branch was run. Labels give healthy count/available seed count under the joint NLL/D_{16}/C criterion; the dashed line marks only its NLL component. The orange UP-RAMP failure is C8S2, the single ramp failure among the six complete seeds. ACTUAL, NEUTRAL, and Warmup500 are separate references, and Warmup500 is not a scalar-window branch.

2 Closest prior work and claim boundary

Temporary early perturbations can have persistent developmental consequences, and the earliest training phase can determine later optimization behavior (Achille et al., 2019; Frankle et al., 2020). Warmup has been linked to Post-LN gradient amplification (Xiong et al., 2020), curvature and stable learning rates (Kalra & Barkeshli, 2024), early Adam angular dynamics (Kosson et al., 2024), adaptive-rate variance (Liu et al., 2020a), and initialization-dependent difficulty (Liu et al., 2020b). Ma and Yarats motivate a practical warmup horizon $2/(1-\beta_2)$ from early adaptive-update magnitude; with fixed $\beta_2 = .98$ this is 100 updates (Ma & Yarats, 2021). Early catapult dynamics and update-scale proxies motivate separating transient geometry from eventual function (Lewkowycz et al., 2020; Wortsman et al., 2024). Most directly, Kosson et al. study finite early large-update treatment, its removal, long-lived damage, random-gradient controls, and stable rank. Our overlap is the persistence question; our distinct test is whether directional and temporally structured scalar treatments rescue functional collapse in a Post-LN language model, with matched-dose controls and a duration/recipe map. Standard Transformer and AdamW references define the baseline (Vaswani et al., 2017; Loshchilov & Hutter, 2019); the corpus follows WikiText practice (Merity et al., 2017). Architecture-level and online stabilization work provides adjacent remedies rather than the same finite-release comparison (Qi et al., 2025; Ramasinghe et al., 2026).

Xu establishes first-moment sufficiency for behavioral trait transfer using optimizer-state component surgery (Xu, 2026); other concurrent work also examines cut-equivalent interventions and delayed

effects (Guo, 2026; Liu & Li, 2026). Our lower-LR continuation tests ask a different question: whether parameter state, complete AdamW state, or isolated moments carry rescue from functional collapse. Parameter state and complete optimizer state are each sufficient at two tested cuts, while isolated m and v are not. This is a setting-dependent contrast, not a contradiction of trait transfer. Analyses of optimizer memory further motivate distinguishing a direct moment coefficient from its trajectory effects (Cattaneo & Shigida, 2025; Huang et al., 2025).

Attention and representation rank can collapse (Dong et al., 2021; Noci et al., 2022; Zhai et al., 2023; Wang et al., 2026). Our D, attention entropy, centered-hidden spectral concentration, and effective rank are not interchangeable with algebraic rank, Q/K update spectra, or every prior aggregation. We therefore use them only as pre-specified descriptive diagnostics. Depth-transfer heuristics motivate the separately pre-specified R8-DA LR recipe, but the relevant transfer-law assumptions do not hold exactly here (Wu et al., 2026; Mlodozeniec et al., 2025); the test is a discrete recipe comparison, not validation of a scaling law.

3 Experimental design

We name experimental cells by phase: R5 is the development pilot and R6 the prospective 24M/6L replication at peak 6×10^{-4}; R9 uses peak 4.5×10^{-4}; R8 and R8-DA use eight layers at peaks 6×10^{-4} and 3.897×10^{-4}, respectively; C8 denotes the independent-seed confirmation extension. Seed identifiers append S1, S2, ...

Let θ_t be parameters and (m_t, v_t) the named AdamW state. Paired branches share initialization, mini-batch order, evaluation data, scheduler phase, RNG state, and an exact update-3 parent. NEUTRAL modifies the native AdamW proposal to preserve familywise update norms and two half-minibatch task-gradient projections while imposing zero first-order change in $\Phi = \log(D + 10^{-12})$. ANGLE is a distance-matched directional control without that Φ constraint. Exact construction and feasibility rules are in the supplement. No directional correction occurs after update 28.

The independent unit is the master seed. We report raw paired outcomes for three-seed transfer cells without fitting a threshold or testing significance. The six-seed original-plus-new confirmation set uses descriptive exact paired sign tests and exact binomial intervals; only seeds with all six terminal branches enter. The operational healthy endpoint is the conjunction $\mathrm{NLL} < 5 \wedge D_{16} > 0.5 \wedge C > 0.5$. Its complete numeric form was documented after some earlier phases but before the new confirmation runs; retrospective labels on earlier endpoints are insensitive over broad threshold intervals given in the supplement. Construction tolerances, RNG handling, and transplant details are also there.

Training uses 15,906 sequences of length 129. Five thousand batches of 32 expose 20,480,000 target-token draws (about 10.06 row-equivalent passes) under a seeded permutation sampler that discards only a final partial batch. Standard validation NLL covers all 1,658 validation sequences and 212,224 targets; D_{16} uses rows 0–15 (2,048 hidden-token positions); C compares one target at absolute position 128 on rows 0–63 under one-token versus 128-token context. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.98)$, $\epsilon = 10^{-8}$, weight decay 0.01, dropout 0.1, normal initialization with standard deviation 0.02, cosine decay to 1% of peak, and no gradient clipping except in the explicit ACTUAL+clip1 control. Pulse branches have no warmup; Warmup500 uses a 500-update linear warmup. Runs use float32 on Apple MPS. Pairing means shared stored state, data order, RNG state, and schedule; it does not claim bitwise-deterministic MPS recomputation.

For scalar-window controls, ℓ_t is the frozen native LR. The ascending multiplier is $k/25$ for $k = 1, \ldots, 25$ at updates 4–28. Its exact weighted dose is $\sum_t \ell_t m_t = 0.007799697439119$. The reverse multiplier is $0.99997654995149(26 - k)/25$, and the constant multiplier is 0.519993902915899; both match that dose to floating precision. Update 29 returns to the native schedule.

Final six-seed confirmation counts are reported below and in Figure 1.

4 C1: Controlled Persistence

In the development pilot, NEUTRAL ends at NLL 4.5763, 4.2688, and 4.2225 rather than the approximately 6.22 failed endpoint. The prospective R6 replication is mixed: two seeds end at 4.2407 and 4.2311 with $D_{16} \approx 0.86$ and $C \approx 1.8$, while R6S1 ends at $(6.2239, 1.99 \times 10^{-7}, -1.86 \times$

Table 1: Scalar-window comparison on the three original R6 master seeds. Sustained release ratio is the native multiplier after release divided by the update-28 multiplier; the SHOCK row separately has a transient update-29 spike. NEUTRAL uses native scalar LR but changes direction, so its LR dose is not directional path length. V0 is approximate-dose.

intervention	first multiplier	sustained release ratio	transient one-step spike	dose/native	healthy
SCALED-RAMP	0.0207998	1.9231	none	0.270394	3/3
UP-RAMP	0.04	1	none	0.519994	3/3
ASCENDING-SHOCK	0.04	1	1.48 at $t = 29$	0.519994	3/3
HALF-LR	0.5	2	none	0.500000	0/3
CONSTANT	0.519994	1.9231	none	0.519994	0/3
REVERSE	0.999977	25.001	none	0.519994	0/3
V0	1	1	none	0.519999	0/3
NEUTRAL	1	1	none	1	2/3

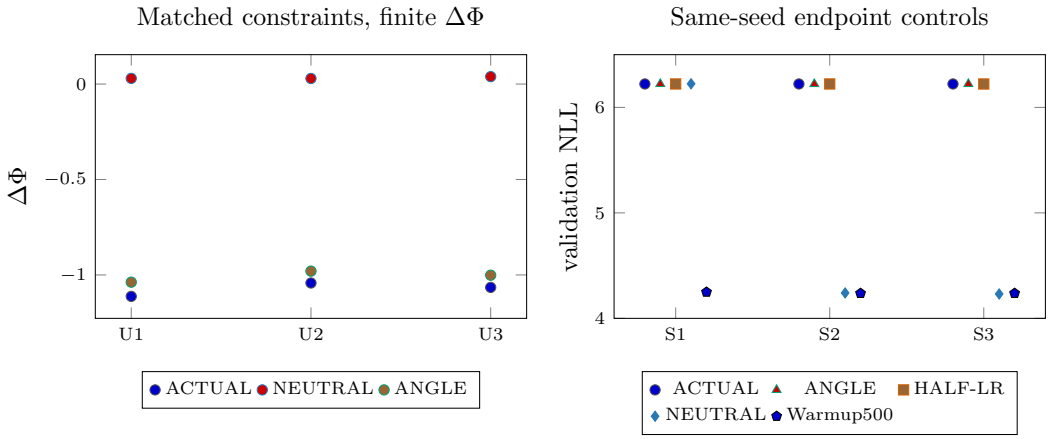

Figure 2: **Local controls.** Matched task projections and update norms do not make all directions equivalent. ANGLE and a same-window half-LR pulse do not reproduce persistent escape.

10^{-8}). At updates 28 and 100, R6S1 reaches only $(\mathrm{NLL}, D_{16}, C) = (6.2668, 0.0706, 0.0029)$ and $(6.2330, 0.0780, 0.0381)$. It therefore shows transient release-state/geometric response without functional escape, not a late re-collapse. Warmup500 is healthy on all three seeds.

ANGLE and the matched-window HALF-LR-PULSE both fail on all three R6 seeds. ANGLE excludes the tested structured matched-angle alternative, and the matched-window half-LR pulse excludes that tested scalar step-size reduction. Arbitrary or random perturbations are not ruled out. These controls do not identify D as the causal mediator: higher-order and unmeasured changes co-vary with the pulse.

An ascending LR ramp confined to updates 4–28 is healthy on all three original R6 seeds (NLL 4.2355–4.2503), so the directional NEUTRAL pulse is not the only route. Reverse and constant profiles match the ramp's weighted LR dose to floating precision, yet both fail on 3/3 seeds near NLL 6.22. The successful ramp remains healthy after a transient multiplier 1.48 shock at update 29 (3/3). A release shock of this tested magnitude therefore does not explain the contrast.

Two further controls separate release behavior from within-window shape. V0 starts at native LR and returns continuously to it, with a weighted dose within 7×10^{-8} of the ascending ramp's, but fails on 3/3 seeds; it is an approximate-, not exact-dose, control. SCALED-RAMP uses 0.5199939029158989 times the ascending profile and consequently has the same approximately $1.9231\times$ sustained release ratio as CONSTANT. It remains healthy on 3/3 original R6 seeds while CONSTANT fails. Thus the tested contrast is not explained by release continuity or sustained release-step magnitude alone. The exact-dose V construction lies outside its admissible multiplier bounds and has no endpoint. ACTUAL with global gradient-norm clipping at 1.0 also fails on 3/3.

A prospective same-parent repetition of the 25-update ramp is healthy on 3/3 and matches each historical healthy label, with absolute terminal-NLL differences below 7.1×10^{-4}. At release, RAMP raises D and lowers normalized attention entropy relative to ACTUAL in all three seeds, yet these changes are smaller than NEUTRAL's. For R6S1, healthy RAMP has $(D, H) = (0.002950, 0.993658)$ while failed NEUTRAL has $(0.038285, 0.965142)$. Neither diagnostic alone predicts the eventual outcome.

Independent-seed confirmation. Six master seeds completed all six branches: the three original R6 seeds and three new seeds. SCALED-RAMP is healthy on 6/6 and CONSTANT on 0/6 (3/3 versus 0/3 on the new seeds alone); UP-RAMP and NEUTRAL are each healthy on 5/6; V0 and ACTUAL are healthy on 0/6. The two routes fail on different seeds: R6S1 is rescued by the ramp but not by NEUTRAL, whereas C8S2 is rescued by NEUTRAL but not by the ramp. SCALED-RAMP rescues both. Because the original R6 outcomes informed subsequent control design, the pooled exact sign test for SCALED-RAMP versus CONSTANT (6 versus 0 discordant pairs, two-sided $p = 0.03125$) is descriptive; exact binomial intervals and stopping details are reported in the supplement.

5 C2: Recipe-Dependent Transfer and State Carriage

5.1 Recipe/LR-Dependent Transfer

Original R8 applies the frozen six-layer recipe to eight layers. NEUTRAL fails on all three seeds (NLL 6.2424, 6.2225, 6.2216), while Warmup500 remains healthy. This is a frozen-recipe transfer failure. R8-DA changes exactly one item: the peak LR becomes $6 \times 10^{-4}(8/6)^{-3/2} = 3.897114317 \times 10^{-4}$. No intervention, seed, or analysis definition changes.

All 12 R8-DA branches reach update 5000 with zero replay. NEUTRAL is healthy on 3/3 seeds. Two adjusted-recipe controls also show partial diagnostic movement: R8DA1 ACTUAL reaches $(\mathrm{NLL}, D_{16}, C) = (6.199684, 0.229638, 0.015668)$ and R8DA3 ANGLE reaches $(6.176960, 0.273881, 1.005483)$. Neither reaches the frozen joint healthy endpoint. Thus the adjusted recipe changes more than a binary NEUTRAL/no-NEUTRAL response, while full functional recovery remains specific to NEUTRAL in the tested R8-DA seeds.

The final same-seed middle-LR cell copies peak LR 4.5×10^{-4} from the tested six-layer lower-LR recipe. ACTUAL fails on 3/3 seeds; NEUTRAL is healthy on two seeds (NLL 4.221114 and 4.222458) and fails on the third (6.222206); Warmup500 is healthy on 3/3. This pre-specified mixed case supports recipe sensitivity and rules out a simple universal lower-LR rescue. Across the three tested eight-layer recipes, NEUTRAL is healthy on 0/3, 2/3, and 3/3 seeds at 6×10^{-4}, 4.5×10^{-4}, and $3.897114317 \times 10^{-4}$, respectively. These are discrete whole-schedule changes: they neither identify a threshold nor isolate pulse compatibility from the continuation environment, and the depth-derived value is not established as uniquely special.

The ascending 25-update ramp is 0/3 at the original eight-layer recipe. Moving its start from update 4 to update 1 is also 0/3, so the first three native updates alone do not explain failure. At the same recipe, a 100-update ramp is healthy on 3/3 and the pre-registered 50-update stage is mixed (2/3). The tested boundary is therefore seed-dependent at 50 updates: 25 fails, 50 is mixed, and 100 succeeds, without defining a mathematical minimum. At the lower-LR R8-DA recipe, the 25-update ramp is healthy on 3/3. Across all four tested cells in which both routes were evaluated, the directional intervention and RAMP25 agree at the cell level: both succeed at 24M/6L/6×10^{-4} and 30M/8L/3.897×10^{-4}, and both fail at 30M/8L/6×10^{-4} and 107M/6L/6×10^{-4}. Seedwise, however, the routes are not nested: R6S1 is rescued by the ramp but not NEUTRAL, while C8S2 shows the opposite dissociation.

