CHROMAMUON: SPECTRAL OPTIMIZER DIAGNOSTICS AND DESIGN WITH RETAINED NEWTON–SCHULZ COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Practical Muon optimizers approximate a polar-like update with a short Newton–Schulz recurrence. Intermediate iterates preserve singular directions while applying distinct spectral responses. Building on prior iterate mixtures, ChromaMuon uses these responses as coordinates for optimizer diagnostics and design. We characterize coefficient-level controls for fitted maps, signed edits, and time-varying paths, and relate each response $f(x)$ to its signal-relative contribution profile $xf(x)$. For gradient inputs, this profile gives first-order modal descent up to scale. Matched language-model experiments show that finite-coordinate fits to bounded weak-tail targets worsen mean loss in the tested family, while static negative corrections have mean effects that depend on the base map. Under native cooldown, checkpoint continuations show a shrinking local advantage of deeper geometry; transient-edit probes show early loss gains followed by endpoint penalties. These observations motivate two representative paths that retain strong matched deep-control performance. On Modded-NanoGPT Track 3, their complete configurations first reach evaluated mean loss below 3.28 before the tested Muon-family references. Transferred to NanoChat without coefficient or schedule retuning, each configuration finishes below Tuned NS12 at all 40 paired endpoints across three separately analyzed token-budget protocols, spanning 537M–2.8B parameters. ChromaMuon connects spectral characterization, matched diagnostics, and complete-path evaluation for practical optimizer design.

1 INTRODUCTION

Muon reshapes the singular values of matrix momentum to approximate a polar-like update, using a short Newton–Schulz recurrence (Higham, 1986; 2008; Jordan et al., 2024b;a). Intermediate iterates preserve singular directions while applying different responses to normalized singular values. Finite iteration therefore selects a spectral shape as well as a numerical approximation. Understanding which features of that shape help training can guide the design of alternatives to the nearly flat Muon response.

Two features make this a diagnostic problem. Changing a power exponent alters the response across the spectrum, so a better final loss does not identify which modes account for the improvement. Timing adds another ambiguity: a map can help from an early checkpoint yet offer little benefit near the target, and an early advantage can be lost over the rest of training. We ask which spectral changes help under matched conditions, and whether their local or early benefits persist in a complete run.

Building on Soft-/Contra-Muon (Nilin, 2026), ChromaMuon retains $X_0, \ldots, X_K$ from one fixed recurrence and forms $\sum_k \alpha_k(t) X_k$. Coefficient identities describe changes to the response $f(x)$ and signal-relative contribution $xf(x)$ (Figure 1), without an explicit SVD. We compare fitted maps,

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.

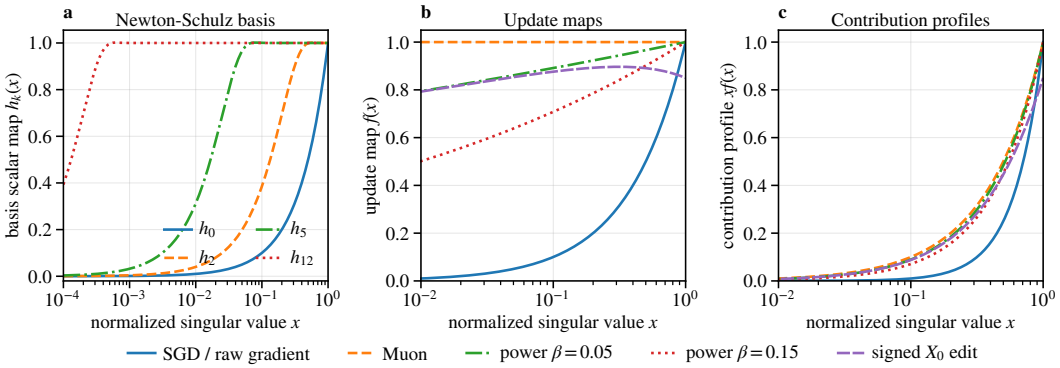

Figure 1: Retained responses as diagnostic instruments. (a) Analytic basis maps h_0, h_2, h_5, h_{12} from one stable Newton–Schulz recurrence over $x \in [10^{-4}, 1]$. (b–c) Five analytic maps are shown as update responses $f(x)$ and signal-relative contribution profiles $xf(x)$ over $x \in [10^{-2}, 1]$; colors match across panels. All panels use logarithmic x-axes and linear y-axes. SGD-like $f(x) = x$ and Muon-like $f(x) = 1$ anchor the power family. The signed example $f_{\text{edit}}(x) = x^{0.05} - 0.15x$ is raw, before norm matching: its linear term changes the top endpoint and slope while preserving the base map's leading weak-tail exponent.

signed edits, and their timing through static experiments, shared-checkpoint continuations, and complete runs in a fixed training stack.

- **Matched diagnostics.** All twelve tested tail-target fits worsen mean loss, negative-edit effects depend on the base, and transient edits can improve early loss yet worsen the endpoint.
- **Spectral characterization.** We derive endpoint and slope controls for the finite retained-response span and relate them to signal-relative contribution.
- **Training paths and transfer.** Two paths track the matched deep control closely, reach the Track 3 target before the tested external references, and transfer to NanoChat without spectral retuning.

Related spectral controls. Soft-Muon uses convex iterate mixtures and power-response approximations; Contra-Muon adds a signed normalized-momentum correction. Their account includes signal-weighted contributions and a Contra-to-Soft schedule (Nilin, 2026). Adaptive Schatten-Muon, Muon^p, Freon, and HTMuon study non-flat responses, while DynMuon schedules a scalar exponent (Massena et al., 2026; Dong & Sawin, 2026; Shumaylov et al., 2026; Pang et al., 2026; Wu et al., 2026). Our interventions probe effects that one exponent couples together. Polar Express optimizes recurrence coefficients (Amsel et al., 2025); we analyze retained outputs of a fixed recurrence. Other work studies Muon's scaling, norm geometry, adaptivity, and curvature corrections (Liu et al., 2025; Pethick et al., 2025; Chen et al., 2025; Si et al., 2025; Du & Su, 2026; Li et al., 2025; Dewulf et al., 2026).

2 Retained Newton–Schulz Iterates as a Spectral Design Space

2.1 Coordinate Representation

Let $Z \neq 0$ be the matrix presented to the recurrence after any upstream preprocessing, and let G be the current gradient. Locally, the host preconditions MLP and attention-value momentum with SOAP; NanoChat instead equilibrates rows (Appendix E). Write

$$Z = U \operatorname{diag}(\sigma_i) V^\top, \qquad x_i = \sigma_i / s(Z) \in [0, 1],$$

where the reported normalization is

$$s(Z) = \|ZZ^\top\|_F^{1/2} = \|Z\|_{S_4} \geq \sigma_{\max}(Z).$$

We work on the active singular subspace; null-space modes receive zero response. Let K be the maximum retained dictionary depth.

Proposition 1 (Retained-coordinate structure)**.** *Let $X_0 = Z/s(Z)$ and apply the reported recurrence*

$$X_{k+1} = 2X_k - \frac{3}{2}(X_k X_k^\top)X_k + \frac{1}{2}(X_k X_k^\top)^2 X_k.$$

Then, in exact arithmetic,

$$X_k(Z) = U\,\mathrm{diag}(h_k(x_i))V^\top.$$

Here $h_0(x) = x$ and

$$h_{k+1}(x) = 2h_k(x) - \frac{3}{2}h_k(x)^3 + \frac{1}{2}h_k(x)^5.$$

For coefficients $\alpha = (\alpha_0, \ldots, \alpha_K)$,

$$\begin{aligned} D_\alpha^{\mathrm{raw}}(Z) &= \sum_{k=0}^{K} \alpha_k X_k(Z) = U\,\mathrm{diag}(f_\alpha(x_i))V^\top, \\ f_\alpha(x) &= \sum_{k=0}^{K} \alpha_k h_k(x). \end{aligned}$$

For this recurrence,

$$f_\alpha(1) = \sum_{k=0}^{K} \alpha_k, \qquad f'_\alpha(0) = \sum_{k=0}^{K} 2^k \alpha_k, \qquad f'_\alpha(1) = \alpha_0.$$

Moreover, $\deg h_k = 5^k$, so $h_0, \ldots, h_K$ are linearly independent and form a basis for their finite span.