At 107M parameters, ACTUAL and NEUTRAL fail on 3/3 seeds under each of three tested whole-schedule peak LRs (6, 4.5, and 3×10^{-4}), while Warmup500 is healthy on 3/3 in every cell. At the original 6×10^{-4} recipe, the tested 25-, 100-, and 200-update ascending ramps also fail on 3/3 completed seeds; the 200-update endpoints have NLL 6.2298–6.2311, $D_{16} < 9 \times 10^{-8}$, and $C \approx 0$. A separate update-1-started 25-update ramp has only two valid failed endpoints and is not counted as a three-seed result. The tested durations show no rescue at this 107M recipe, not a universal minimum. Every scale receives approximately 20.48M sampled target-token draws, so model width is confounded with the model/data/optimization-budget ratio. Execution exclusions and exact-resume checks are detailed in the supplement.

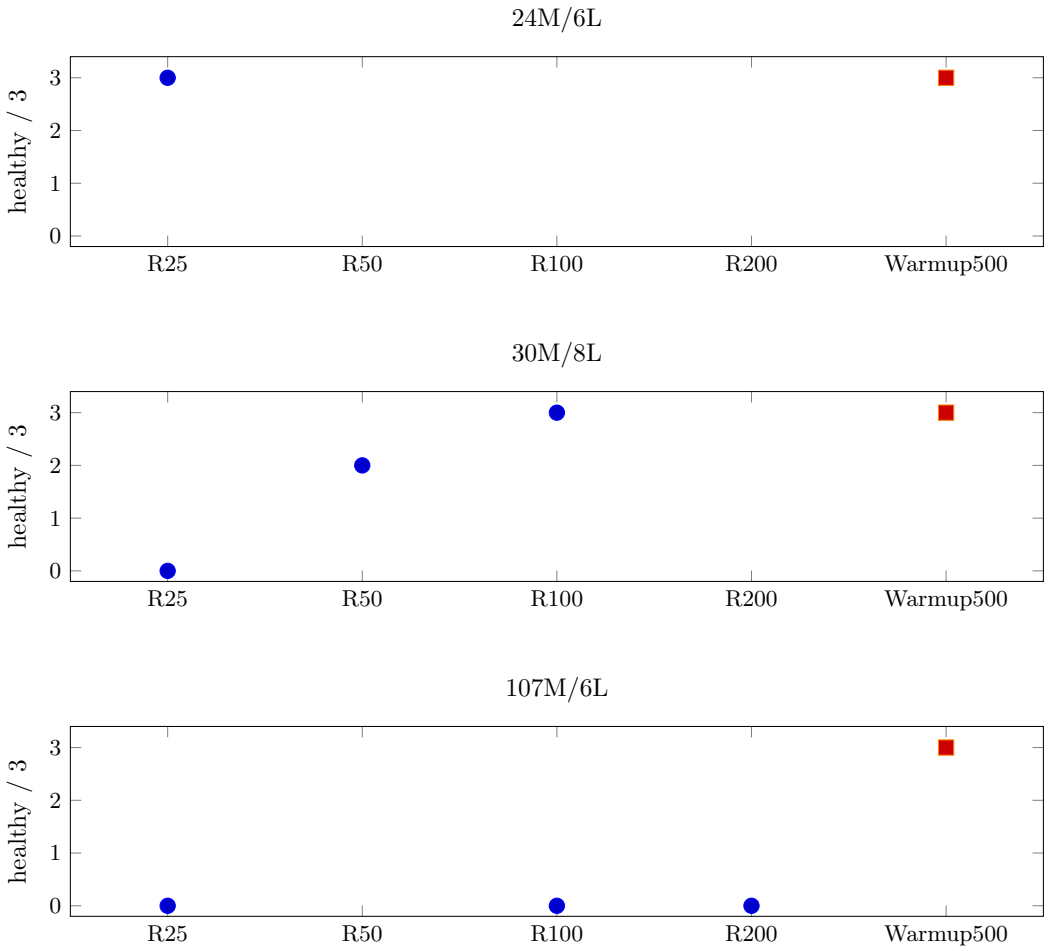

Figure 3: **Tested ramp durations across recipes.** Blue circular markers show completed multiplier-ramp cells; square markers show the distinct Warmup500 reference. Blank positions are untested, and no duration threshold is interpolated. The eight-layer LR-transfer figure is in the supplement.

Table 2: Core cross-recipe tested-point summary. Cells are healthy seeds / tested seeds; – means untested. The full control matrix is in the supplement.

recipe	ACTUAL	NEUTRAL	RAMP25	Warmup500
24M 6L, peak 6×10^{-4}	0/3	2/3	3/3	3/3
24M 6L, peak 4.5×10^{-4}	0/3	3/3	–	3/3
30M 8L, peak 6×10^{-4}	0/3	0/3	0/3	3/3
30M 8L, peak 4.5×10^{-4}	0/3	2/3	–	3/3
30M 8L, peak $3.897114317 \times 10^{-4}$	0/3	3/3	3/3	3/3
107M 6L, peak 6×10^{-4}	0/3	0/3	0/3	3/3
107M 6L, peak 4.5×10^{-4}	0/3	0/3	–	3/3
107M 6L, peak 3×10^{-4}	0/3	0/3	–	3/3

5.2 Joint-Moment State Carriage

At the exact update-28 cut, donor constructions have identical receiving parameters, logits, and loss before continuation. At high LR, the factorial has healthy NEUTRAL-parameter receiver corners NA

and NN, but ACTUAL parameters with full NEUTRAL moments (AN) fail; this does not support a general high-LR moment-carriage claim. R9S1 is the lower-LR decomposition, reconstructed from the frozen update-3 parent with every recorded batch, metric, state component, and cut-equivalence check verified against the transcript. Its complete four-corner result is AA failed (NLL 6.222094), AN healthy (4.243446), NA healthy (4.246761, $D_{16} = 0.854332$, $C = 1.853460$), and NN healthy (4.244900). Thus parameter state and complete optimizer state are redundant sufficient routes in this tested lower-LR release state. The moment decomposition then asks what is sufficient inside the successful AN route: corrected first moment only ends at 6.222021, while corrected second moment only reaches 6.116861 with descriptive partial movement; both fail the joint endpoint.

R9S2 is the independent prospective confirmation at a second qualified lower-LR state. Its completed NA corner is healthy (NLL 4.237875, $D_{16} = 0.849911$, $C = 1.829794$), as is AN (4.227257), whereas first-moment-only and second-moment-only end at 6.223699 and 6.151374. Thus both tested lower-LR states support redundant parameter/full-optimizer sufficiency, and inside the AN route neither isolated moment is sufficient. This recurrence does not imply quantitative invariance or algebraic necessity. The failed m-only branch has update-29 proposal norm 14.1993 versus 0.5897 for healthy full state, ruling out the monotonic explanation "larger first continued update implies healthier continuation." It does not distinguish informational insufficiency of m alone from destructive mismatch-induced overshoot when donor m is combined with receiver v.

In the failed ACTUAL state, attention moment scales are attenuated. Combining NEUTRAL's larger first moment with ACTUAL's smaller second-moment denominator yields large normalized proposals, consistent with a mismatch-induced overshoot hypothesis. The low-v attention coordinates are chiefly first-block Q/K weights, not a fixed key-bias or other structurally null slice; removing them barely changes the hybrid norm. These zero-update measurements do not establish that overshoot caused the m-only failure.

6 C3: Descriptive Release-State Formation

Release-state measurements help explain why one scalar diagnostic cannot stand in for the whole intervention. In the prospective R6 ramp replication, both successful tested routes raise D and lower normalized attention entropy relative to ACTUAL, but a failed NEUTRAL seed can show a larger change than a healthy RAMP seed. Effective rank moves differently: RAMP gives 52.2/28.4/29.9 across seeds, compared with ACTUAL's 30.4/14.9/20.9 and NEUTRAL's 2.15/1.31/2.33. Shared update-3 reference values appear in the supplement.

The native AdamW normalized-moment quantity $m/(\sqrt{v} + \epsilon)$ also differs by route at release. Across six parameter families, RAMP/ACTUAL norm ratios span 1.02–4.25, while RAMP/NEUTRAL spans 0.266–1.05 in POSITION, ATTENTION, and MLP and exceeds 1 in TOKEN_IO and LN_SCALE. Healthy endpoints can thus follow distinct measured release-state organizations. These zero-update comparisons are descriptive; they do not establish a mediator or a shared attractor.

For the three R6 NEUTRAL seeds from update 3 to 28, ΔD is $+0.020055$, $+0.316399$, and $+0.312499$, while normalized-attention-entropy changes are -0.026744, -0.176121, and -0.302753. Spectral concentration and effective rank remain comparatively similar. These exact values are an analysis finding, not a predictor. A frozen literature-inspired audit also finds descriptive separation in token similarity, pre-normalization residual norms, and layerwise gradient contraction; normalization and measured-object mismatches prohibit treating it as mechanism validation.

In both mixed cells, the failing seed happened to have the largest realized correction count; we treat this only as a post-hoc observation.

7 Limitations and conclusion

The independent-seed confirmation contains six complete master seeds (three original and three new); two planned new seeds were not completed, and transfer cells remain three-seed tested points. The study is restricted to Post-LN language modeling on WikiText-2 with float32 Apple MPS execution; bitwise-deterministic recomputation is not claimed. Wider models receive the same token-draw budget, confounding width with data and optimization budget, and at 107M the multiplier-

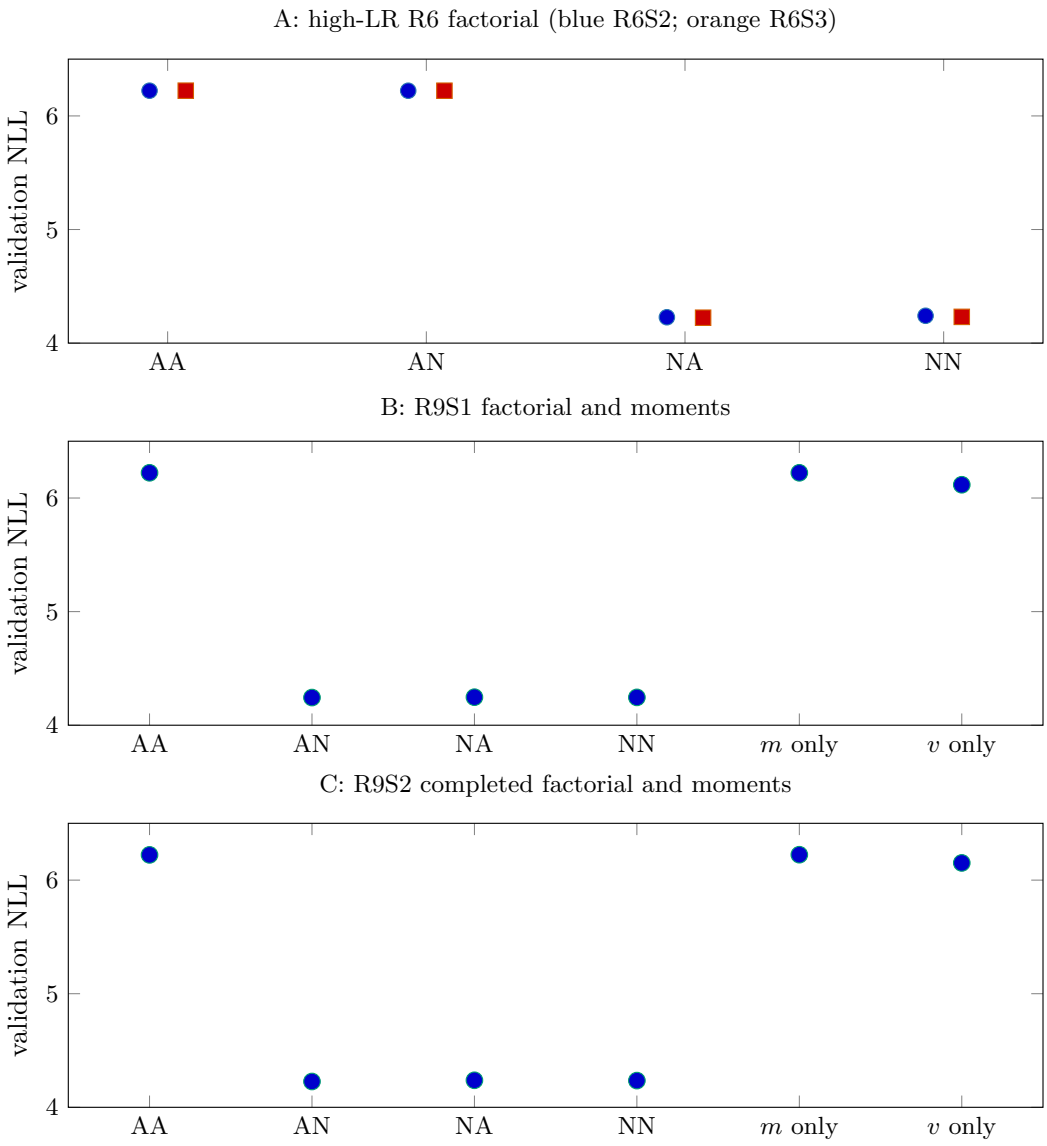

Figure 4: **State-carriage interventions.** Categorical points show the two qualified high-LR R6 seeds separately and the complete lower-LR R9S1/R9S2 factorial and moment conditions. No connecting line or ordinal path is implied.

defined 500-update ramp was not run. V0 matches the ascending weighted dose only approximately (relative difference about 9×10^{-6}). The isotropic/random-direction control produced no formal endpoint, so arbitrary perturbations are not excluded. Release-state and optimizer measurements remain descriptive, and isolated-m failure does not distinguish informational insufficiency from mismatch-induced overshoot. Warmup500 is more robust than either brief treatment across the tested recipes.

Untested hypothesis. Across the tested cells, rescue is more reliable when the earliest updates are smaller. At 24M, SCALED-RAMP is healthy on 6/6 versus 5/6 for the unscaled ramp and rescues the one seed the latter misses. At eight layers, longer ramps reduce the earliest multipliers and move outcomes from 0/3 at 25 updates to 2/3 at 50 and 3/3 at 100, while lowering the whole-schedule peak LR also restores RAMP25. One possible unifying interpretation is that larger architectures require stronger early-step reduction rather than longer intervention duration per se. This has not

been tested directly; for example, a 0.25-scaled 25-update ramp at the original eight-layer recipe would distinguish these explanations.

Brief early treatments can select long-run function after complete release in this unstable Post-LN regime, through at least two routes that reach different release states and rescue different seeds. Matched-dose and release-transition controls show that within-window temporal organization, rather than integrated dose or release transition alone, separates the tested scalar outcomes. Depth, width, LR, and duration tests bound where a 25-update treatment suffices. The persistent object carried into later ordinary optimization is therefore better described as the joint release state than by any single scalar diagnostic.