Induction gives the shared singular directions; the scalar recurrence gives the endpoint identities and distinct degrees (Appendix D.1.1). Functional independence does not guarantee numerical conditioning or independent matrices on one spectrum. One-hot coefficients select depth; general coefficients give interpolation, fitted maps, signed edits, and paths. These identities concern the raw direction. Scalar norm matching preserves its direction; non-scalar wrappers can change it. Host preprocessing may use accumulated curvature information (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2024).

2.2 Contribution Profiles

For any response f, direct expansion in the singular vectors gives

$$\left\langle G, D_f^{\mathrm{raw}}(Z) \right\rangle_F = \sum_i (u_i^\top G v_i) f(x_i).$$

If the transformed signal is the gradient, $Z = G$, then

$$\left\langle G, D_f^{\mathrm{raw}}(G) \right\rangle_F = s(G) \sum_i x_i f(x_i).$$

We therefore call $C_f(x) = x f(x)$ the *signal-relative contribution profile*. For $Z = G$, it decomposes signed first-order modal descent up to the common scale $s(G)$. For momentum, true descent instead depends on $(u_i^\top G v_i) f(x_i)$, so C_f is only signal-relative and gradient–momentum alignment remains essential. Raw-gradient-like response has $C(x) = x^2$, while ideal Muon has $C(x) = x$. Under a common top-end normalization $f(1) = 1$, increasing contribution at a fixed weak mode relative to the top endpoint requires increasing its response $f(x)$.

2.3 Partially Decoupled Signed Control

A signed retained-X_0 edit adds cX_0 to the raw direction, giving

$$\widetilde{f}(x) = f_\alpha(x) + cx, \qquad \widetilde{C}(x) = C_\alpha(x) + cx^2.$$

Negative c suppresses the top-weighted x^2 component without directly constructing a weak-tail target.

For an analytic power base x^β, $\beta < 1$, a nonzero cx edit preserves its leading weak-tail exponent but changes its endpoint and slope; it cannot be another scaled power (Proposition A3 in Appendix D.1.1). Finite retained fits instead satisfy $f_\alpha(x) = f'_\alpha(0)x + O(x^3)$: their power labels describe targets on $[10^{-4}, 1]$, not exact asymptotics at zero. The control is partially decoupled: cx affects the full spectrum and norm matching couples scale.

2.4 Coefficient Paths and Selected Depth

A coefficient path is $\alpha(t) = (\alpha_0(t), \ldots, \alpha_K(t))$, with

$$D_t^{\mathrm{raw}}(Z_t) = \sum_{k=0}^{K} \alpha_k(t) X_k(Z_t).$$

Let e_k select X_k. For $q \in [0, K]$, set $d(q) = e_q$ when q is an integer. Otherwise, with $m = \lfloor q \rfloor$ and $r = q - m$, define $d(q) = (1 - r)e_m + re_{m+1}$. Thus $q(t)$ is the selected or interpolated depth. For $q \geq 1$, every $d(q)$ preserves $f(1) = 1$ and $f'(1) = 0$, while varying $f'(0)$ and the interior response. An e_0 correction instead changes both endpoint and slope. These identities distinguish the probes; the reference endpoint $x = 1$ need not occur under Schatten-4 normalization.

We use *coefficient path* for the mathematical object $\alpha(t)$ and *configuration* for its complete optimizer realization within the surrounding training stack. Both reported ChromaMuon configurations use $K = 12$ and compute $X_0, \ldots, X_{12}$ at every optimizer step. Changing $q(t)$ therefore selects a different response within the same retained dictionary. Signed edits are norm-matched to their pre-edit base; the fitted pulse is part of that base (Appendix E).

3 From Spectral Diagnostics to Phase-Aware Design

3.1 Setup

All controlled experiments use the Modded-NanoGPT-style FineWeb10B setup (Jordan et al., 2024a): a 12-layer decoder-only Transformer with model dimension 768, sequence length 1024, and a global batch of 524,288 tokens. Each validation evaluates 10,485,760 tokens. Matrix parameters inside Transformer blocks use the tested Muon or ChromaMuon map with momentum 0.95 and matrix learning rate 0.0375; embeddings, the output projection, and one-dimensional parameters use a common auxiliary AdamW stack (Loshchilov & Hutter, 2019). Within each controlled comparison, the retained-coefficient map is the changed factor and the surrounding training system is held fixed.

Static-map experiments run for 6000 steps under a horizon-adjusted cooldown. Complete schedule runs use 3040 steps; checkpoint continuations follow the native cooldown, whose rates reach zero at step 3105. In the power and bounded-target sweeps, "static" denotes the time-independent fitted base response: every row shares the same early fading X_0 envelope, leaving that response as the only between-row factor. All controlled settings in Figures 2 and 3 use four seeds. Seed-matched designs are analyzed through paired differences: independent-setting error bars are sample SDs, whereas paired contrasts use the sample SD of seed-level differences. All small-sample comparisons are descriptive rather than formal significance claims; checkpoints, envelopes, and token-budget protocols are not independent replications. We express small validation-loss differences in milli-loss, where one milli-loss is 10^{-3} validation-loss units.

The official Track 3 target is validation loss strictly below 3.28. We report the first evaluated crossing that remains below this target under the rounding model in Appendix B.4; no interpolation is used. In schedule comparisons, the deep and shallow controls fix $q(t) = 12$ and $q(t) = 5$, respectively, while retaining the fading X_0 envelope used by Depth–X_0. This internal deep control differs from the external Tuned NS12 comparator. Plain fixed-k labels denote controls without that envelope; plain $k = 5$ is a deliberately separated diagnostic anchor.