REPRODUCIBILITY STATEMENT

Supplement Sections 5–8 specify the NEUTRAL and ANGLE constructions, tolerances, RNG handling, checkpoint/resume rules, and state-transplant procedure; Section 21 lists the implementation state that must be frozen, and the supplement reports all endpoint values used in the paper. Runs used float32 PyTorch 2.13.0 on one Apple M4 Pro (24 GB) through MPS; checkpoints are omitted because of size and their recorded hashes are provided where available. An anonymized supplementary code archive is included.

AI USE STATEMENT

Generative AI assisted literature discovery, hypothesis and experiment-design feedback, method implementation and debugging, derivations, figure code, result interpretation, drafting, editing, and adversarial review. The authors checked AI-assisted code and scientific claims against the recorded protocols, source data, and execution logs. AI output was not treated as scientific ground truth; the authors take responsibility for the final text, claims, code, and artifacts.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019.

Matias Cattaneo and Boris Shigida. How memory in optimization algorithms implicitly modifies the loss. In *NeurIPS*, 2025. doi: 10.52202/085713-5215.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2021.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Hkl1iRNFwS`.

Jinhui Guo. Delayed optimizer-state transport shapes short-horizon training decisions. *arXiv preprint arXiv:2608.24593*, 2026.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. Spam: Spike-aware adam with momentum reset for stable llm training. In *ICLR*, 2025.

Dayal Kalra and Maissam Barkeshli. Why warmup the learning rate? underlying mechanisms and improvements. In *NeurIPS*, 2024.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Analyzing and reducing the need for learning rate warmup in gpt training. In *NeurIPS*, 2024.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Kang Liu and Suyan Li. Finite-horizon input-output dynamics of minibatch perturbations in adamw. *arXiv preprint arXiv:2608.19762*, 2026.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020a.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5747–5763. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.emnlp-main.463.

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

Jerry Ma and Denis Yarats. On the adequacy of untuned warmup for adaptive optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8828–8836, 2021. doi: 10.1609/aaai.v35i10.17069.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *International Conference on Learning Representations*, 2017.

Bruno Mlodozeniec, Pierre Ablin, Louis Béthune, Dan Busbridge, Michal Klein, Jason Ramapuram, and Marco Cuturi. Completed hyperparameter transfer across modules, width, depth, batch and duration. *arXiv preprint arXiv:2512.22382*, 2025.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *NeurIPS*, 2022.

Xianbiao Qi, Yelin He, Jiaquan Ye, et al. Taming transformer without using learning rate warmup. In *ICLR*, 2025.

Sameera Ramasinghe, Thalaiyasingam Ajanthan, Hadi Mohaghegh Dolatabadi, et al. Taming curvature: Architecture warm-up for stable transformer training. *arXiv preprint arXiv:2606.16768*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Xingjian Wang, Qingyu Han, Xiaodong Luo, and Yin Zhang. Why post-norm transformers collapse: Attention amplification and gradient repair failure. *arXiv preprint arXiv:2608.09417*, 2026.

Mitchell Wortsman et al. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations*, 2024.

Shenxi Wu, Haosong Zhang, Xingjian Ma, Shirui Bian, Yichi Zhang, Xi Chen, and Wei Lin. Hyperparameter transfer laws for non-recurrent multi-path neural networks. *arXiv preprint arXiv:2602.07494*, 2026.

Ruibin Xiong, Yunchang Yang, Di He, et al. On layer normalization in the transformer architecture. In *ICML*, 2020.

Qinyang Xu. Stored in optimizer state, valued by later training: A causal account of subliminal trait transfer. *arXiv preprint arXiv:2608.20442*, 2026.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *ICML*, 2023.

SUPPLEMENT: BRIEF TREATMENTS SELECT LONG-RUN FUNCTION

Anonymous authors
Paper under double-blind review

1 SCOPE AND FROZEN EVIDENCE HIERARCHY

This supplement specifies the interventions and state transplants for independent implementation and exposes the endpoints used in the paper. The evidence order is recorded result tables and receipts, protocol and source manifests, generated tables, then narrative prose. Manuscript generation and the stated zero-update checks did not themselves train models. Development, prospective confirmation, transfer, and diagnostic cells are not pooled as exchangeable replications.

Table 1: Forty-one-row final evidence and dependency crosswalk. Linked rows are not pooled as exchangeable replications.

phase	question	source	role	supports/status
R4	development precursor	frozen records	development	early persistence
R5	pilot pulse	frozen endpoints	development	persistent selection
R6	prospective pulse	frozen endpoints	confirmation	mixed persistence
R7-R1	high-LR carrier	qualified R6 donors	mechanism	factorial contrast
R8	original 8L transfer	frozen endpoints	transfer	negative cell
R8-DA	adjusted 8L recipe	frozen endpoints	transfer	positive NEUTRAL cell
R8-LR45	middle 8L recipe	frozen endpoints	transfer	mixed NEUTRAL cell
R9	lower-LR states	frozen endpoints	selection	qualified donors
R10-R2	R9S1 reconstruction	verified transcript	carriage	verified continuation cut
R9S1 m/v	moment decomposition	cut-equivalent branches	mechanism	isolated moments fail
R9S2	prospective moments	qualified donor	confirmation	full moments healthy
R9S2 NA	missing corner completion	exact continuation	confirmation	NA healthy
P0 HALF-LR	matched-window scalar	frozen endpoints	control	0/3 healthy
P1-R3 107M 6e-4	original width cell	frozen endpoints	transfer	pulse negative
107M 4.5e-4	lower-LR width cell	exact branches	transfer	pulse negative
107M 3e-4	conditional width cell	exact branches	transfer	pulse negative
WINDOW-RAMP R6	ascending scalar window	exact branches	control	3/3 healthy
ACTUAL+clip1	global clipping	exact branches	control	0/3 healthy
8L WINDOW-RAMP	original R8 parents	exact branches	transfer	0/3 healthy
107M WINDOW-RAMP	original P1-R3 parents	exact branches	transfer	0/3 healthy
REVERSE-RAMP	exact dose match	exact branches	shape control	0/3 healthy
CONSTANT MATCHED	exact dose match	exact branches	shape control	0/3 healthy
P2-R1	R6 release diagnostics	resident t3/t28 states	descriptive	four frozen diagnostics
R8-DA external P2	8L release diagnostics	resident states	descriptive	external reuse
historical ramp release	original R6 t28 object	missing historical checkpoint	provenance	not substituted by prospective state
repair-signature	literature-inspired probes	zero-update records	descriptive	association only
ISO / ISO-R1	isotropic control	qualification only	incomplete	no formal endpoint
final width audit	parameterization scope	code/config audit	admissibility	no transfer-law validation
P1 ascending shock	release-shock control	3 exact branches	shape control	3/3 healthy
P1 V bridge	dose plus continuity	analytic solution	admissibility	b outside bounds
P2 ramp replication	prospective R6 ramp	3 exact branches	confirmation	3/3 reproduced
P2 ramp release	ACTUAL/NEUTRAL/RAMP t28	exact resident states	descriptive	four frozen diagnostics
P3 low-LR ramp	R8-DA route transfer	3 exact branches	transfer	3/3 healthy
P4 8L start	START1 RAMP25	3 exact branches	boundary	0/3 healthy
P4 8L length	RAMP100 then RAMP50	6 exact branches	duration	3/3 then 2/3
P4 107M start	START1 RAMP25	2 valid plus 1 invalid	incomplete	0/2 valid
P4 107M length	RAMP100 then RAMP200	6 exact; 2 invalid excluded	duration	0/3 then 0/3
P5 V0 control	approximate-dose continuous V	3 exact branches	shape control	0/3 healthy
P6 SCALED-RAMP	scaled ascending step-up	3 exact branches	shape control	3/3 healthy
C8 complete	independent-seed six-branch extension	keyed endpoints	confirmation	3/5 new complete; combined n=6
C8 excluded	incomplete concurrent attempt and unstarted seed	partial/unstarted records	incomplete	2 new seeds not pooled

2 DECISION RULE AND PROVENANCE

We use the strict conjunction $\mathrm{NLL} < 5 \wedge D_{16} > 0.5 \wedge C > 0.5$ as the frozen operational endpoint classifier for the R2 analyses and all newly executed confirmation cells. Earlier R5, R6, R7-R1, R8, and R9 phases predate the complete documented numeric form. The first located executable analysis script is dated 2026-09-20; the first located complete written statement is dated 2026-09-22. R9S2 and

the final-expansion cells inherit it prospectively. Applying the later rule to earlier frozen update-5000 endpoints is insensitive over the open componentwise intervals reported in the threshold-sensitivity audit; this robustness is not retrospective preregistration.

Table 2: Threshold sensitivity over 90 frozen update-5000 rows (39 healthy, 51 failed). Componentwise intervals preserve each metric's healthy/failed separation; joint-label intervals vary one cutoff with the other two fixed. All intervals are open.

metric	worst healthy	best failed	componentwise interval	joint-label interval
NLL ($<$)	4.576301	6.116861	$(4.576301, 6.116861)$	$(4.576301, \infty)$
D_{16} ($>$)	0.825412	0.437469	$(0.437469, 0.825412)$	$(-\infty, 0.825412)$
C ($>$)	1.192630	1.005483	$(1.005483, 1.192630)$	$(-\infty, 1.192630)$

Table 3: Realized NEUTRAL intervention dose. Updates 4–28 provide 25 eligible opportunities; identity/native and infeasible assignments are not corrected updates.

phase	unit	eligible	corrected	identity	infeasible	blocked
R5	P1	25	15	9	1	0
R5	P2	25	16	9	0	0
R5	P3	25	16	9	0	0
R6	R6S1	25	18	6	1	0
R6	R6S2	25	10	14	1	0
R6	R6S3	25	10	14	1	0
R8	R8S1	25	20	5	0	0
R8	R8S2	25	13	11	1	0
R8	R8S3	25	15	9	1	0
R9	R9S1	25	8	16	1	0
R9	R9S2	25	12	12	1	0
R9	R9S3	25	11	13	1	0
R8-DA	R8DA1	25	15	10	0	0
R8-DA	R8DA2	25	17	8	0	0
R8-DA	R8DA3	25	15	9	1	0
R8-LR45	S1	25	13	11	1	0
R8-LR45	S2	25	16	8	1	0
R8-LR45	S3	25	19	5	1	0

Table 4: Exact prospective NEUTRAL trajectory. R6S1 never satisfies the joint endpoint criterion.

seed	update	NLL	D_{16}	C	healthy
R6S1	28	6.266755	0.0705805	0.00288866	False
R6S1	100	6.232994	0.0779771	0.0381271	False
R6S1	500	6.241209	1.30341e-05	3.2261e-06	False
R6S1	1000	6.232646	1.27342e-07	-8.56817e-08	False
R6S1	2000	6.226583	8.66749e-08	-1.15484e-07	False
R6S1	5000	6.223885	1.98515e-07	-1.86265e-08	False
R6S2	28	6.185224	0.346818	0.0711884	False
R6S2	100	5.716943	0.722969	0.137164	False
R6S2	500	4.876540	0.792378	0.634475	True
R6S2	1000	4.553322	0.821119	0.970124	True
R6S2	2000	4.318256	0.837588	1.52059	True
R6S2	5000	4.240685	0.861327	1.88219	True
R6S3	28	6.113983	0.348816	-0.0589858	False
R6S3	100	5.672062	0.738384	0.0461663	False
R6S3	500	4.848174	0.793243	0.530805	True
R6S3	1000	4.537002	0.83186	1.2092	True
R6S3	2000	4.302938	0.827081	1.48553	True
R6S3	5000	4.231073	0.857906	1.80805	True

3 Model, data, and native update

Experiments use decoder-only Post-LN Transformers with tied token embedding and output weights. The frozen tokenizer/vocabulary, token arrays, split hashes, model hyperparameters, dropout, sequence length, batch size, AdamW hyperparameters, schedule, evaluation batch, and seed are recorded in each phase config and source manifest. A training batch has shape 32×129: rows 0–15 form half A, rows 16–31 half B, tokens 0–127 are inputs, and tokens 1–128 are targets. Cross-entropy uses `ignore_index=0` and mean reduction.

At optimizer update t, the cosine schedule is indexed as `get_lr(t-1,5000,peak,0)` unless the branch is Warmup500. Native PyTorch AdamW uses $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$, and weight decay 0.01. The native optimizer step occurs first on the current minibatch, producing θ_{nom} and native moment/step state. Let θ_{parent} denote the pre-step parameters and $v = \theta_{\mathrm{nom}} - \theta_{\mathrm{parent}}$. Any pulse overwrite changes parameters only; it retains the moments and optimizer step produced by the native update. Buffers are untouched.

Table 5: Exact training and evaluation scope. The deterministic sampler discards a final partial batch before starting the next seeded permutation.

object	sequence scope	evaluation rule	token scope
training arrays	15,906 rows	128 targets/row	2,035,968 targets
validation arrays	1,658 rows	128 targets/row	212,224 targets
training budget	5,000 updates	32 rows/update	20,480,000 sampled targets
standard NLL	1,658 rows	26 ordered batches	212,224 targets
D_16	rows 0–15	128 inputs/row	2,048 hidden-token positions
C	rows 0–63	one target/row	64 targets per context condition

The training array has shape $15{,}906 \times 129$ and the validation array $1{,}658 \times 129$; neither target array contains padding zeros. The deterministic sampler draws seeded permutations without within-permutation repeats, discards a final partial batch of fewer than 32 rows, then begins the next permutation. Standard validation uses 26 ordered batches and all 212,224 targets. Context benefit uses validation rows 0–63, target position 128, and frozen context spans $[127, 127]$ and $[0, 127]$. All reported training runs use float32 on Apple MPS; construction vectors use CPU float64. Linear and embedding weights are initialized normally with standard deviation 0.02; linear biases are zero; LayerNorm begins with unit scale and zero shift. Dropout is 0.1 and gradient clipping is disabled except in the explicit ACTUAL+clip1 control. The cosine endpoint is 1% of peak LR. Pairing means shared stored state, data order, schedule, and RNG state; deterministic algorithms were not recorded as enabled, so no bitwise-recomputation claim is made.

4 PARAMETER FAMILIES AND DIAGNOSTIC

The exact ordered families are TOKEN_IO, POSITION, ATTENTION, MLP, LN_SCALE, and LN_SHIFT. `token_emb.weight` is TOKEN_IO; the tied `lm_head.weight` is the same object and does not appear twice in named parameters. `pos_emb.weight` is POSITION; `blocks.*.attn.*` is ATTENTION; `blocks.*.ff.*` is MLP; block `ln1/ln2.weight` and `ln1/ln2.bias` are LN_SCALE and LN_SHIFT. There are no unmapped trainable parameters.

For last-block hidden states X, centered-token diversity is

$$D(X) = \frac{\|X - \mathrm{mean}_{\mathrm{token}}(X)\|_F}{\|X\|_F + 10^{-12}}, \qquad \Phi = \log(D + 10^{-12}).$$

During construction the model is in evaluation mode. Two forward passes and three `autograd.grad` calls produce half-batch task gradients h_A, h_B and $p_A = \nabla_\theta \Phi$ without populating optimizer gradients. Python, NumPy, CPU, and MPS RNG states and the previous train/eval mode are saved and restored. Construction vectors are converted to float64 on CPU.