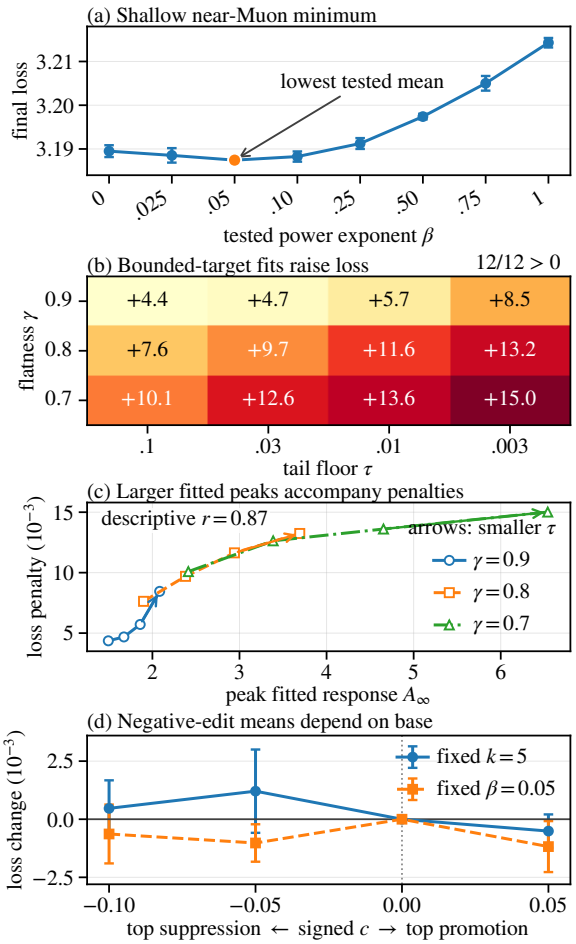

Figure 2: Static spectral probes. Experiments (a–c) use static bases under a shared fading X_0 correction. (a) Four-seed final loss across categorical tested exponents; the arrow marks the lowest mean and bars are sample SDs. (b) Mean loss penalty $10^3[L_{\gamma,\tau} - L_{\gamma=1}]$ for twelve finite-coordinate fits to bounded weak-tail targets; positive values are worse. (c) The same penalty against peak fitted response $A_\infty = \max_{x \in \mathcal{G}} f_\alpha(x)$, evaluated on the fixed normalized grid with $f_\alpha(1) = 1$; colors denote γ, arrowheads point toward smaller τ, and r is the descriptive Pearson correlation across the twelve fitted targets. (d) Paired loss change $10^3[L_c - L_{c=0}]$ for constant signed-X_0 edits on two base maps; negative values favor the edit and bars are paired sample SDs.

3.2 Static Power Maps Identify a Near-Muon Basin

Power-response sweep. Across rows, we vary the target x^β of the static fitted base over $\beta \in \{0, 0.025, 0.05, 0.10, 0.25, 0.50, 0.75, 1\}$, where $\beta = 0$ is the Muon-like endpoint. The best tested mean occurs at $\beta = 0.05$, with final validation loss 3.18744 ± 0.00052, versus 3.18950 ± 0.00135 at $\beta = 0$, a 2.06-milli-loss reduction; larger moves toward the gradient-like $\beta = 1$ endpoint are progressively worse (Figure 2a). The lowest means lie near the Muon-like endpoint. Four seeds do not identify a unique best exponent.

The exponent changes response across the spectrum. Figure A2 documents this coupling; the sweep does not isolate which part accounts for the loss improvement.

3.3 Static Probes Test Tail and Top-Side Interventions

Promotion beyond Muon. For $0 < x < 1$, the favorable power target $x^{0.05} < 1$ reduces response relative to ideal Muon. To test promotion beyond Muon, we instead fit

$f_{\gamma,\tau}(x) = ((x+\tau)/(1+\tau))^{\gamma-1}$. Each fit uses 512 logarithmically spaced points over $[10^{-4}, 1]$ and is evaluated over that interval. We compare twelve (γ, τ) targets with $\gamma < 1$ against the $\gamma = 1$ control. Let $f_\alpha(x) = \sum_k \alpha_k h_k(x)$ be the fitted map, normalized by $f_\alpha(1) = 1$. We call $A_\infty = \max_{x\in\mathcal{G}} f_\alpha(x)$ its peak fitted response, where $\mathcal{G}$ is a fixed 2048-point logarithmic grid over $[10^{-4}, 1]$. Thus A_∞ is neither the analytic target supremum nor a maximum observed from training matrices. Every intervention worsens four-seed mean final loss, by 4.36 to 15.01 milli-loss, while A_∞ ranges from 1.49 to 6.54. Across this designed grid, penalty and A_∞ co-vary (descriptive Pearson $r = 0.873$; Figure 2b–c). Stronger tail targets therefore give no monotone improvement in this fitting-and-training pipeline. The fits violate analytic upper bounds and sometimes $f \geq 1$ (Appendix Table A14), so these penalties cannot separate intended reweighting from approximation effects or explain the power-sweep gain.

Static signed corrections. We next hold either fixed $k = 5$ or fixed $\beta = 0.05$ as the base map and change only a constant coefficient cX_0 over the full 6000-step run. This adds cx to the response and cx^2 to the contribution profile, so $c < 0$ suppresses the top side while $c > 0$ promotes it. On the $k = 5$ base, $c = +0.05$ changes paired final loss by -0.510 ± 0.717 milli-loss; on the power base, $c = -0.05$ and $c = +0.05$ give -1.023 ± 0.806 and -1.178 ± 1.097 milli-loss, respectively (Figure 2d). The negative edit instead worsens mean loss on $k = 5$; $c = +0.05$ has the lowest tested mean on both bases. The negative edit's mean effect is thus base-dependent, while these four-seed contrasts do not establish a universal preferred sign. We next probe timing.

3.4 Timing Diagnostics and Complete Training Runs

Checkpoint continuations. We branch checkpoints from Depth–X_0 into paired 300-step continuations, disabling the signed edit in all branches and varying only the spectral map. Relative to plain fixed $k = 5$, the paired step-averaged continuation-loss reduction from selected $k = 12$ is 17.879 ± 0.544 milli-loss at checkpoint 1000 and 6.778 ± 0.047 at checkpoint 2000, but contracts to 0.184 ± 0.008 at 2875 and -0.039 ± 0.010 at 3010 (Figure 3a). A fixed power-$\beta = 0.05$ map and the independently specified bounded-target fit show the same early-to-late contraction. All parameter-group rates clamp to zero at step 3105; the last two branches have 230 and 95 nonzero-rate updates. This measures map value under the native cooldown, without separating phase from learning-rate exposure or establishing a better shallow endpoint.

A separate experiment measures the accumulated effects of transient signed edits. Holding the selected-depth schedule and training stack fixed, we compare twelve negative X_0 envelopes with paired no-X_0 controls. Averaged across envelopes, the edit reduces loss by 3.035 milli-loss over steps 125–500 and by 1.488 over steps 125–1000, but increases it by 0.827 over steps 1000–2000 and by 0.789 at step 3040 (Figure 3b). These are effects of whole training paths: every envelope is zero by 2500, so the final gaps reflect earlier trajectory changes. Transience thus does not guarantee an endpoint gain. This experiment neither isolates late application nor tests a matched positive transient.

Two training paths. The continuation results favor deep maps early under cooldown, while the transient-edit results expose a tradeoff between early loss and the endpoint. We evaluate two ways to combine these controls over a full run (Figure 3c):

$$\begin{aligned} \alpha_{X_0}(t) &= d(q(t)) + c_{X_0}(t)e_0, \\ \alpha_{\mathrm{FP}}(t) &= [1 - w_{.025}(t)]d(q(t)) + w_{.025}(t)p_{.025}, \end{aligned}$$

where $p_{.025}$ is the fixed $K = 12$ coefficient vector fitted to $x^{0.025}$. Both retain selected depth $q(t) = 12$ through step 2500 and interpolate to $q(t) = 5$ by 3010. Depth–X_0 fades its signed top-side coefficient $c_{X_0}(t)$ from -0.2 to zero by step 2000. Depth–Fitted-Power instead sets that coefficient to zero, ramps the fitted-map weight $w_{.025}(t)$ from zero to 0.25 over steps 500–1000, holds it through 1500, and removes it by 2000. Thus one path uses a transient signed correction and the other a mild fitted-response pulse. The diagnostics do not determine their constants or breakpoints; complete runs must test these choices.