5 NEUTRAL CONSTRUCTION

For groups g, NEUTRAL solves

$$\min_u \sum_g \|u_g - v_g\|_2^2$$

subject to $\|u_g\|_2 = \|v_g\|_2$ for every family, $h_A^\top u = h_A^\top v$, $h_B^\top u = h_B^\top v$, and $\sum_g p_{A,g}^\top u_g = 0$. The task subspace is formed from normalized h_A, h_B with two reorthogonalization passes; vectors below rank tolerance 10^{-10} are removed. The norm/projection relative tolerance is 10^{-5}; diversity-neutrality tolerance is $10^{-7} \max(\sum_g \|p_g\| \|v_g\|, 10^{-12})$. The h_B degeneracy, $p_\perp$ degeneracy, and denominator tolerances are 10^{-12}. Scalar clipping tolerance is 10^{-9}; feasibility bracketing permits 60 expansions and bisection 100 iterations.

If the native diversity projection already satisfies the frozen tolerance, NEUTRAL is the identity/native proposal. If the constrained target is analytically infeasible, it is also native. A numerically unresolved construction is a blocking error, never a silent native fallback. A feasible solution is cast to the parameter dtype and installed as $\theta_{\mathrm{parent}} + u$ after the native optimizer step.

6 ANGLE CONSTRUCTION

ANGLE first computes NEUTRAL. Within each family it preserves the native task-subspace component and diversity projection while replacing the remaining orthogonal direction. If z_N is NEUTRAL's complement and $v_\perp$ the native complement, it uses $k = \langle v_\perp, z_N \rangle / \|v_\perp\|^2$ to match the NEUTRAL-to-native distance subject to familywise norm and task constraints. A Gaussian direction is drawn from NumPy PCG64 initialized by `SeedSequence([master_seed,5005])`, projected off the task subspace and $v_\perp$, normalized, and rotated by the analytic coefficient. Random draws occur only for actual construction; exact resume advances by the logged draw count. The ANGLE norm and degeneracy tolerances are 10^{-12}, with 10^{-9} clipping. Its Φ response is intentionally unconstrained. If

the complement is degenerate, the frozen code follows the documented identity/native rule; numerical non-resolution blocks.

7 EXECUTABLE PSEUDOCODE

```
for update in 1..5000:
    batch = next(frozen_sampler)
    lr = cosine(update-1, total=5000, peak=branch_peak, warmup=0)
    theta_parent = clone(named_parameters)
    loss(batch).backward(); adamw.step(); adamw.zero_grad()
    theta_nom = clone(named_parameters)
    # Native moments and step now represent this exact minibatch.
    if 4 <= update <= 28 and branch in {NEUTRAL, ANGLE}:
        save(mode, python_rng, numpy_rng, cpu_rng, mps_rng)
        eval_mode(); hA, hB, pA = diagnostic_grads(batch)
        restore(mode, python_rng, numpy_rng, cpu_rng, mps_rng)
        v = float64_cpu(theta_nom - theta_parent)
        if branch == NEUTRAL:
            u = solve_neutral(v,hA,hB,pA,tolerances)
        else:
            u = solve_angle(v,hA,hB,pA,seed,drawing_log)
        assert frozen_constraints(u,v,hA,hB,pA)
        copy_parameters(theta_parent + cast_like(u))
        # Keep moments, step, scheduler, buffers, and sampler native.
    checkpoint(update, model, named_optimizer_state, scheduler,
               sampler, all_rng_states, construction_draw_count)
```

Resume accepts a checkpoint only after byte hash, recorded completed update, optimizer step, scheduler, sampler cursor, all RNG states, and construction draw count agree with the provenance manifest. It resumes at completed update plus one. Re-execution of a completed optimizer update is forbidden and replay count must remain zero.

8 STATE TRANSPLANTS

At update 28, A denotes ACTUAL and N denotes NEUTRAL. The receiver supplies model parameters and buffers; the optimizer donor supplies state matched by exact parameter name. The scheduler remains at step 28, continuation minibatches and RNG states are identical, and no optimizer update is used to verify the cut.

Condition	parameters/buffers	`exp_avg` m	`exp_avg_sq` v
AA	A	A	A
AN (full m+v)	A	N	N
NA	N	A	A
NN	N	N	N
m only	A	N	A
v only	A	A	N

The AdamW `step` is 28 and is checked equal across donors. Unrecognized, duplicated, or missing names block construction. Before continuation, exact receiving parameters, logits, and loss must match the designated receiver; transplanted moment tensors must hash to the designated donors. Buffers are copied from the parameter receiver. This makes AN a full corrected optimizer-state intervention and the two hybrids exact first- or second-moment interventions.

9 RAW R8-DA AND PAIRED R8 COMPARISON

Table 6: R8-DA raw update-5000 endpoints. All 12 branches completed with zero replay.

seed	branch	NLL	D_{16}	C
R8DA1	ACTUAL	6.199684	0.229638	0.015668
R8DA1	ANGLE	6.222149	6.20771e-08	-0.001265
R8DA1	NEUTRAL	4.221161	0.853284	2.064538
R8DA1	WARMUP500	4.212842	0.848077	1.886530
R8DA2	ACTUAL	6.222379	7.32137e-08	-0.000000
R8DA2	ANGLE	6.222455	7.39627e-08	0.000000
R8DA2	NEUTRAL	4.221482	0.853679	1.676797
R8DA2	WARMUP500	4.223559	0.853932	1.879583
R8DA3	ACTUAL	6.222059	7.95317e-08	-0.000000
R8DA3	ANGLE	6.176960	0.273881	1.005483
R8DA3	NEUTRAL	4.221333	0.853319	1.730516
R8DA3	WARMUP500	4.222146	0.850713	1.839137

Table 7: Exact same-seed R8 versus R8-DA comparison. Difference is adjusted minus original.

seed	branch	R8 NLL	R8-DA NLL	difference
R8S1	ACTUAL	6.222033	6.199684	-0.022349
R8S1	ANGLE	6.222060	6.222149	+0.000089
R8S1	NEUTRAL	6.242355	4.221161	-2.021194
R8S1	WARMUP500	4.229280	4.212842	-0.016438
R8S2	ACTUAL	6.222484	6.222379	-0.000105
R8S2	ANGLE	6.222299	6.222455	+0.000156
R8S2	NEUTRAL	6.222545	4.221482	-2.001064
R8S2	WARMUP500	4.230670	4.223559	-0.007111
R8S3	ACTUAL	6.221811	6.222059	+0.000248
R8S3	ANGLE	6.221925	6.176960	-0.044965
R8S3	NEUTRAL	6.221600	4.221333	-2.000268
R8S3	WARMUP500	4.229924	4.222146	-0.007778

All 12 R8-DA branches completed update 5000. Accounting reports 60,000 branch updates, zero replay, and verified endpoint checkpoints. The power-interrupted R8DA3-WARMUP500 branch was resumed only from the verified update-4200 checkpoint, at update 4201, with optimizer/scheduler/sampler/RNG state restored and no completed update re-executed. NEUTRAL is healthy on all three seeds. The three exact NLL interaction contrasts are -1.998845, -2.000959, and -2.000515.

R8DA1 ACTUAL has NLL 6.199684, $D_{16} = 0.229638$, and $C = 0.015668$; R8DA3 ANGLE has NLL 6.176960, $D_{16} = 0.273881$, and $C = 1.005483$. The source CSV and endpoint logs agree. R8DA3 ANGLE's resident terminal checkpoint matches its creation hash; the R8DA1 ACTUAL

checkpoint was not locally available, so its bytes and live hash could not be re-verified and only its creation receipt is reported. No recorded execution or provenance anomaly explains either endpoint. This partial diagnostic movement is a frozen observation, but neither branch is healthy under the joint criterion.

10 FINAL R8-LR45 PROTOCOL AND THREE-RECIPE COMPARISON

The final cell retains exact R8 architecture, data, optimizer, schedule form, update budget, and master seeds, changing only peak LR from 6×10^{-4} to 4.5×10^{-4}. ACTUAL and NEUTRAL share a newly constructed t3 parent under the new recipe; WARMUP500 follows its canonical route. Exactly nine branches were run, with 45,000 represented branch updates, 44,991 unique optimizer calls due to shared prefixes, 64 route-comparison optimizer calls, and zero replay.

Table 8: R8-LR45 raw endpoints. All nine branches completed with zero replay.

seed	branch	NLL	D_{16}	C	healthy
R8LR451	ACTUAL	6.222072	8.15575e-08	-0.000000	false
R8LR451	NEUTRAL	4.221114	0.856073	1.730746	true
R8LR451	WARMUP500	4.212642	0.852257	1.847580	true
R8LR452	ACTUAL	6.222270	6.85971e-08	-0.000000	false
R8LR452	NEUTRAL	4.222458	0.854857	1.964229	true
R8LR452	WARMUP500	4.221931	0.858717	1.822402	true
R8LR453	ACTUAL	6.222172	8.76335e-08	0.000000	false
R8LR453	NEUTRAL	6.222206	3.18259e-07	0.000000	false
R8LR453	WARMUP500	4.216716	0.852722	1.685336	true

Table 9: Same-seed eight-layer NLL at the three discrete tested LR recipes; no interpolation or threshold fit.

seed	branch	6e-4	4.5e-4	3.897e-4
2026092101	ACTUAL	6.222033	6.222072	6.199684
2026092101	NEUTRAL	6.242355	4.221114	4.221161
2026092101	WARMUP500	4.229280	4.212642	4.212842
2026092102	ACTUAL	6.222484	6.222270	6.222379
2026092102	NEUTRAL	6.222545	4.222458	4.221482
2026092102	WARMUP500	4.230670	4.221931	4.223559
2026092103	ACTUAL	6.221811	6.222172	6.222059
2026092103	NEUTRAL	6.221600	6.222206	4.221333
2026092103	WARMUP500	4.229924	4.216716	4.222146

The pre-specified outcome is mixed: ACTUAL fails 3/3, NEUTRAL is healthy 2/3, and Warmup500 is healthy 3/3. The exact interpretation is tested-point recipe sensitivity. No threshold, interpolation, phase boundary, curve, or scaling law is estimated.

11 FINAL 107M SEQUENTIAL LR PROBE

The 107M architecture has six layers, width 1152, 18 heads, feed-forward width 4608, vocabulary 10,000, and tied TOKEN_IO. The original 6×10^{-4} recipe was frozen before the final expansion.

The first new cell changed only the whole-schedule peak LR to 4.5×10^{-4}. Because ACTUAL failed 3/3 and NEUTRAL was healthy on 0/3, the pre-specified gate authorized exactly one further cell at 3×10^{-4}, half the original peak. No other LR was run.

Table 10: Final 107M lower-LR endpoints. Every cell has zero replay; both tested lower-LR cells retain failed ACTUAL/NEUTRAL and healthy Warmup500.

study	seed	branch	NLL	D_{16}	C	healthy
4.5e-4	P1S1	ACTUAL	6.226369	7.80975e-08	-0.000000	false
4.5e-4	P1S1	NEUTRAL	6.226287	9.16608e-08	0.000000	false
4.5e-4	P1S1	WARMUP500	4.374907	0.876806	1.930463	true
4.5e-4	P1S2	ACTUAL	6.227020	8.65124e-08	-0.000000	false
4.5e-4	P1S2	NEUTRAL	6.225877	9.06526e-08	0.000000	false
4.5e-4	P1S2	WARMUP500	4.371767	0.876016	1.667557	true
4.5e-4	P1S3	ACTUAL	6.225662	8.45958e-08	-0.000000	false
4.5e-4	P1S3	NEUTRAL	6.225801	9.93971e-08	0.000000	false
4.5e-4	P1S3	WARMUP500	4.377445	0.876086	1.640323	true
3e-4	P1S1	ACTUAL	6.222036	1.09272e-07	-0.000000	false
3e-4	P1S1	NEUTRAL	6.221899	1.03502e-07	0.000000	false
3e-4	P1S1	WARMUP500	4.376619	0.881257	2.012225	true
3e-4	P1S2	ACTUAL	6.223395	1.03795e-07	0.000000	false
3e-4	P1S2	NEUTRAL	6.223759	0.0878505	0.024207	false
3e-4	P1S2	WARMUP500	4.390716	0.879109	1.955721	true
3e-4	P1S3	ACTUAL	6.222984	1.0398e-07	0.000000	false
3e-4	P1S3	NEUTRAL	6.224228	1.13994e-07	0.000000	false
3e-4	P1S3	WARMUP500	4.393425	0.874848	1.644372	true

At both lower LRs, ACTUAL and NEUTRAL fail on all three seeds, while Warmup500 is healthy on all three. The original recipe has the same healthy-count pattern. The result is therefore negative for NEUTRAL recovery across these three exact recipes but positive for architecture trainability under Warmup500. It estimates neither an optimal LR nor a width law. A power interruption in the 3×10^{-4} cell occurred after update 1950; resume began at update 1951 with zero replay. Accounting corrections removed a double-counted projection and aggregate count, yielding exactly 44,991 unique formal updates plus 64 qualification updates.

The earlier admissibility audit remains relevant only to claim scope: WARM uses standard parameterization rather than μP, and the tied token embedding/output tensor prevents direct adoption of prescriptions assigning separate input/output roles without another modeling intervention. The empirical probes are discrete recipe tests, not validation of those transfer laws.

12 Final R6 controls

12.1 Deadline-bounded independent-seed confirmation

The C8 extension was initially planned for five new master seeds. During execution, after C8 results had begun to accumulate, the extension was stopped under the submission-time cutoff and execution-throughput constraints after three new seeds had completed all six branches. C8S4 was launched only as a concurrent-throughput attempt; concurrent execution reduced throughput and the attempt ended without a valid terminal endpoint. Its partial progress was visible operationally but no terminal classification was included. C8S5 was not started. Neither seed enters any denominator. Consequently, pooled exact tests over the six complete seeds are reported descriptively, not as fully pre-data confirmatory inference. The primary SCALED-RAMP versus CONSTANT contrast has 6 SCALED-only and 0 CONSTANT-only healthy pairs among 6 independent master seeds (exact two-sided sign $p = 0.03125$). The 95% Clopper–Pearson interval for SCALED win probability

Table 11: Completed six-seed confirmation. Only complete six-branch master seeds enter the denominator; 95% intervals are exact Clopper–Pearson intervals.

Branch	Healthy / n	95% interval	New healthy / complete new
ACTUAL	0/6	[0.000,0.459]	0/3
UP-RAMP	5/6	[0.359,0.996]	2/3
CONSTANT	0/6	[0.000,0.459]	0/3
SCALED-RAMP	6/6	[0.541,1.000]	3/3
V0	0/6	[0.000,0.459]	0/3
NEUTRAL	5/6	[0.359,0.996]	3/3

Table 12: Descriptive exact paired contrasts over the same six complete seeds. Counts give healthy only under the first or second named branch; tests use discordant pairs.