Matched complete runs compare both constructions with the deep fixed-depth control. The Depth–X_0 contrast primarily tests late depth contraction because the control shares its early X_0 fade; the Depth–Fitted-Power contrast instead compares complete configurations. Both track the deep control

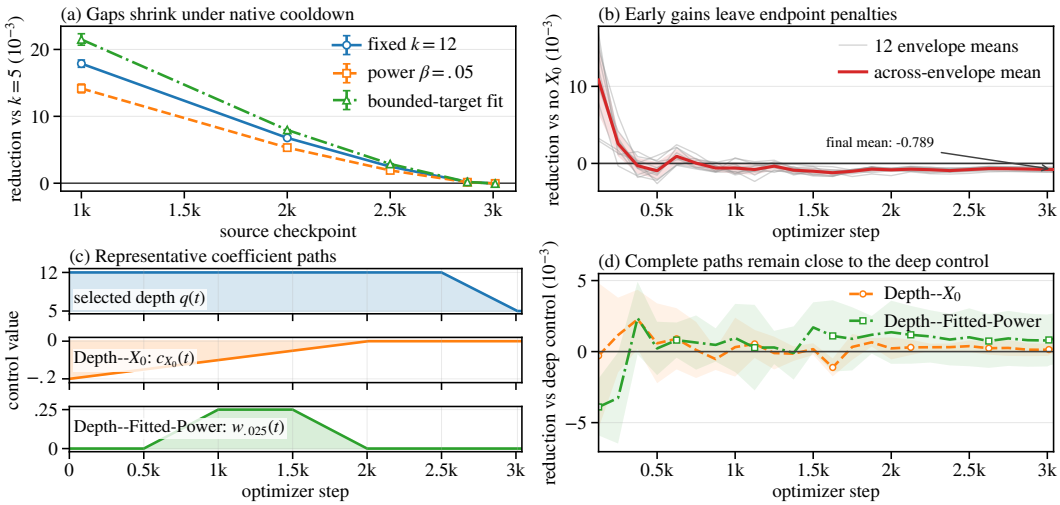

Figure 3: Spectral measurements and representative paths. Loss reductions use $\Delta L = L_{\text{reference}} - L_{\text{tested}}$, so positive values favor the tested map or configuration and negative values favor its named reference. (a) Four-seed paired 300-step continuation reductions for three maps relative to plain fixed $k = 5$, placed at their source checkpoints under the native cooldown; bars are paired sample SDs. (b) Time-resolved reductions for twelve negative-X_0 envelopes relative to their paired no-X_0 controls. Faint curves are four-seed paired means, red is the mean across the twelve envelopes, and the band is their descriptive SD. These trajectory effects include the history of each edit. (c) Representative selected depth $q(t)$, signed coefficient $c_{X_0}(t)$, and fitted-map weight $w_{.025}(t)$ define Depth–X_0 and Depth–Fitted-Power. (d) Four-seed paired complete-run reductions at 28 checkpoints relative to the deep control; bands are paired sample SDs. The deep control fixes $q(t) = 12$ while sharing Depth–X_0's fading envelope. Depth–Fitted-Power omits that edit, making its contrast a complete-configuration comparison. The diagnostics distinguish local map effects from training-path effects; they do not select optimal schedule constants or breakpoints.

closely (Figure 3d). A four-seed schedule ablation reported in Appendix Table A8 gives a 1.025-milli-loss target-window penalty for an abrupt $12 \to 5$ switch, while the smooth depth schedule's paired target-window mean differs from fixed $k = 12$ by only 0.033 milli-loss (paired SD 0.466). The linear transition differs from the abrupt switch in duration and depth exposure, so this comparison cannot isolate smoothness. In this same ablation, disabling X_0 gives a slightly lower final mean; the early signed-edit benefit therefore does not establish an endpoint gain. A five-rate sweep also keeps both paths near the deep control, with their best tested means at the reference rate (Appendix Table A13).

3.5 Target Crossing on Modded-NanoGPT Track 3

We also evaluate Tuned NS12, Newton–Muon, NorMuon, and Aurora locally (Du & Su, 2026; Li et al., 2025; Dewulf et al., 2026). Under the rounding criterion in Appendix B.4, we conservatively report step 3020 for both Depth–Fitted-Power and Depth–X_0, 155 optimizer steps (4.9%) before Aurora, the nearest tested external reference at step 3175. NorMuon, Newton–Muon, and Tuned NS12 cross at 3250, 3275, and 3325, giving leads of up to 305 steps (9.2%). Percentages use each reference's reported crossing as the denominator. NorMuon has the lowest displayed endpoint mean; endpoint loss and first evaluated target crossing are distinct metrics (Table 1; Figure 4a–b).

These aggregate comparisons are descriptive, not paired tests: grids, run counts, horizons, and seed identities differ. The matched deep control crosses at 3030 under the shared fading edit and training shell; that comparison does not enter the reference percentages. External Tuned NS12 settings are only partly recoverable (Appendix E.3).

Table 1: Track 3 target crossing: first evaluated checkpoint below 3.28 under the rounding model in Appendix B.4, without interpolation. Endpoint loss is mean ± sample SD at the listed final step; n is run count and q selected depth. The appendix also gives the original stored-record crossings and their precision audit.

Method	Reported crossing	Final step	Endpoint loss	n
Constructed configurations				
Depth–Fitted-Power	3020	3040	$3.27824 \pm .00112$	24
Depth–X_0	3020	3040	$3.27851 \pm .00120$	20
Internal geometry controls				
Deep control ($q = 12$)	3030	3040	$3.27915 \pm .00181$	4
Shallow control ($q = 5$)	–	3040	$3.28529 \pm .00263$	4
Muon-family references				
Aurora	3175	3200	$3.27830 \pm .00060$	12
NorMuon	3250	3300	$3.27733 \pm .00130$	12
Newton–Muon	3275	3300	$3.27860 \pm .00122$	12
Tuned NS12	3325	3350	$3.27824 \pm .00065$	12

4 Transfer to NanoChat without Retuning

We transfer both ChromaMuon configurations to NanoChat (Karpathy, 2025) without spectral retuning. Schedules use source time $t_{\mathrm{src}} = 3040t/T$, with $T = 3000$. Model, data, evaluation, and auxiliary optimizer settings match the transferred Tuned NS12 comparator, but matrix-update wrappers differ: ChromaMuon retains equilibration, renormalization, variance reduction, and cautious decay; Tuned NS12 uses direct updates. This tests transfer of complete implementations and cannot attribute gains to spectral paths alone (Appendix E).

We evaluate D16–D32 models (537M–2.82B parameters) for 3,000 optimizer steps at sequence length 2,048, using four paired seeds per depth. The depth-specific batch protocol validates every 100 steps over 8,388,608 tokens; comparisons are paired within depth and protocol. Appendix B.5 gives batches, tokenizer provenance, and evaluation protocols.

At step 3000, each configuration wins 4/4 paired seeds at every depth, or all 16 depth-specific-batch endpoints. Paired advantages span 5.818–9.606 milli-BPB (Table 2).

Table 2: Paired NanoChat validation at optimizer step 3000. Tuned NS12 is absolute BPB (mean ± sample SD). $\Delta = 10^3(\mathrm{BPB}_{\text{Tuned NS12}} - \mathrm{BPB}_{\text{configuration}})$ is paired mean ± sample SD over four seed differences, in milli-BPB; positive values favor ChromaMuon. Each configuration wins 4/4 seeds per depth.

Model	Tuned NS12	Δ Depth–X_0	Δ Fitted-Power
D16	0.80749 ± 0.00298	9.545 ± 2.489	9.606 ± 2.673
D20	0.75737 ± 0.00377	7.794 ± 3.428	7.953 ± 3.623
D24	0.73963 ± 0.00261	5.824 ± 2.465	5.818 ± 2.573
D32	0.69114 ± 0.00043	8.385 ± 0.744	8.310 ± 0.642

At 196.6M and 786.4M tokens, paired advantages are 6.106–8.791 and 4.009–7.902 milli-BPB, respectively. Each configuration wins 12/12 paired endpoints in each fixed-budget protocol. Across ten model–protocol settings, both improve all four paired endpoints: 40/40 is a consistency count, not 40 independent replications. Appendix B.5 gives trajectories and endpoint summaries; the code/data ZIP contains the underlying records.