Contrast	First only	Second only	two-sided p
SCALED-RAMP vs CONSTANT	6	0	0.03125
UP-RAMP vs CONSTANT	5	0	0.0625
UP-RAMP vs V0	5	0	0.0625
NEUTRAL vs ACTUAL	5	0	0.0625

among discordant pairs is [0.541, 1.000]. The original R6 outcomes informed control design, so the pooled sign-test value is descriptive rather than a fully pre-data discovery probability.

Table 13: Complete original and prospective confirmation endpoints at update 5000. H is the frozen conjunction NLL < 5, $D_{16} > 0.5$, $C > 0.5$.

Seed	Branch	NLL	D_{16}	C	H
R6S1	ACTUAL	6.2222	8.75e-08	-0.000177	N
R6S1	UP-RAMP	4.2368	0.861	1.89	Y
R6S1	CONSTANT	6.2254	9.39e-08	7.82e-08	N
R6S1	SCALED-RAMP	4.2508	0.859	1.79	Y
R6S1	V0	6.2223	8.61e-08	-1.49e-07	N
R6S1	NEUTRAL	6.2239	1.99e-07	-1.86e-08	N
R6S2	ACTUAL	6.2221	8.58e-08	-0.00092	N
R6S2	UP-RAMP	4.2355	0.859	1.97	Y
R6S2	CONSTANT	6.2220	1.09e-07	-0.00011	N
R6S2	SCALED-RAMP	4.2340	0.86	1.93	Y
R6S2	V0	6.2219	9.02e-08	-5.59e-08	N
R6S2	NEUTRAL	4.2407	0.861	1.88	Y
R6S3	ACTUAL	6.2221	8.35e-08	1.86e-08	N
R6S3	UP-RAMP	4.2503	0.858	1.86	Y
R6S3	CONSTANT	6.2220	7.63e-08	1.49e-08	N
R6S3	SCALED-RAMP	4.2363	0.86	1.93	Y
R6S3	V0	6.2219	7.91e-08	-4.84e-08	N
R6S3	NEUTRAL	4.2311	0.858	1.81	Y
C8S1	ACTUAL	6.2221	8.79e-08	-5.96e-08	N
C8S1	UP-RAMP	4.2452	0.859	1.86	Y
C8S1	CONSTANT	6.2076	0.0649	0.00437	N
C8S1	SCALED-RAMP	4.2465	0.859	1.95	Y
C8S1	V0	6.2218	7.02e-08	2.37e-05	N
C8S1	NEUTRAL	4.2444	0.858	1.86	Y
C8S2	ACTUAL	6.2222	9.09e-08	1.65e-05	N
C8S2	UP-RAMP	6.2220	8.74e-08	-7.45e-09	N
C8S2	CONSTANT	6.2220	8.77e-08	6.79e-05	N
C8S2	SCALED-RAMP	4.2385	0.861	1.93	Y

Seed	Branch	NLL	D_{16}	C	H
C8S2	V0	6.2221	1.13e-07	-5.96e-08	N
C8S2	NEUTRAL	4.2605	0.855	1.91	Y
C8S3	ACTUAL	6.2220	2.58e-07	-2.61e-08	N
C8S3	UP-RAMP	4.2363	0.857	1.91	Y
C8S3	CONSTANT	6.2218	7.21e-08	0.000224	N
C8S3	SCALED-RAMP	4.2444	0.862	1.8	Y
C8S3	V0	6.2216	8.4e-08	-0.00135	N
C8S3	NEUTRAL	4.2516	0.858	1.96	Y

WINDOW_RAMP shares the exact R6 update-3 parent and multiplies the native LR over updates 4–28 by $k/25$, where $k = 1$ at update 4 and $k = 25$ at update 28; update 29 onward is the native no-warmup schedule. ACTUAL+clip1 starts from the exact R6 state zero and clips the complete global gradient norm to 1.0 after backward and before every AdamW step. No NEUTRAL+clip or second ramp shape was run.

Table 14: Final R6 controls. WINDOW-RAMP is the same-window scalar schedule intervention; ACTUAL+clip1 uses global gradient-norm clipping throughout.

study	seed	branch	NLL	D_{16}	C	healthy
ramp	R6S1	WINDOW_RAMP	4.236773	0.860513	1.891704	true
ramp	R6S2	WINDOW_RAMP	4.235526	0.858653	1.971818	true
ramp	R6S3	WINDOW_RAMP	4.250256	0.858082	1.855909	true
clip1	R6S1	ACTUAL_CLIP1	6.222251	8.06269e-08	0.000440	false
clip1	R6S2	ACTUAL_CLIP1	6.221720	9.37795e-08	0.000000	false
clip1	R6S3	ACTUAL_CLIP1	6.222070	7.92654e-08	0.000000	false

WINDOW_RAMP is healthy on 3/3 seeds, whereas ACTUAL+clip1 is healthy on 0/3. Thus the tested same-window scalar schedule intervention can also select healthy continuation, while clipping at 1.0 does not remove the tested failure regime. Neither result is universal over warmup or clipping choices.

13 EXACT-DOSE SHAPE CONTROLS AND CROSS-SCALE RAMP TRANSFER

The native cosine schedule changes slightly over updates 4–28, so controls are matched using $\sum_t \ell_t m_t$, not the unweighted multiplier sum. The exact ascending dose is 0.007799697439118994 and the native-window dose is 0.014999594024817781. The frozen reverse scalar is 0.9999765499514898 and the constant multiplier is 0.5199939029158989. Neither multiplier was fitted to outcomes.

The ascending R6 ramp is healthy on 3/3 seeds, while its exact-dose reverse and constant controls are each healthy on 0/3. ASCENDING-SHOCK adds multiplier 1.48 at update 29 after the successful ramp and remains healthy on 3/3. The historical exact-dose V-profile solution is $b = -9.950150963155328 \times 10^{-6}$, outside its frozen $[0, 1]$ admissibility interval; it was therefore stopped before any optimizer update and retains no endpoint.

V0 is a new pre-registered approximate-dose, release-continuous control, not a repair of the historical V-BRIDGE. At updates 4–28 its multiplier traces a continuous V: $1 - (k - 1)/12$ for $k = 1, \ldots, 13$ and $(k - 13)/12$ for $k = 14, \ldots, 25$, reaching 0 at the midpoint and returning to 1 at update 28 with no release discontinuity. Its weighted dose is 0.007799769078464119, within 7.2×10^{-8} of the exact ascending dose; it is not exact-dose and we do not claim exact-dose equivalence. V0 is healthy on 0/3 seeds: all three endpoints are near NLL 6.222, $D_{16} \approx 8$–9×10^{-8}, and $C \approx 0$, with zero replay. Release continuity alone is therefore insufficient in the tested R6 control.

Table 15: Scalar-window comparison on the three original R6 master seeds. Sustained release ratio is the native multiplier after release divided by the update-28 multiplier; the SHOCK row separately has a transient update-29 spike. NEUTRAL uses native scalar LR but changes direction, so its LR dose is not directional path length. V0 is approximate-dose.

intervention	first multiplier	sustained release ratio	transient one-step spike	dose/native	healthy
SCALED-RAMP	0.0207998	1.9231	none	0.270394	3/3
UP-RAMP	0.04	1	none	0.519994	3/3
ASCENDING-SHOCK	0.04	1	1.48 at $t = 29$	0.519994	3/3
HALF-LR	0.5	2	none	0.500000	0/3
CONSTANT	0.519994	1.9231	none	0.519994	0/3
REVERSE	0.999977	25.001	none	0.519994	0/3
V0	1	1	none	0.519999	0/3
NEUTRAL	1	1	none	1	2/3

Table 16: Historical V-BRIDGE is an analytic inadmissibility result, not a failed run.

candidate	exact-dose midpoint	frozen admissible bound	status
historical V-BRIDGE	$-9.950150963155328 \times 10^{-6}$	$0 \le b \le 1$	inadmissible; no endpoint

SCALED-RAMP multiplies the ascending profile by the frozen constant $c = 0.5199939029158989$, so $m(k) = ck/25$ at updates 4–28 with the native schedule from update 29; its weighted dose is 0.004055795112930627 and its release step-up is $1/c \approx 1.9231$. It remains healthy on 3/3 seeds (NLL 4.234046, 4.236285, 4.250774; $D_{16} \approx 0.860$; $C \approx 1.79$–1.93), with zero replay. Despite a sustained approximately $1.92\times$ release step-up, the scaled ascending profile remains healthy on all tested seeds; release step-up alone is therefore insufficient to explain the historical constant-profile failure. Together with V0, the executed controls support a role for within-window temporal ordering beyond release continuity alone.

The prospective R6 repetition is healthy on 3/3, matches every historical label, and differs from the corresponding historical terminal NLL by 0.000173, 0.000705, and 0.000150. Its exact update-28 checkpoints and component digests support the pre-registered zero-update ACTUAL/NEUTRAL/RAMP comparison:

The prospective-table ACTUAL entries read the byte-identical historical P2-R1 update-28 ACTUAL checkpoints (SHA-256 respectively `3485320077449eb6ca2c3fa0a6af0c2d75b58d6f4f6d11276e2f19cc3c5ed084`, `0d5392fc197a6957d27f3e3b29258ed49d44d400496baae3f478536fb1cabb68`, and `b490cb9fe1f245b195189689c6d22629901db4aa233fdafa489bb952ca11f1d0`). Both audits use Apple MPS and the same first-16-row frozen validation batch (array SHA-256 `eab47bff512ac607b6afb449fa80063604472c3ac9ff89c06956e3dbc9d8df50`); the small numerical differences arise from separate nondeterministic MPS forward recomputations under the historical and prospective diagnostic scripts (SHA-256 `4afc44a344e4e4972931cba6ceb8e74888cb72c54c31422aefb0684365d7d08e` and `6381d6d80c2db70928594f7db664999134fe0458583ba0d0ea76f2b388f26b74`), not from regenerated ACTUAL states. Endpoint labels and qualitative comparisons are unchanged.

Table 17: Completed window-shape, replication, transfer, start-position, and duration endpoints. Inadmissible or integrity-blocked branches have no endpoint and are described in text.

seed	branch	NLL	D_{16}	C	healthy
R8S1	WINDOW_RAMP	6.225716	1.02781e-07	-0.002180	false
R8S2	WINDOW_RAMP	6.222860	7.84755e-08	0.000000	false
R8S3	WINDOW_RAMP	6.222073	7.75869e-08	-0.000000	false
P1S1	WINDOW_RAMP	6.231933	7.22305e-08	-0.000000	false
P1S2	WINDOW_RAMP	6.234258	7.93237e-08	-0.000000	false
P1S3	WINDOW_RAMP	6.233256	9.05226e-08	-0.000000	false
R6S1	REVERSE_RAMP	6.222122	7.89255e-08	0.000000	false
R6S2	REVERSE_RAMP	6.221845	9.65335e-08	0.000000	false
R6S3	REVERSE_RAMP	6.221979	8.23067e-08	0.008196	false
R6S1	CONSTANT_MATCHED_DOSE	6.225388	9.39246e-08	0.000000	false
R6S2	CONSTANT_MATCHED_DOSE	6.221954	1.09266e-07	-0.000110	false
R6S3	CONSTANT_MATCHED_DOSE	6.222043	7.62588e-08	0.000000	false
R6S1	ASCENDING_SHOCK	4.246812	0.861069	1.712379	true
R6S2	ASCENDING_SHOCK	4.232932	0.860471	1.841573	true
R6S3	ASCENDING_SHOCK	4.253058	0.85545	1.700045	true
R6S1	PROSPECTIVE_UP_RAMP	4.236600	0.859815	1.834369	true
R6S2	PROSPECTIVE_UP_RAMP	4.234820	0.858416	1.936968	true
R6S3	PROSPECTIVE_UP_RAMP	4.250406	0.857419	1.907941	true
R8DA1	UP_RAMP	4.214382	0.851264	2.189521	true
R8DA2	UP_RAMP	4.229490	0.852097	2.023353	true
R8DA3	UP_RAMP	4.215667	0.853712	1.819617	true
R8S1	START1_RAMP25	6.221856	8.00211e-08	0.000000	false
R8S2	START1_RAMP25	6.229111	9.88236e-08	-0.008535	false
R8S3	START1_RAMP25	6.222092	7.98114e-08	-0.001236	false
R8S1	START4_RAMP100	4.219524	0.864031	1.823843	true
R8S2	START4_RAMP100	4.230845	0.86055	2.004395	true
R8S3	START4_RAMP100	4.222764	0.863977	1.784110	true
R8S1	START4_RAMP50	4.207544	0.857487	1.797088	true
R8S2	START4_RAMP50	4.236145	0.860739	1.895997	true
R8S3	START4_RAMP50	6.222093	8.51105e-08	0.000385	false
P1S1	START4_RAMP100	6.230378	8.37264e-08	-0.000000	false
P1S2	START4_RAMP100	6.233390	7.98209e-08	-0.000000	false
P1S3	START4_RAMP100	6.229627	7.30124e-08	-0.000000	false
R6S1	V0	6.222268	8.6054e-08	-0.000000	false
R6S2	V0	6.221912	9.0166e-08	-0.000000	false
R6S3	V0	6.221883	7.90514e-08	-0.000000	false
R6S1	SCALED_RAMP	4.250774	0.859279	1.787996	true
R6S2	SCALED_RAMP	4.234046	0.86007	1.926642	true
R6S3	SCALED_RAMP	4.236285	0.860009	1.934257	true
P1S1	START4_RAMP200	6.229882	8.8152e-08	-0.000000	false
P1S2	START4_RAMP200	6.229763	8.32527e-08	0.000000	false
P1S3	START4_RAMP200	6.231143	7.12581e-08	0.000000	false

Table 18: Complete tested-point matrix. Cells are healthy seeds / tested seeds; – means untested.

recipe	ACTUAL	NEUTRAL	ANGLE	HALF	UP-RAMP	REVERSE	CONST	CLIP1	Warmup500	SHOCK	V0	SCALED
24M 6L, peak 6×10^{-4}	0/3	2/3	0/3	0/3	3/3	0/3	0/3	0/3	3/3	3/3	0/3	3/3
24M 6L, peak 4.5×10^{-4}	0/3	3/3	0/3	–	–	–	–	–	3/3	–	–	–
30M 8L, peak 6×10^{-4}	0/3	0/3	0/3	–	0/3	–	–	–	3/3	–	–	–
30M 8L, peak 4.5×10^{-4}	0/3	2/3	–	–	–	–	–	–	3/3	–	–	–
30M 8L, peak $3.897114317 \times 10^{-4}$	0/3	3/3	0/3	–	3/3	–	–	–	3/3	–	–	–
107M 6L, peak 6×10^{-4}	0/3	0/3	–	–	0/3	–	–	–	3/3	–	–	–
107M 6L, peak 4.5×10^{-4}	0/3	0/3	–	–	–	–	–	–	3/3	–	–	–
107M 6L, peak 3×10^{-4}	0/3	0/3	–	–	–	–	–	–	3/3	–	–	–

Table 19: R6 shared update-3 PRE reference and update-28 ACTUAL/NEUTRAL/RAMP zero-update diagnostics. All states use the same frozen batch and definitions; PRE is the historical resident t3 object.

seed	state	D	norm. entropy	spectral conc.	effective rank
R6S1	PRE	0.0182301	0.991886	0.057943	81.6557
R6S1	ACTUAL	1.96529e-06	0.997203	0.278366	30.3910
R6S1	NEUTRAL	0.0382853	0.965142	0.866726	2.1527
R6S1	RAMP	0.00294998	0.993658	0.186292	52.1581
R6S2	PRE	0.0232768	0.990819	0.076651	77.7785
R6S2	ACTUAL	6.27419e-06	0.995294	0.439585	14.9072
R6S2	NEUTRAL	0.339676	0.814698	0.956586	1.3144
R6S2	RAMP	0.047474	0.977271	0.417898	28.3781
R6S3	PRE	0.0245203	0.991580	0.062852	80.5013
R6S3	ACTUAL	5.82675e-06	0.996635	0.377276	20.8579
R6S3	NEUTRAL	0.337019	0.688828	0.786294	2.3307
R6S3	RAMP	0.0265385	0.988127	0.384981	29.8766

Both tested successful routes raise D and lower normalized attention entropy relative to ACTUAL across all three seeds, but these changes are not sufficient and their magnitude does not track outcome. Effective rank moves in opposite directions: RAMP gives 52.2/28.4/29.9, ACTUAL 30.4/14.9/20.9, and NEUTRAL 2.15/1.31/2.33. Distinct release-state organizations can lead to healthy functional endpoints. These quantities are descriptive; they define no predictor, threshold, mediator, or common attractor.