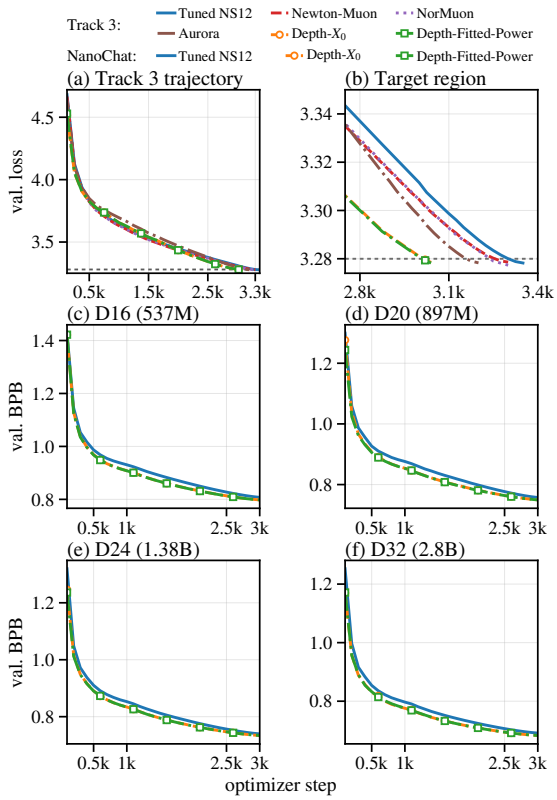

Figure 4: Track 3 and NanoChat validation. (a–b) Stored mean losses on method-specific grids; Table 1 applies conservative rounding at target 3.28. (c–f) Both configurations beat Tuned NS12 at all 30 post-initial checkpoints per depth (mean BPB, four paired seeds). Matrix wrappers differ.

5 Discussion: What the Diagnostics Establish

Spectral shape and training time pose distinct design questions. The power sweep favors near-Muon behavior, yet all twelve bounded-tail fits worsen mean loss, and signed corrections depend on the base. Analytic targets therefore need testing through their realized finite-coordinate responses. Under native cooldown, deeper maps lose much of their local advantage; transient edits can improve early loss yet incur endpoint penalties.

Inspecting $f(x)$ and $xf(x)$, comparing interventions from shared checkpoints, and testing complete paths address the realized map, its effect at a given state, and its accumulated effect on training. Our two paths retain matched deep-control performance and transfer without spectral retuning; a substantial schedule gain over that control remains unestablished.

The identities concern raw maps; $xf(x)$ alone does not determine descent under momentum. NanoChat compares complete configurations with different wrappers. Training horizons are short; both paths compute all $K = 12$ iterates, so earlier crossings establish no wall-clock advantage.

6 Conclusion

ChromaMuon links interpretable spectral controls to matched training experiments. The diagnostics expose base-dependent effects and a separation between local or early gains and final outcomes. Two paths preserve strong deep-control performance, reach the Track 3 target before tested external references, and transfer without spectral retuning. Retained Newton–Schulz coordinates thus support both controlled investigation and evaluation of complete spectral optimizer designs.

AI USE STATEMENT

Generative AI tools assisted with manuscript revision; retrieval, discovery, and comparison of related work; and research ideation and analysis, including framing research questions, interpreting experimental findings, and auditing preserved results. AI assistance also supported preparation of submission artifacts. During revision, AI-assisted checks included comparison with the source manuscript and implementation, verification of reported statistics against preserved records and plotted data, recomputation of packaged result summaries, and compilation and inspection of the PDF. These revision checks did not rerun the training experiments. Responsibility for the submitted text, claims, and artifacts rests with the authors.

REPRODUCIBILITY STATEMENT

Appendices C and E describe the training protocols, optimizer pipeline, fitting procedure, and complete coefficient schedules; Appendix D gives assumptions and proofs. Appendix B records the extended experimental boards and comparison boundaries. All appendices are included in this PDF. The accompanying `ChromaMuon_ICLR2027_code_data.zip`, rooted at `ChromaMuon_code_data/`, contains optimizer implementations, host integration and command generators, paper-facing summaries and validation curves, and audit scripts for the reported Track 3 crossings and NanoChat paired comparisons. It contains NanoChat seed endpoints, continuation validation curves, and temporal signed-edit records, but is not a complete archive of original training logs, checkpoints, or all diagnostic coefficient arrays. Full training also requires the public host projects and datasets described in its `README.md`; all package paths in this paper are relative to that archive root.

REFERENCES

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025. URL https://arxiv.org/abs/2505.16932.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025. doi: 10.48550/arXiv.2506.15054. URL https://arxiv.org/abs/2506.15054.

Alec Dewulf, Dhruv Pai, Li Yang, Ashley Zhang, and Ben Keigwin. Aurora: A leverage-aware spectral optimizer. *arXiv preprint arXiv:2606.27715*, 2026. URL https://arxiv.org/abs/2606.27715.

Yihe Dong and Will Sawin. Muon^p: Muon with fractional spectral powers. *arXiv preprint arXiv:2606.13867*, 2026.

Zhehang Du and Weijie Su. The newton–muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026. doi: 10.48550/arXiv.2604.01472. URL https://arxiv.org/abs/2604.01472.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, volume 80, pp. 1842–1850, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079. URL https://doi.org/10.1137/0907079.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, 2008. doi: 10.1137/1.9780898717778. URL https://doi.org/10.1137/1.9780898717778.

Keller Jordan, Jeremy Bernstein, Bo Rappazzo, et al. modded-nanogpt: Speedrunning the NanoGPT baseline. https://github.com/KellerJordan/modded-nanogpt, 2024a.

Keller Jordan, Yuchen Jin, Vladimir Boza, Jiacheng You, Franz Louis Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon, 2024b.

Andrej Karpathy. nanochat: The best ChatGPT that $100 can buy. https://github.com/karpathy/nanochat, 2025. URL https://github.com/karpathy/nanochat.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025. URL https://arxiv.org/abs/2510.05491.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. doi: 10.48550/arXiv.2502.16982. URL https://arxiv.org/abs/2502.16982.

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

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, volume 37, pp. 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Thomas Massena, Corentin Friedrich, and Mathieu Serrurier. From SGD to Muon: Adaptive optimization via Schatten-p norms. *arXiv preprint arXiv:2605.19781*, 2026. doi: 10.48550/arXiv.2605.19781. URL https://arxiv.org/abs/2605.19781.

Nilin. Contra-muon and soft-muon. https://nilin.github.io/contra-muon-and-soft-muon/, 2026. Non-archival blog post; first version May 4, revised May 14; accessed 2026-07-22.

Tianyu Pang, Yujie Fang, Zihang Liu, Shenyang Deng, Lei Hsiung, Shuhua Yu, and Yaoqing Yang. HTMuon: Improving muon via heavy-tailed spectral correction. *arXiv preprint arXiv:2603.10067*, 2026. URL https://arxiv.org/abs/2603.10067.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. *arXiv preprint arXiv:2502.07529*, 2025. URL https://arxiv.org/abs/2502.07529.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026. doi: 10.48550/arXiv.2605.11181. URL https://arxiv.org/abs/2605.11181.

Chongjie Si, Debing Zhang, and Wei Shen. AdaMuon: Adaptive Muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025. URL https://arxiv.org/abs/2507.11005.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. *arXiv preprint arXiv:2409.11321*, 2024. URL https://arxiv.org/abs/2409.11321.