Table 20: Prospective R6 update-28 moment audit: family norms. Native AdamW quantity $m/(\sqrt{v}+\epsilon)$ with $\epsilon = 10^{-8}$; 54 keyed rows (3 seeds × 3 states × 6 families). Zero optimizer updates.

seed	state	family	m L2	v L2	norm. L2	$v \le \epsilon^2$ frac
R6S1	ACTUAL	TOKEN_IO	0.2163	0.01147	852.4	0
R6S1	ACTUAL	POSITION	0.002869	2.213e-05	15.01	0
R6S1	ACTUAL	ATTENTION	0.01272	2.275e-05	186.7	0.08333
R6S1	ACTUAL	MLP	0.02921	2.446e-05	380.6	0
R6S1	ACTUAL	LN_SCALE	0.006246	8.881e-05	11.07	0
R6S1	ACTUAL	LN_SHIFT	0.004332	3.652e-05	7.621	0
R6S1	NEUTRAL	TOKEN_IO	0.2186	0.01299	847.7	0
R6S1	NEUTRAL	POSITION	0.01612	0.0002041	31.67	0
R6S1	NEUTRAL	ATTENTION	0.05784	0.0002596	459	0.06931
R6S1	NEUTRAL	MLP	0.04553	6.136e-05	566.8	0
R6S1	NEUTRAL	LN_SCALE	0.00728	0.0001007	13.95	0
R6S1	NEUTRAL	LN_SHIFT	0.008043	6.38e-05	10.78	0
R6S1	RAMP	TOKEN_IO	0.6649	0.0243	1257	0
R6S1	RAMP	POSITION	0.002602	2.213e-05	15.32	0
R6S1	RAMP	ATTENTION	0.02419	2.82e-05	304.2	0.08333
R6S1	RAMP	MLP	0.03415	2.651e-05	561.7	0
R6S1	RAMP	LN_SCALE	0.03154	0.0002263	19.1	0
R6S1	RAMP	LN_SHIFT	0.02414	7.186e-05	18.04	0
R6S2	ACTUAL	TOKEN_IO	0.2223	0.01077	857.9	0
R6S2	ACTUAL	POSITION	0.002671	1.72e-05	14.78	0
R6S2	ACTUAL	ATTENTION	0.01268	1.917e-05	194.1	0.08333
R6S2	ACTUAL	MLP	0.02912	2.229e-05	391.3	0
R6S2	ACTUAL	LN_SCALE	0.005842	7.208e-05	10.45	0
R6S2	ACTUAL	LN_SHIFT	0.003933	3.326e-05	7.401	0
R6S2	NEUTRAL	TOKEN_IO	0.396	0.04748	868.3	0
R6S2	NEUTRAL	POSITION	0.2442	0.01368	81.86	0
R6S2	NEUTRAL	ATTENTION	0.6224	0.008385	787.3	0.006726
R6S2	NEUTRAL	MLP	0.4532	0.002739	1086	0
R6S2	NEUTRAL	LN_SCALE	0.04195	0.0007361	24.42	0
R6S2	NEUTRAL	LN_SHIFT	0.08524	0.002271	21.78	0
R6S2	RAMP	TOKEN_IO	0.7731	0.02519	1286	0
R6S2	RAMP	POSITION	0.003583	1.725e-05	34.74	0
R6S2	RAMP	ATTENTION	0.02252	2.506e-05	825.8	0.08333
R6S2	RAMP	MLP	0.02947	2.405e-05	828.2	0
R6S2	RAMP	LN_SCALE	0.0337	0.0001776	28.41	0
R6S2	RAMP	LN_SHIFT	0.0271	6.677e-05	20.12	0
R6S3	ACTUAL	TOKEN_IO	0.2207	0.01194	868.5	0
R6S3	ACTUAL	POSITION	0.002744	1.997e-05	14.8	0
R6S3	ACTUAL	ATTENTION	0.01296	2.4e-05	188.2	0.08333
R6S3	ACTUAL	MLP	0.02893	2.296e-05	385.8	0
R6S3	ACTUAL	LN_SCALE	0.005778	8.653e-05	10.49	0
R6S3	ACTUAL	LN_SHIFT	0.004109	3.54e-05	7.411	0
R6S3	NEUTRAL	TOKEN_IO	0.225	0.01489	883.2	0
R6S3	NEUTRAL	POSITION	0.03565	0.000191	95.09	0
R6S3	NEUTRAL	ATTENTION	0.06322	0.0001722	802.3	0.04763
R6S3	NEUTRAL	MLP	0.0578	5.342e-05	978.1	0
R6S3	NEUTRAL	LN_SCALE	0.007291	8.872e-05	20.7	0
R6S3	NEUTRAL	LN_SHIFT	0.008574	5.616e-05	12.59	0
R6S3	RAMP	TOKEN_IO	0.7849	0.02665	1294	0
R6S3	RAMP	POSITION	0.002767	1.999e-05	25.33	0
R6S3	RAMP	ATTENTION	0.02187	3.054e-05	500.6	0.08333
R6S3	RAMP	MLP	0.02899	2.507e-05	693.1	0
R6S3	RAMP	LN_SCALE	0.03604	0.0002117	24.96	0
R6S3	RAMP	LN_SHIFT	0.02802	7.084e-05	20.28	0

Table 21: Prospective R6 update-28 moment audit: distributional summaries. Columns are v quantiles, denominator $\sqrt{v}+\epsilon$ quantiles, and absolute normalized-coordinate quantiles per keyed row.

seed	state	family	v med	v p90	v p99	v max	den med	den p90	den p99	den max	nrm med	nrm p90	nrm p99	nrm max
R6S1	ACTUAL	TOKEN_IO	1.985e-09	1.848e-08	1.74e-07	0.001678	4.456e-05	0.0001359	0.0004171	0.04096	0.2741	0.5944	0.959	1.617
R6S1	ACTUAL	POSITION	1.564e-08	7.594e-08	3.034e-07	3.409e-06	0.0001251	0.0002756	0.0005508	0.001846	0.05939	0.07931	0.09118	0.141
R6S1	ACTUAL	ATTENTION	2.061e-11	4.748e-09	3.331e-08	9.748e-07	4.549e-06	6.892e-05	0.0001825	0.0009873	0.05396	0.08384	0.265	1.29
R6S1	ACTUAL	MLP	4.448e-10	3.805e-09	2.476e-08	1.738e-06	2.11e-05	6.17e-05	0.0001574	0.001318	0.06122	0.1166	0.4376	1.453
R6S1	ACTUAL	LN_SCALE	6.835e-09	8.354e-08	3.5e-06	3.713e-05	8.268e-05	0.000289	0.001871	0.006094	0.07057	0.1731	0.6075	1.143
R6S1	ACTUAL	LN_SHIFT	8.456e-08	4.992e-07	2.01e-06	6.983e-06	0.0002908	0.0007066	0.001418	0.002643	0.06615	0.1137	0.3804	0.788
R6S1	NEUTRAL	TOKEN_IO	1.996e-09	1.868e-08	1.778e-07	0.001951	4.468e-05	0.0001367	0.0004217	0.04417	0.2716	0.5924	0.9445	1.619
R6S1	NEUTRAL	POSITION	5.063e-08	3.476e-07	2.372e-06	4.534e-05	0.000225	0.0005895	0.00154	0.006734	0.1214	0.1715	0.2055	0.6158
R6S1	NEUTRAL	ATTENTION	5.764e-10	4.395e-08	3.79e-07	1.526e-05	2.402e-05	0.0002096	0.0006156	0.003906	0.1403	0.2821	0.5321	1.338
R6S1	NEUTRAL	MLP	1.49e-09	1.209e-08	6.342e-08	9.11e-06	3.861e-05	0.00011	0.0002518	0.003018	0.1179	0.2225	0.4925	1.443
R6S1	NEUTRAL	LN_SCALE	3.566e-08	3.787e-07	4.267e-06	4.404e-05	0.0001888	0.0006154	0.002066	0.006636	0.1386	0.2326	0.5745	1.162
R6S1	NEUTRAL	LN_SHIFT	2.365e-07	1.133e-06	3.177e-06	1.334e-05	0.0004864	0.001064	0.001782	0.003653	0.1272	0.1796	0.3296	0.6498
R6S1	RAMP	TOKEN_IO	2.535e-09	1.919e-08	1.862e-07	0.003446	5.036e-05	0.0001385	0.0004315	0.05871	0.4489	0.8592	1.117	1.636

seed	state	family	v med	v p90	v p99	v max	den med	den p90	den p99	den max	nrm med	nrm p90	nrm p99	nrm max
R6S1	RAMP	POSITION	1.567e-08	7.598e-08	3.034e-07	3.409e-06	0.0001252	0.0002757	0.0005508	0.001846	0.05129	0.08208	0.1492	0.5933
R6S1	RAMP	ATTENTION	2.135e-11	5.667e-09	4.288e-08	1.064e-06	4.63e-06	7.529e-05	0.0002071	0.001031	0.05521	0.1645	0.4905	1.363
R6S1	RAMP	MLP	4.932e-10	4.293e-09	2.721e-08	1.738e-06	2.222e-05	6.553e-05	0.000165	0.001318	0.0686	0.2447	0.5924	1.364
R6S1	RAMP	LN_SCALE	7.685e-09	1.046e-07	8.751e-06	0.0001054	8.768e-05	0.0003234	0.002958	0.01027	0.07608	0.4243	0.8148	1.055
R6S1	RAMP	LN_SHIFT	9.022e-08	5.558e-07	4.495e-06	1.626e-05	0.0003004	0.0007455	0.00212	0.004032	0.06149	0.3554	0.8261	1.023
R6S2	ACTUAL	TOKEN_IO	2.046e-09	1.843e-08	1.775e-07	0.001507	4.524e-05	0.0001358	0.0004213	0.03882	0.2831	0.5979	0.95	1.634
R6S2	ACTUAL	POSITION	1.594e-08	6.979e-08	2.415e-07	2.328e-06	0.0001263	0.0002642	0.0004915	0.001526	0.05784	0.07874	0.0926	0.1326
R6S2	ACTUAL	ATTENTION	2.047e-11	4.381e-09	3.012e-08	8.924e-07	4.535e-06	6.62e-05	0.0001736	0.0009447	0.05317	0.08718	0.2888	1.262
R6S2	ACTUAL	MLP	4.321e-10	3.492e-09	2.285e-08	1.181e-06	2.08e-05	5.91e-05	0.0001512	0.001087	0.06057	0.1262	0.4544	1.342
R6S2	ACTUAL	LN_SCALE	6.791e-09	7.883e-08	3.629e-06	3.981e-05	8.242e-05	0.0002808	0.001905	0.00631	0.06748	0.1775	0.5388	1.047
R6S2	ACTUAL	LN_SHIFT	7.944e-08	4.205e-07	2.045e-06	6.712e-06	0.0002819	0.0006485	0.00143	0.002591	0.06492	0.1167	0.3379	0.7203
R6S2	NEUTRAL	TOKEN_IO	2.178e-09	2.143e-08	2.949e-07	0.008609	4.668e-05	0.0001464	0.000543	0.09278	0.294	0.6077	0.9161	1.627
R6S2	NEUTRAL	POSITION	4.749e-07	1.675e-05	0.0002199	0.003828	0.0006892	0.004093	0.01483	0.06187	0.2779	0.4469	0.8231	1.353
R6S2	NEUTRAL	ATTENTION	2.33e-08	1.326e-06	1.255e-05	0.0004211	0.0001527	0.001151	0.003543	0.02052	0.293	0.392	0.7437	1.533
R6S2	NEUTRAL	MLP	2.358e-08	4.037e-07	2.729e-06	0.000202	0.0001536	0.0006354	0.001652	0.01421	0.288	0.3915	0.7258	1.561
R6S2	NEUTRAL	LN_SCALE	7.862e-07	8.112e-06	3.718e-05	0.0002573	0.0008867	0.002848	0.006098	0.01604	0.2905	0.3947	0.7509	1.191
R6S2	NEUTRAL	LN_SHIFT	3.846e-06	3.948e-05	0.0001245	0.000302	0.001961	0.006283	0.01116	0.01738	0.2865	0.3279	0.4555	0.8281
R6S2	RAMP	TOKEN_IO	2.678e-09	1.921e-08	1.85e-07	0.004146	5.176e-05	0.0001386	0.0004302	0.06439	0.4718	0.8742	1.102	1.601
R6S2	RAMP	POSITION	1.611e-08	7.005e-08	2.424e-07	2.328e-06	0.0001269	0.0002647	0.0004923	0.001526	0.05904	0.1932	0.5264	1.312
R6S2	RAMP	ATTENTION	2.58e-11	5.469e-09	4.049e-08	9.754e-07	5.09e-06	7.396e-05	0.0002012	0.0009876	0.1125	0.5685	1.115	1.606
R6S2	RAMP	MLP	4.92e-10	3.98e-09	2.526e-08	1.181e-06	2.219e-05	6.31e-05	0.0001589	0.001087	0.1064	0.3681	0.8293	1.572
R6S2	RAMP	LN_SCALE	8.133e-09	1.071e-07	8.655e-06	9.979e-05	9.019e-05	0.0003272	0.002942	0.00999	0.1563	0.7203	1.06	1.448
R6S2	RAMP	LN_SHIFT	8.606e-08	4.56e-07	4.521e-06	1.484e-05	0.0002934	0.0006753	0.002126	0.003853	0.05578	0.3418	0.8929	1.2
R6S3	ACTUAL	TOKEN_IO	1.985e-09	1.834e-08	1.789e-07	0.001592	4.457e-05	0.0001354	0.000423	0.0399	0.2855	0.6079	0.9554	1.608
R6S3	ACTUAL	POSITION	1.614e-08	7.117e-08	2.778e-07	4.393e-06	0.0001271	0.0002668	0.000527	0.002096	0.0577	0.07936	0.09266	0.1645
R6S3	ACTUAL	ATTENTION	1.972e-11	4.525e-09	3.337e-08	2.07e-06	4.45e-06	6.728e-05	0.0001827	0.001439	0.05349	0.08575	0.2717	1.127
R6S3	ACTUAL	MLP	4.289e-10	3.526e-09	2.42e-08	1.218e-06	2.072e-05	5.939e-05	0.0001556	0.001104	0.06123	0.1235	0.4446	1.342
R6S3	ACTUAL	LN_SCALE	7.006e-09	8.273e-08	3.688e-06	2.93e-05	8.371e-05	0.0002876	0.00192	0.005413	0.0693	0.1728	0.5116	1.22
R6S3	ACTUAL	LN_SHIFT	7.338e-08	4.424e-07	2.085e-06	6.114e-06	0.0002709	0.0006651	0.001444	0.002473	0.06634	0.1175	0.3292	0.9671
R6S3	NEUTRAL	TOKEN_IO	2.052e-09	1.858e-08	1.901e-07	0.00176	4.531e-05	0.0001363	0.000436	0.04195	0.2969	0.6164	0.9285	1.631
R6S3	NEUTRAL	POSITION	7.137e-08	3.351e-07	2.239e-06	3.668e-05	0.0002672	0.0005789	0.001496	0.006057	0.2338	0.63	0.965	1.326
R6S3	NEUTRAL	ATTENTION	8.78e-10	3.13e-08	2.389e-07	8.838e-06	2.964e-05	0.0001769	0.0004888	0.002973	0.2003	0.5248	0.9384	1.504
R6S3	NEUTRAL	MLP	1.698e-09	1.101e-08	5.388e-08	5.093e-06	4.122e-05	0.0001049	0.0002321	0.002257	0.1771	0.4411	0.8209	1.555
R6S3	NEUTRAL	LN_SCALE	3.669e-08	3.195e-07	4.1e-06	2.836e-05	0.0001916	0.0005653	0.002025	0.005325	0.1687	0.4177	0.7916	1.292
R6S3	NEUTRAL	LN_SHIFT	2.146e-07	1.066e-06	2.847e-06	6.317e-06	0.0004632	0.001033	0.001687	0.002513	0.141	0.2289	0.3725	0.6428
R6S3	RAMP	TOKEN_IO	2.603e-09	1.916e-08	1.896e-07	0.003546	5.103e-05	0.0001384	0.0004354	0.05954	0.4698	0.8777	1.13	1.645
R6S3	RAMP	POSITION	1.623e-08	7.135e-08	2.781e-07	4.394e-06	0.0001274	0.0002671	0.0005274	0.002096	0.05061	0.1368	0.3603	1.267
R6S3	RAMP	ATTENTION	2.129e-11	5.568e-09	4.322e-08	2.064e-06	4.624e-06	7.463e-05	0.0002079	0.001437	0.07189	0.2951	0.7921	1.62
R6S3	RAMP	MLP	4.805e-10	3.96e-09	2.589e-08	1.219e-06	2.193e-05	6.294e-05	0.0001609	0.001104	0.08884	0.3079	0.6867	1.602
R6S3	RAMP	LN_SCALE	8.187e-09	1.041e-07	8.123e-06	7.653e-05	9.049e-05	0.0003226	0.00285	0.008748	0.1256	0.5646	0.9495	1.448
R6S3	RAMP	LN_SHIFT	7.921e-08	4.923e-07	4.411e-06	1.418e-05	0.0002814	0.0007016	0.0021	0.003766	0.05267	0.362	0.9079	1.078