Fangzhou Wu, Rikhav Shah, Sandeep Silwal, and Qiuyi Zhang. DynMuon: A dynamic spectral shaping view of Muon. *arXiv preprint arXiv:2605.17109*, 2026.

A GUIDE TO THE SUPPLEMENT

Table A1 summarizes the experiments and their scope. The remaining sections give the complete results, protocols, proofs, implementation, and exact schedules.

Table A1: Summary of the diagnostic evidence. Schedule constants are empirical choices, assessed in complete training runs.

Experiment	Finding and scope
Coordinate capacity	$K = 12$ reduces observed-spectrum fit error and supports non-power fits. Compute-optimal depth is untested (Figure A1; Table A2).
Power responses	The lowest mean loss occurs at $\beta = .05$, near the Muon-like endpoint. This sweep does not select the downstream .025 pulse or its timing (Figure A2; Table A3).
Weak-tail fits	All 12 fits worsen mean loss. Approximation errors prevent attribution of these penalties to the target reweighting alone (Tables A4, A14).
Static signed edits	The mean effect of $c = -.05$ changes sign across the two bases; $c = +.05$ has the lowest tested mean on both (Table A5).
Timing	The local deep-map advantage contracts under cooldown. Transient signed edits give early gains but leave endpoint penalties after they vanish (Tables A6, A7).
Depth transition	The linear transition tracks the deep control closely; the abrupt switch is worse. Duration and depth exposure also differ, so smoothness is not isolated (Table A8).
Complete runs	Both configurations cross earlier than the external Track 3 references and improve all 40 paired NanoChat endpoints. Differing NanoChat wrappers prevent attribution to spectral paths alone (Tables A9–A12).
Learning rate	Both paths have their lowest tested mean at the reference rate and remain near the deep control across five rates (Figure A4; Table A13).

A.1 NOTATION AND COMPARISON BOUNDARIES

Here K sets dictionary capacity and $q(t)$ selects a member. In experimental labels, fixed $k = r$ means $q(t) \equiv r$; uppercase K denotes maximum retained dictionary depth, while lowercase k otherwise indexes recurrence iterates. The reported configurations are Depth–X_0 and Depth–Fitted-Power. Tuned NS12 denotes the Track 3-tuned Muon comparator whose matrix map is the terminal output of a fixed 12-step recurrence; NanoChat transfers that map into its host optimizer stack. Deep and shallow controls instead fix the retained-coordinate construction at $q(t) = 12$ and $q(t) = 5$ under the fading envelope; plain controls omit it.

Benefit $B(m; r) = \mathcal{L}(r) - \mathcal{L}(m)$ is positive when m has lower loss; NanoChat analogously uses $\Delta_{\mathrm{BPB}} = \mathrm{BPB}_{\text{Tuned NS12}} - \mathrm{BPB}_{\text{configuration}}$. One milli-loss is 10^{-3} loss units. Scalar norm matching preserves relative spectral shape, but non-scalar wrappers may alter directions; continuations are local.

B EXTENDED EXPERIMENTAL EVIDENCE

B.1 COORDINATE APPROXIMATION AND RETAINED-DEPTH CAPACITY

Figure A1 illustrates finite-coordinate approximation of the two power targets used downstream and a smooth non-power target $0.5x^{0.25} + 0.5x^{0.75}$. On the plotted 900-point logarithmic grid, the normalized $K = 12$ non-power fit has RMSE 0.00330, establishing finite non-power expressivity alongside the fitted power maps.

Maximum retained depth K controls dictionary capacity, whereas selected depth $q(t)$ selects a response inside that dictionary. They are not the same intervention.

Table A2: Basis-depth continuation audit. Entries are four-seed mean final loss; parentheses are benefit relative to $K = 12$, $B(K; K{=}12) = L(K{=}12) - L(K)$, so positive means lower loss.

Source	$K = 5$	$K = 7$	$K = 9$	$K = 12$
1000	3.59280 (-.00712)	3.58685 (-.00117)	3.58510 (+.00059)	3.58568
2500	3.30028 (-.00029)	3.29994 (+.00005)	3.29997 (+.00002)	3.29999
2875	3.27744 (+.00006)	3.27750 (0.00000)	3.27751 (0.00000)	3.27750

For archived spectrum s, normalized singular values $x_{s,i}$, target g, and fitted response f_K, the observed-spectrum weighted response RMSE is

$$E_s(K) = \left[\sum_i w_{s,i} \big(f_K(x_{s,i}) - g(x_{s,i}) \big)^2 \right]^{1/2},$$
$$w_{s,i} = \frac{x_{s,i}}{\sum_j x_{s,j}}.$$

These weights emphasize larger singular values; this error does not certify uniform weak-tail accuracy or coverage below the fitting cutoff.

The visible low-x oscillations in Figure A1(a) motivate the archived-spectrum audit in panel (c), which averages $E_s(K)$ across 54 spectra: six checkpoints times nine matrices. For the fitted $x^{0.025}$ target, this mean falls from 0.26439 at $K = 5$, to 0.12209 at 7, 0.04777 at 9, and 0.00478 at 12. An offline-only $K = 16$ fit reaches 0.00023; no reported training run uses $K = 16$.

For each K, the $x^{0.025}$ target is independently fitted and normalized in $h_0, \ldots, h_K$; paired continuation branches otherwise restore the same source state and host settings. Table A2 reports the 300-step continuation audit. At source step 1,000, $K = 12$ reduces final loss by 0.00712 relative to $K = 5$; by source steps 2,500 and 2,875, all tested dictionaries are separated by small mean gaps. Together, the fit and continuation audits support retaining the expressive $K = 12$ dictionary while treating selected depth $q(t)$ as a separate phase-dependent control; $K = 7$ and $K = 9$ already capture most of the measured late continuation behavior.

B.2 Static Spectral Interventions

B.2.1 Fixed Power Maps

Table A3 reports the full fixed-map board, while Figure A2 connects its training trajectories to weak-response and contribution-support diagnostics. Let x_j be 2,048 logarithmically spaced points on $[10^{-4}, 1]$, and define normalized positive contribution mass

$$p_j = \frac{[x_j f(x_j)]_+}{\max\{\sum_\ell [x_\ell f(x_\ell)]_+, 10^{-12}\}}.$$

The mean weak-mode response over $x_j < 0.01$ is

$$\bar{f}_{\text{weak}} = \frac{1}{|\{j : x_j < 0.01\}|} \sum_{j : x_j < 0.01} f(x_j),$$

before norm matching. Entropy effective support is

$$N_{\text{eff}} = \exp\left[-\sum_j p_j \log \max\{p_j, 10^{-12}\} \right].$$

Larger N_{eff} means that normalized contribution mass is spread across more points on the fixed grid. It is a grid-dependent concentration statistic, not matrix rank.

The lowest tested mean occurs at $\beta = 0.05$, 2.06 milli-loss below the Muon-like $\beta = 0$ endpoint, with nearby exponents forming a local near-Muon basin. Effective support contracts monotonically as β increases. The exponent therefore acts as a coupled top-to-tail allocation axis, motivating interventions that separate which part of that coupling is useful.

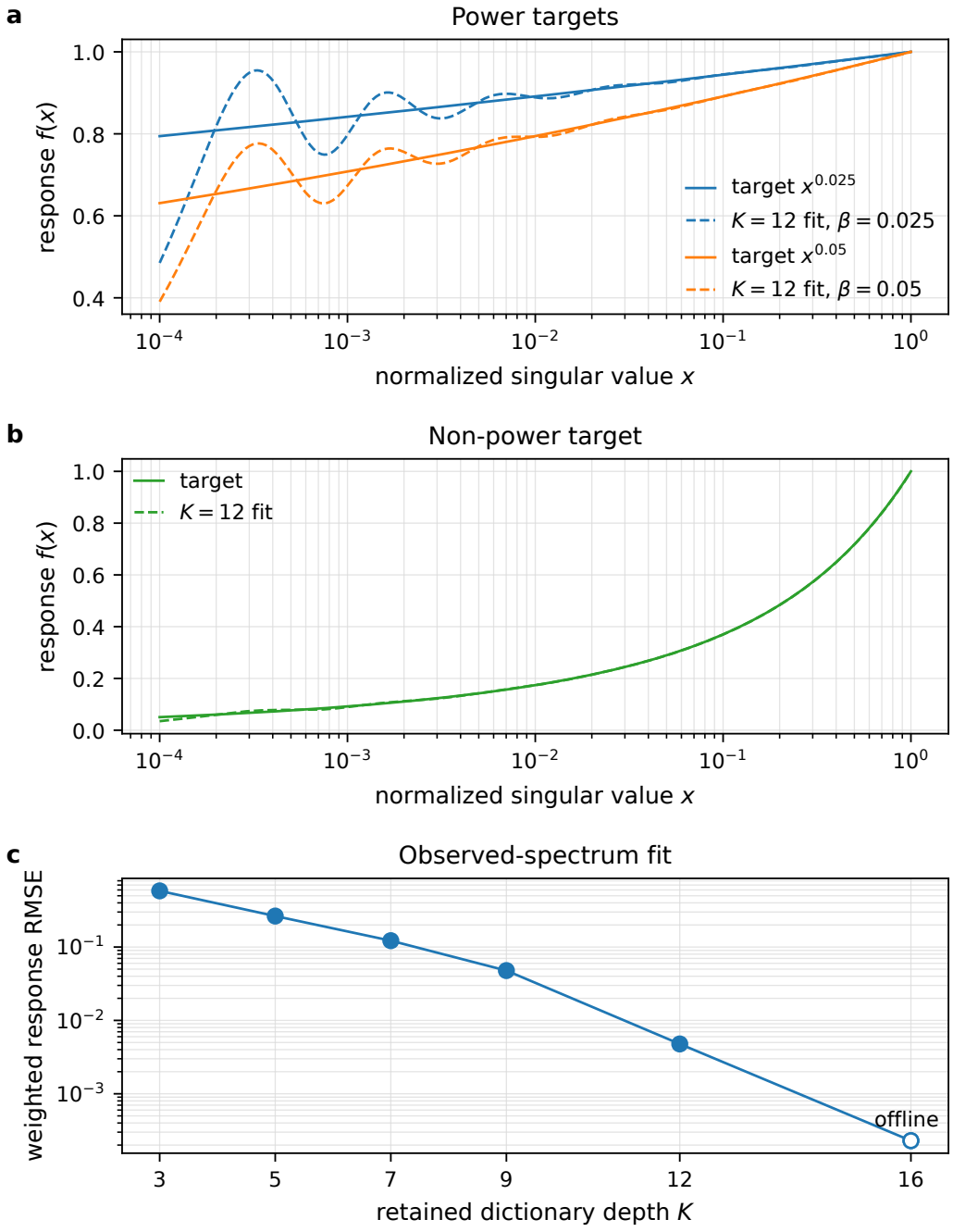

Figure A1: Retained-coordinate approximation and capacity over $x \in [10^{-4}, 1]$. Panel (a) compares the two downstream power targets with their $K = 12$ fits; panel (b) fits the smooth non-power target $0.5x^{0.25} + 0.5x^{0.75}$. Solid lines are analytic targets and dashed lines are retained-coordinate fits. Panel (c) reports observed-spectrum weighted response RMSE across 54 audited spectra; $K = 16$ is offline-only.

B.2.2 Bounded Weak-Tail Targets

The bounded target family $f_{\gamma,\tau}(x) = ((x + \tau)/(1 + \tau))^{\gamma-1}$ tests promotion beyond ideal Muon: for $\gamma < 1$, it has $f(x) \geq 1$ on $[0, 1]$. This moves in the opposite direction from the favorable power

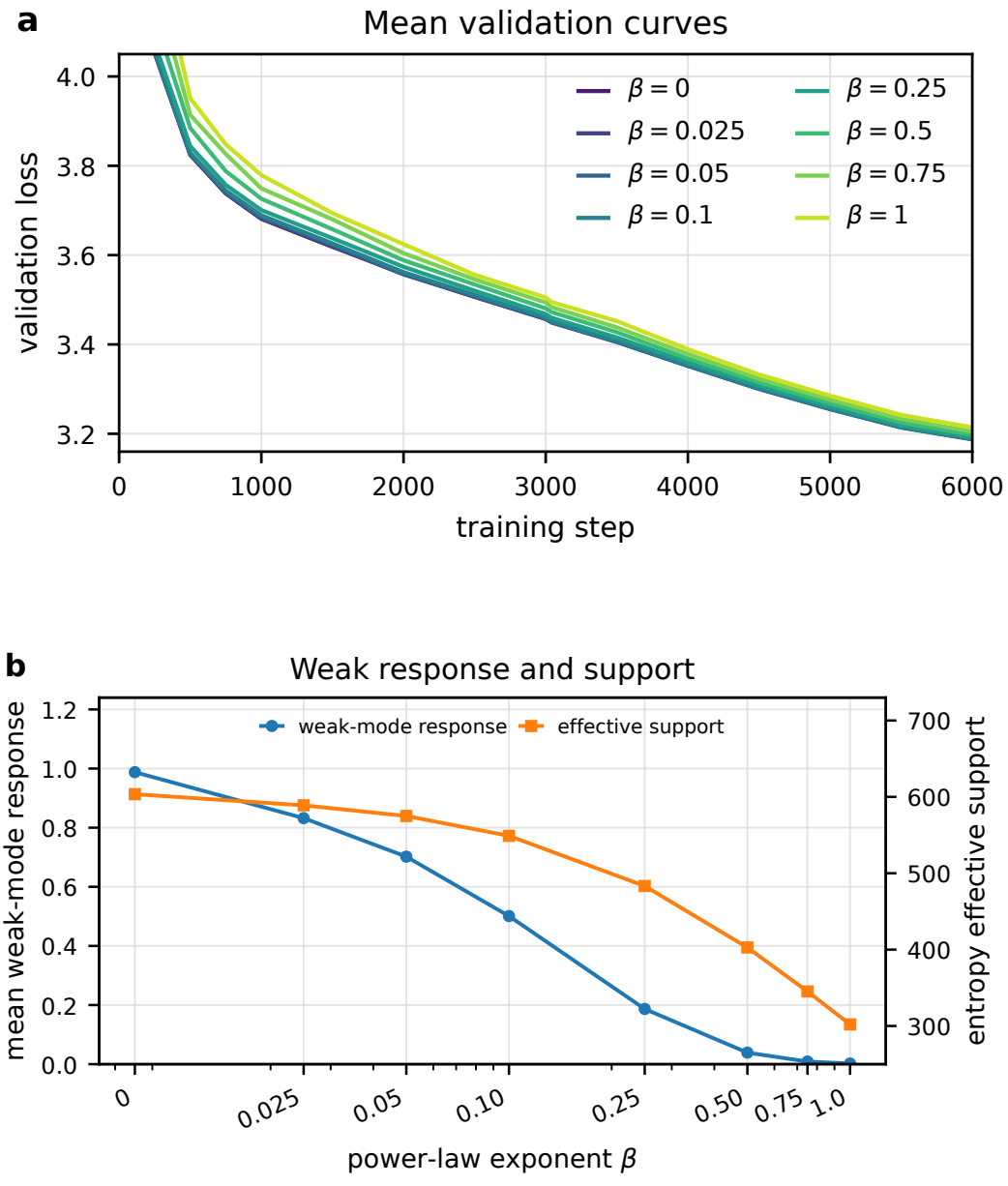

Figure A2: Complete power-base audit under the shared early fading X_0 envelope. Panel (a) shows four-seed mean validation trajectories. Panel (b) shows mean weak-mode response over $x < 0.01$ before norm matching and entropy effective support N_{eff} of normalized contribution mass.