A companion zero-update audit reads the frozen update-28 ACTUAL, NEUTRAL, and prospective-RAMP checkpoints for all three R6 master seeds and reports the native AdamW quantity $m/(\sqrt{v}+\epsilon)$ ($\epsilon = 10^{-8}$) under the Table-19 family partition, with m and v norms, denominator summaries, the fraction of $v \le \epsilon^2$ coordinates, and normalized-coordinate quantiles (54 keyed rows above). Healthy RAMP escapes the uniformly attenuated ACTUAL normalized-moment regime—its normalized-family L2 norm exceeds ACTUAL's in every family and seed, with RAMP/ACTUAL ratios spanning 1.02–2.37 (R6S1), 1.50–4.25 (R6S2), and 1.49–2.74 (R6S3)—but it does not converge to NEUTRAL's optimizer state: RAMP/NEUTRAL is 0.484/0.424/0.266 in POSITION, 0.663/1.05/0.624 in ATTENTION, and 0.991/0.763/0.709 in MLP, while RAMP exceeds NEUTRAL consistently in TOKEN_IO and LN_SCALE. The audit executes no optimizer update, introduces no new metric, and supports no mediation, necessity, or common-attractor claim.

At the original eight-layer recipe, START1-RAMP25 is healthy on 0/3, START4-RAMP100 is healthy on 3/3, and the conditional START4-RAMP50 stage is mixed at 2/3. At the R8-DA lower-LR recipe, the original 25-update ramp is healthy on 3/3. At 107M, START4-RAMP100 is 0/3; START1-RAMP25 has two valid failed endpoints and one replay-invalid branch, and no valid clean-rerun slot remained inside the submission-time compute cutoff, so that branch stays incomplete rather than failed. The conditional RAMP200 stage is complete at 0/3. Two attempts are excluded and supply no endpoint: the original first-seed run had multiple writers, and the second seed's first clean rerun ended after an infrastructure logging fault that left inconsistent rolling checkpoints. All three valid RAMP200 seeds were executed from the exact frozen update-3 parents under the identical protocol with zero replay and verified endpoint records: NLL 6.229882, 6.229763, and 6.231143; D_{16} 8.8, 8.3, and 7.1×10^{-8}; $C \approx 0$. Neither invalid attempt is counted as a failure.

Table 22: Final tested window-duration and boundary summary. Blocked, inadmissible, and incomplete cells are not treated as failures.

recipe	R25	START1-R25	R50	R100	R200	SHOCK	V0/V
24M 6L, 6×10^{-4}	3/3	–	–	–	UNTESTED	3/3	V0 0/3; historical V inadmissible
30M 8L, 6×10^{-4}	0/3	0/3	2/3	3/3	–	–	–
30M 8L, $3.897114317 \times 10^{-4}$	3/3	–	–	–	–	–	–
107M 6L, 6×10^{-4}	0/3	0/2 valid	–	0/3	0/3	–	–

Table 23: Co-transfer cells in which both NEUTRAL and RAMP25 were evaluated. The two routes have the same cell-level success/failure direction in all four cells.

cell	NEUTRAL	RAMP25
24M 6L, 6×10^{-4}	2/3	3/3
30M 8L, 6×10^{-4}	0/3	0/3
30M 8L, $3.897114317 \times 10^{-4}$	3/3	3/3
107M 6L, 6×10^{-4}	0/3	0/3

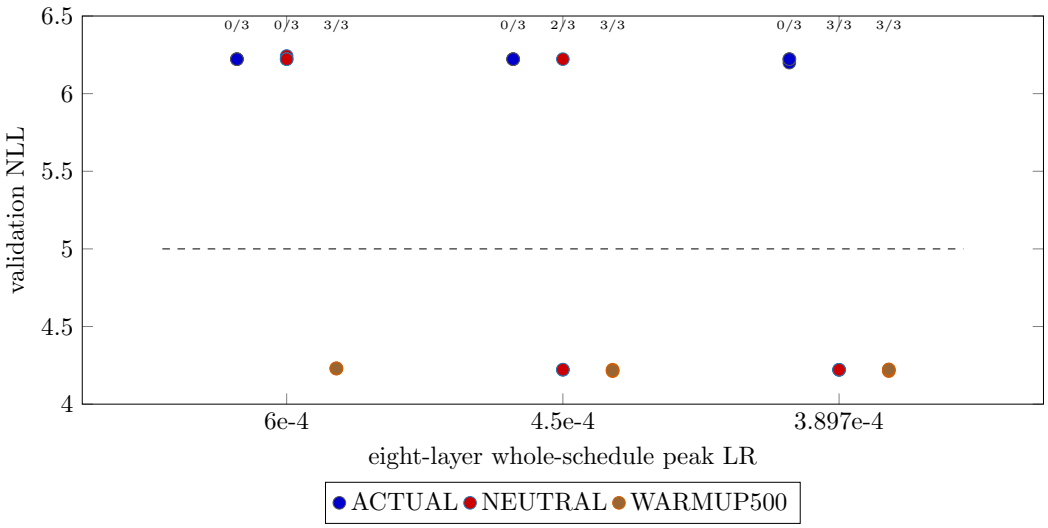

Figure 1: Original eight-layer whole-schedule LR comparison, retained as a supplementary tested-point figure; no interpolated threshold.

At the cell level, the two tested routes have the same success/failure direction in all four cells where both were evaluated. Post hoc, R8 master seed 2026092103 is the failing seed for NEUTRAL at 4.5×10^{-4} and for RAMP50 at 6×10^{-4}; this coincidence is not a predictor.

The eight-layer evidence supports a tested 25-failed, 50-mixed, 100-healthy pattern, not a mathematical minimum. At 107M, the tested 25-, 100-, and 200-update ramps all fail at the original recipe; this is a tested-duration failure set at one exact recipe, not a mathematical minimum and not a width law. Moving the start to update 1 does not rescue eight layers; the 107M start-position result remains incomplete. Ma and Yarats' $2/(1-\beta_2) = 100$ rule motivates the 100-update horizon through early adaptive-update magnitude, but fixed β_2 cannot explain the cross-recipe difference; the main paper gives the full citation.

No cross-cell sign test is reported. Several eight-layer cells reuse master seeds, development and confirmation stages differ, the program is sequential, and the cells are not independent exchangeable Bernoulli trials. Raw cell-wise outcomes remain primary.

14 Raw moment-carrier results

Table 24: Moment-carrier endpoints. At both lower-LR states, full $m+v$ and the completed NA route are healthy; isolated m or v transfer is not.

study	condition	NLL	D_{16}	C
R9S1	AA	6.222094	8.30272e-08	0.000000
R9S1	AN	4.243446	0.852044	1.904771
R9S1	NA	4.246761	0.854332	1.853460
R9S1	NN	4.244900	0.856158	2.072903
R9S1	A_mN_vA	6.222021	7.59083e-06	-0.000017
R9S1	A_mA_vN	6.116861	0.437469	0.037873
R9S2	AA	6.222198	1.22689e-07	0.000000
R9S2	NN	4.235148	0.853598	1.665971
R9S2	AN	4.227257	0.848872	1.801619
R9S2	A_mN_vA	6.223699	0.0193055	0.005803
R9S2	A_mA_vN	6.151374	0.26834	-0.034577
R9S2	NA	4.237875	0.849911	1.829794

R9S1 and R9S2 both pass cut-equivalence and endpoint-checkpoint verification. R9S1 is the original tested decomposition using the verified R10-R2 transcript-compatible release-state construction; R9S2 is the independent prospective confirmation. Its original decomposition records 24,913 verified updates and zero replay; the final NA completion adds exactly updates 29–5000 with zero replay and exact cut equivalence.

Table 25: Completed R9S2 NA endpoint.

study	seed	branch	NLL	D_{16}	C	healthy
R9S2	NA	NA	4.237875	0.849911	1.829794	true

R9S2 NA is healthy, as is its frozen AN corner. R9S2 therefore reproduces the redundant parameter/full-optimizer sufficiency pattern observed in R9S1, while the isolated m and v corners remain failed. In R9S1, the v-only endpoint (NLL 6.116861, $D_{16} = 0.437469$, $C = 0.037873$) is descriptive partial movement only and does not meet functional success.

Here "transcript-compatible" does not mean that a historical byte-identical t28 checkpoint was available. R10-R2 reconstructed ACTUAL and NEUTRAL independently from the frozen R9S1 update-3 parent by executing updates 4–28 under the frozen source code, batches, sampler/RNG state, LR schedule, and intervention transcript. It verified all 52 required pulse/D16/validation/probe comparisons, exact continuation-state agreement, named optimizer coverage and step 28, model/buffer integrity, serialization, and zero replay. This establishes equivalence to the frozen recorded transcript and a verified cut for continuation; it does not claim byte identity with a missing historical t28 file or universal equivalence beyond the verified components.

15 ZERO-UPDATE EXTERNAL RELEASE-STATE CHECK

Table 26: Zero-update R8-DA release-state diagnostics under the frozen P2 definitions.

seed	state	D	entropy	spectral	eff. rank
R8DA1	PRE	0.040620	0.990851	0.057949	81.5282
R8DA1	ACTUAL	0.000037	0.996371	0.375426	25.6150
R8DA1	NEUTRAL	0.236657	0.877449	0.972677	1.1827
R8DA2	PRE	0.039061	0.990880	0.062768	80.5282
R8DA2	ACTUAL	0.000022	0.996083	0.400129	20.0628
R8DA2	NEUTRAL	0.405192	0.881994	0.730050	3.0430
R8DA3	PRE	0.032702	0.990918	0.057618	81.5118
R8DA3	ACTUAL	0.000015	0.995900	0.244950	36.4017
R8DA3	NEUTRAL	0.221571	0.916064	0.634341	4.1393

The check reads only each exact shared $t3$ parent and the ACTUAL and NEUTRAL $t28$ checkpoints. It uses exactly D, normalized attention entropy, centered-hidden spectral concentration, and effective rank, with the same frozen batch, definitions, and aggregation as P2-R1. It performs no training and defines no new metric, threshold, predictor, significance test, or classifier. The observed direction of separation agrees qualitatively with P2-R1 under these frozen definitions.

16 METRIC-OBJECT CROSSWALK

D is a centered-token Frobenius-energy ratio, not algebraic rank. Normalized attention entropy is aggregated under the frozen WARM procedure and is not numerically interchangeable with every entropy definition in prior work. Spectral concentration is computed from centered hidden states, not from attention Q/K update matrices. Effective rank is entropy-based and is not algebraic rank. Token-similarity formulae are comparable to recent Post-Norm diagnostics, but residual and gradient objects remain only analogous; LayerNorm/RMSNorm differences prohibit identity claims.

17 MOMENT MEMORY AND UPDATE-SCALE DIAGNOSTICS

For $\beta_1 = 0.9$ and $\beta_2 = 0.98$, nominal half-lives $\log(1/2)/\log\beta$ are 6.5788 and 34.3096 updates. At continuation update 4972, direct donor coefficients are negligible. The endpoint consequence is therefore naturally interpreted as trajectory-mediated rather than as direct persistence of the initial moment tensors. At update 29, proposal norms are 14.1993 for failed m-only and 0.5897 for healthy full $m{+}v$. This rules out the monotonic explanation that a larger first continued update implies healthier continuation. It does not distinguish informational insufficiency of m alone from destructive mismatch-induced overshoot when donor m is combined with receiver v, and it does not make the large m-only step irrelevant. No clipped, rescaled, or $m/\sqrt{v}$ hybrid was run.

A zero-update family audit at R9S1 update 28 reports all six families for ACTUAL A, NEUTRAL N, and the N-m/A-v hybrid. Ratios are explicitly oriented: A/N means ACTUAL divided by NEUTRAL, while N/A means the reverse. The hybrid normalized-family norms are 46,588 for ATTENTION, 15,959 for MLP, and 611.65 for POSITION. In those families, donor m N/A is 10.24, 4.42, and 27.70, while receiver-to-donor median denominator A/N is 0.0850, 0.3520, and 0.4354. Hybrid 99th-percentile absolute normalized coordinates are 69.46, 17.75, and 9.13, with maxima 3246.5, 642.5, and 70.9. Attention has 8.33% of receiver v coordinates at or below ϵ^2, but a structural audit localizes them almost entirely to the first block's Q/K projection *weights*. The actual attention implementation has no key bias or other fixed softmax-invariant parameter slice. No complete attention tensor has mean $v \le \epsilon^2$; the 8.33% is coordinate-weighted, not per-tensor. Removing even all low-v coordinates changes the R9S1 hybrid attention norm only from 46,588.1 to 46,586.4. The generic key-bias-null

Table 27: Native and hybrid normalized-family references at R9S1 update 28. Ratios use the displayed numerator divided by denominator.

family	A native	N native	N-m/A-v	hyb/A	hyb/N	median	p90	p99	max	$v \le \epsilon^2$
TOKEN_IO	911.67	968.42	1164.6	1.277	1.203	0.3468	0.7456	1.491	27.96	0
POSITION	15.296	97.451	611.65	39.99	6.276	0.7317	2.939	9.129	70.9	0
ATTENTION	264.64	1038.6	46588	176	44.86	1.894	14.91	69.46	3247	0.08333
MLP	559.31	1429	15959	28.53	11.17	0.8594	4.828	17.75	642.5	0
LN_SCALE	13.376	33.399	334.09	24.98	10	1.198	5.652	17.55	69.9	0
LN_SHIFT	9.9439	27.137	198.55	19.97	7.316	0.8439	2.976	8.877	72.76	0

explanation is false for this model; the remaining functional-coordinate mismatch is a diagnostic hypothesis, not proof of a destructive update or mediation.

18 Sanity checks and incomplete controls

The full-validation empirical unigram NLL is 6.223748521448295: vocabulary 10,000, additive smoothing 0.5, counts from the frozen training corpus, all 1,658 standard validation sequences, and all 212,224 non-padding targets under the endpoint ignore-index semantics. Every frozen data hash matched. The 24M failed cluster is near/slightly below this value and the 107M failed cluster is above it. This remains a sanity comparison and does not identify either failed model as literally unigram.

The isotropic-noise control never entered formal science. Its repaired nonzero-dose qualification failed because the update-17 LN_SCALE realized-norm relative error was 1.3259714×10^{-5}, above the frozen 10^{-5} tolerance. There are no formal endpoints; the scientific outcome is UNKNOWN rather than negative.

The R8DA1 ACTUAL update-5000 checkpoint was not locally available, so its bytes and live hash could not be re-verified. State-digest checks are therefore unavailable, and equality to the recorded creation SHA-256 `e530c2b4356e7b09e899845d9501a56f4ad879353c3295744dac98fcfc0913d8` remains unknown. No other checkpoint was touched and no model evaluation was run during this check.

19 Supplementary descriptive audit

19.1 `LITERATURE_INSPIRED_REPAIR_SIGNATURE_AUDIT`

This frozen zero-update audit uses exactly three probe families. For a block output $X \in \mathbb{R}^{n \times d}$, token/representation similarity is

$$\mathrm{sim}(X) = \frac{\|\Pi_1 X\|_F^2}{\|X\|_F^2}, \qquad \Pi_1 = \frac{\mathbf{1}\mathbf{1}^\top}{n}.$$

It is computed per sequence at every layer and then averaged across sequences. Pre-normalization residual norm is $\sqrt{\mathrm{mean}_h\, y_h^2}$ for each token vector entering `ln1` or `ln2`, averaged across batch and tokens at each site and layer, without folding in learned LayerNorm affine parameters. Gradient contraction uses activation-gradient RMS, $\|\partial L/\partial X\|_F/\sqrt{\mathrm{numel}(X)}$, at the input and output boundary of every block; the audit reports all boundary values, adjacent backward ratios, and the boundary-0-to-boundary-L early/late ratio.

All probes use the same frozen validation array (SHA-256 `eab47bff512ac607b6afb449fa80063604472c3ac9ff89c06956e3dbc9d8df50`): the first 16 rows, columns 0–127 as inputs and 1–128 as targets, with mean next-token cross-entropy, `ignore_index=0`, evaluation mode, and dropout disabled. The R6 panel is seed 1 at state zero, the common completed-update-3 parent, ACTUAL and NEUTRAL at update 28, and their resident update-5000 endpoints. The R8-DA panel comprises each seed's common update-3 parent and ACTUAL/NEUTRAL update-28 release states when resident and readable; 11 states were included, while four unavailable checkpoint placeholders were excluded without requesting further local bytes.

Qualitatively, failed R6S1 states show near-one late-layer token similarity and severe early-gradient attenuation; the resident R8DA2 release pair separates in the same quantities and in pre-normalization residual profiles. R6S1 NEUTRAL nevertheless shows that a healthier release signature is not sufficient for persistent success. These are descriptive, literature-inspired associations only. WARM uses affine LayerNorm with $\epsilon = 10^{-5}$, whereas the closest comparison analyzes a simplified non-affine, epsilon-free RMSNorm block; architecture, intervention, and gradient/residual objects also differ. These normalization and object mismatches prohibit mechanism-validation, universal-collapse, or causal-mediator claims.

20 ADDITIONAL ENDPOINT TABLES

Table 28: R5 pilot endpoints; development evidence, not pooled with R6.

seed	branch	NLL	D_{16}	C
P1	ACTUAL	6.222020	7.189e-08	-0.0002
P1	NEUTRAL	4.576301	0.8254	1.1926
P1	ANGLE	6.222172	8.544e-08	-0.0000
P2	ACTUAL	6.221673	8.003e-08	0.0005
P2	NEUTRAL	4.268838	0.8526	1.8288
P2	ANGLE	6.221642	7.888e-08	0.0002
P3	ACTUAL	6.221949	8.201e-08	0.0000
P3	NEUTRAL	4.222483	0.8643	1.7188
P3	ANGLE	6.222024	1.906e-07	-0.0000

Table 29: R6 frozen prospective replication endpoints.

seed	branch	NLL	D_{16}	C
R6S1	ACTUAL	6.222221	8.753e-08	-0.0002
R6S1	NEUTRAL	6.223885	1.985e-07	-0.0000
R6S1	ANGLE	6.222226	0.01682	-0.0019
R6S1	WARMUP500	4.249132	0.8585	1.7718
R6S2	ACTUAL	6.222100	8.582e-08	-0.0009
R6S2	NEUTRAL	4.240685	0.8613	1.8822
R6S2	ANGLE	6.221817	9.928e-08	-0.0005
R6S2	WARMUP500	4.237302	0.8616	2.0061
R6S3	ACTUAL	6.222130	8.355e-08	0.0000
R6S3	NEUTRAL	4.231073	0.8579	1.8081
R6S3	ANGLE	6.222328	8.556e-08	-0.0004
R6S3	WARMUP500	4.237714	0.8596	1.5727

Table 30: R8 depth-transfer endpoints.

seed	branch	NLL	D_{16}	C
R8S1	ACTUAL	6.222033	8.685e-08	0.0000
R8S1	NEUTRAL	6.242355	9.05e-08	-0.0177
R8S1	ANGLE	6.222060	8.888e-08	0.0000
R8S1	WARMUP500	4.229280	0.8609	1.8817
R8S2	ACTUAL	6.222484	8.797e-08	0.0000
R8S2	NEUTRAL	6.222545	7.875e-08	0.1503
R8S2	ANGLE	6.222299	7.508e-08	-0.0002
R8S2	WARMUP500	4.230670	0.864	1.7808
R8S3	ACTUAL	6.221811	7.832e-08	0.0000
R8S3	NEUTRAL	6.221600	7.911e-08	0.0000
R8S3	ANGLE	6.221925	7.892e-08	0.0000
R8S3	WARMUP500	4.229924	0.8624	1.5510

Table 31: R9 lower-LR endpoints.

seed	branch	NLL	D_{16}	C
R9S1	ACTUAL	6.222094	8.303e-08	0.0000
R9S1	NEUTRAL	4.244900	0.8562	2.0729
R9S1	ANGLE	6.222019	8.632e-08	-0.0015
R9S1	WARMUP500	4.236458	0.8509	1.9841
R9S2	ACTUAL	6.221973	1.886e-07	-0.0000
R9S2	NEUTRAL	4.235148	0.8536	1.6660
R9S2	ANGLE	6.221921	9.299e-08	-0.0000
R9S2	WARMUP500	4.234187	0.8517	1.6544
R9S3	ACTUAL	6.221975	7.65e-08	0.0026
R9S3	NEUTRAL	4.238055	0.8541	1.7887
R9S3	ANGLE	6.221855	7.84e-08	-0.0126
R9S3	WARMUP500	4.238138	0.8506	1.7953

Table 32: P0 exact half-LR-pulse endpoints.

seed	NLL	D_{16}	C
R6S1	6.222262	8.138e-08	-0.0002
R6S2	6.222270	8.374e-08	-0.0004
R6S3	6.221932	8.145e-08	0.0000

Table 33: P1-R3 107M width-scale endpoints; all nine branches complete, zero replay.

seed	branch	NLL	D_{16}	C
P1S1	ACTUAL	6.233728	7.566e-08	0.0000
P1S1	NEUTRAL	6.234157	7.128e-08	0.0000
P1S1	WARMUP500	4.314842	0.861	1.7354
P1S2	ACTUAL	6.235699	8.99e-08	0.0000
P1S2	NEUTRAL	6.236288	7.971e-08	-0.0000
P1S2	WARMUP500	4.313156	0.8582	1.6574
P1S3	ACTUAL	6.234816	9.43e-08	-0.0000
P1S3	NEUTRAL	6.233819	9.086e-08	-0.0000
P1S3	WARMUP500	4.314874	0.8595	1.8016

Table 34: Qualified carrier-factorial endpoints. R6S1 is excluded because its NEUTRAL donor never achieves functional escape.

seed	corner	NLL	D_{16}	C	donor
R6S2	AA	6.222100	8.582e-08	-0.0009	qualified
R6S2	NN	4.240685	0.8613	1.8822	qualified
R6S2	AN	6.221839	7.749e-08	-0.0004	qualified
R6S2	NA	4.227688	0.863	1.9180	qualified
R6S3	AA	6.222130	8.355e-08	0.0000	qualified
R6S3	NN	4.231073	0.8579	1.8081	qualified
R6S3	AN	6.221885	7.652e-08	0.0002	qualified
R6S3	NA	4.223627	0.8593	1.8401	qualified

Table 35: R6S1 retained for provenance but not qualified for carrier-sufficiency interpretation.

seed	corner	NLL	D_{16}	C	donor
R6S1	AA	6.222221	8.753e-08	-0.0002	not qualified
R6S1	NN	6.223885	1.985e-07	-0.0000	not qualified
R6S1	AN	6.222225	6.843e-08	-0.0000	not qualified
R6S1	NA	6.256648	0.135	-0.0055	not qualified

Table 36: P2-R1 exact PRE/ACTUAL/NEUTRAL raw values.

seed	state	t	D	entropy	spectral	eff. rank
R6S1	PRE	3	0.018230	0.991886	0.057943	81.6557
R6S1	ACTUAL	28	0.000002	0.997203	0.278385	30.3957
R6S1	NEUTRAL	28	0.038285	0.965142	0.866727	2.1526
R6S2	PRE	3	0.023277	0.990819	0.076651	77.7785
R6S2	ACTUAL	28	0.000006	0.995294	0.439600	14.9056
R6S2	NEUTRAL	28	0.339676	0.814698	0.956586	1.3144
R6S3	PRE	3	0.024520	0.991580	0.062852	80.5013
R6S3	ACTUAL	28	0.000006	0.996635	0.377257	20.8557
R6S3	NEUTRAL	28	0.337019	0.688828	0.786294	2.3307

Table 37: Exact within-seed t3-to-t28 deltas; no threshold or predictor fit.

seed	branch	ΔD	Δ entropy	Δ spectral	Δ rank
R6S1	NEUTRAL	0.020055	-0.026744	0.808784	-79.5031
R6S2	NEUTRAL	0.316399	-0.176121	0.879935	-76.4641
R6S3	NEUTRAL	0.312499	-0.302752	0.723442	-78.1706

21 Reproducibility and provenance checklist

An external implementation must freeze: repository revision and file hashes; model and tied-weight semantics; dataset/tokenizer/vocabulary and split hashes; batch construction and sampler state; validation batch; precision and device; native AdamW semantics; schedule indexing; parameter-family order; diagnostic formula and layer; autograd/mode/RNG preservation; all construction tolerances and degeneracy rules; ANGLE seed and draw accounting; buffer handling; named optimizer-state mapping; cut-equivalence tests; and checkpoint hashes for model, optimizer, scheduler, sampler, RNG, and completed update. A run is excluded if any completed update is replayed or if checkpoint/provenance verification cannot establish the next legitimate update.

22 Code, data, and compute availability

An anonymized supplementary code archive contains the model definition, NEUTRAL and ANGLE solvers, scalar schedules, named optimizer-state transplantation and cut-equivalence checks, zero-update diagnostics, endpoint tables, and deterministic table/figure generation scripts. Checkpoints are omitted because of size; their recorded SHA-256 hashes are provided where available.

23 AI ASSISTANCE

Generative AI assisted literature discovery, hypothesis and design feedback, implementation, figures, result interpretation, drafting, editing, and adversarial review. The authors verified claims against source records, treated no AI output as scientific ground truth, and remain responsible.