Table A3: Complete 6,000-step power-base board under the shared early fading X_0 envelope. Final loss is mean $\pm$ sample standard deviation over four seeds.

β	Final validation loss	Effective support
0	3.18950 ± 0.00135	603.52
0.025	3.18853 ± 0.00166	588.91
0.05	3.18744 ± 0.00052	574.98
0.10	3.18826 ± 0.00117	548.97
0.25	3.19124 ± 0.00124	483.25
0.50	3.19740 ± 0.00065	402.76
0.75	3.20501 ± 0.00168	345.23
1.00	3.21429 ± 0.00109	302.07

Table A4: Complete board of fits to bounded weak-tail targets under the same shared early fading X_0 envelope. Benefits relative to the $\gamma = 1$ control are milli-loss, with positive meaning lower loss; A_∞ is the peak fitted response defined above.

γ	τ	Benefit	A_∞
0.9	0.003	-8.453	2.083
0.9	0.010	-5.715	1.860
0.9	0.030	-4.683	1.673
0.9	0.100	-4.363	1.494
0.8	0.003	-13.240	3.692
0.8	0.010	-11.638	2.942
0.8	0.030	-9.695	2.380
0.8	0.100	-7.627	1.898
0.7	0.003	-15.013	6.540
0.7	0.010	-13.608	4.653
0.7	0.030	-12.633	3.387
0.7	0.100	-10.088	2.413

target $x^{0.05} < 1$ on $(0, 1)$, so it is not a mediation test of that gain. We fit $f_\alpha(x) = \sum_{k=0}^{12} \alpha_k h_k(x)$ to each target by least squares on 512 logarithmically spaced points over $[10^{-4}, 1]$, then normalize the coefficients so that $f_\alpha(1) = 1$. This unconstrained fit does not automatically preserve the analytic target's bounds; the independent validity audit appears in Table A14. The reported peak fitted response is

$$A_\infty = \max_{x \in \mathcal{G}_{2048}} f_\alpha(x),$$
$$\mathcal{G}_{2048} = \mathrm{logspace}(10^{-4}, 1, 2048).$$

Thus A_∞ is the implemented fitted-map maximum on the stated grid; because $f_\alpha(1) = 1$, it measures peak scalar response relative to the top normalized coordinate before global learning-rate scaling. Table A4 gives all 12 interventions. The contrast is the four-seed mean benefit relative to the $\gamma = 1$ control, whose mean is 3.189565; positive values mean lower loss.

Every tested target has negative benefit while stronger weak-tail targets increase A_∞. Across these designed targets, A_∞ and loss increase $-B$ co-vary with descriptive Pearson $r = 0.873$. Thus the tested fits to bounded weak-tail targets do not give a successful monotone design rule in this pipeline. Top-side edits provide a separate intervention. The independent audit in Table A14 finds positive contribution derivatives on the audited grid but shows that not every fitted response preserves $f(x) \geq 1$, and all twelve exceed the analytic upper bound. The conclusion therefore concerns finite-coordinate fits, without separating target reweighting from approximation effects.

B.2.3 Static Signed Top-Side Edits

Tables A3 and A5 use separate runs with different X_0 controls. The fixed-power sweep retains the shared early fading X_0 envelope, whereas the signed-edit board disables that envelope and applies only the listed constant coefficient c. Thus the $c = 0$ rows in Table A5 are paired controls for the

signed-edit board, not duplicate estimates of Table A3. The mean effect of $c = -0.05$ changes with the base: it worsens loss on fixed $k = 5$ but improves loss on the power base. However, $c = +0.05$ has the lowest tested mean on both bases. These are different observations: the negative edit's mean effect is base-dependent, while four-seed contrasts do not establish a universal preferred sign. We therefore also probe timing.

B.3 Phase-Resolved Diagnostics and Configuration Construction

The checkpoint experiment branches common source states at steps 1,000, 2,000, 2,500, 2,875, and 3,010 into four-seed, 300-step continuations. These branches diagnose conditional local preference. The source configuration is Depth–X_0: $q = 12$ through 2500, linear interpolation to 5 by 3010, and $c(t) = -0.2(1 - t/2000)_+$. Each branch restores the source model, both optimizers (including momentum and variance state), data-loader, random-number, and global-step state. Paired branches share data order and host settings and differ only in the selected spectral response. The three plotted alternatives and plain $k = 5$ all disable the signed correction after branching. The native cooldown clamps every parameter-group rate to zero at step 3,105 and thereafter. Branches still execute all 300 optimizer steps, including the zero-rate portion; their step-averaged loss includes that portion. Cross-checkpoint magnitudes are conditional on this schedule, not equalized by learning-rate exposure. In the clean selected-depth contrast of Table A6, the deep-over-shallow benefit contracts from 17.879 milli-loss at step 1,000 to -0.039 at step 3,010. Under this schedule, the observed advantage is largest early and becomes small near the target. Continuation means use normalized trapezoidal step averages. Temporal-envelope and schedule-ablation windows instead average the evaluated checkpoints equally within each seed, then summarize across seeds; their final columns report endpoints. These metrics differ on nonuniform validation grids.

For the matrix group, the rate at update index t is $\eta_t = \min\{0.0375,\ 3.3169534699576625 \times 10^{-6}(3105 - t)_+^{1.2}\}$. Updates use $t = t_0, \ldots, t_0 + 299$, followed by validation at $t_0 + 300$. For source steps $1000, 2000, 2500, 2875, 3010$, respectively, nonzero-rate update counts are $300, 300, 300, 230, 95$, and $\sum_t \eta_t$ is $8.860786, 3.754621, 1.548608, 0.237789, 0.034223$. The preserved logs confirm zero rates for all four groups and unchanged sampled validation losses beyond the horizon. Optimizer buffers can still advance; zero-rate steps do not update model parameters.

Independent bounded continuation target. The third curve in Figure 3(a) is a logistic-window target, distinct from the (γ, τ) board. With $z = \log_{10} x$ and $\ell(v) = (1 + e^{-v})^{-1}$, define

$$u(x) = 1 + 0.35\,\ell\left(\frac{-2.2 - z}{0.25}\right)\ell\left(\frac{z + 3.5}{0.20}\right), \qquad g(x) = u(x)/u(1).$$

It uses $K = 12,512$ logarithmically spaced points on $[10^{-4}, 1]$, float64 scalar fitting, and uniform fitting weights. The fitted vector minimizes $\sum_j(\sum_k \alpha_k h_k(x_j) - g(x_j))^2 + 10^{-7}\|\alpha\|_2^2$, then is divided by its coefficient sum. The branch uses the full fitted map without a blend ramp or signed edit; matrix recurrence execution remains bfloat16. These settings specify the intervention independently of the bounded-tail sweep.

The broader temporal board fixes the selected-depth schedule and varies only a negative X_0 envelope. Table A7 reports all 12 tested schedules under the positive-benefit convention. Every envelope helps over steps 125–1,000 and reverses over steps 1,000–2,000 and at the final checkpoint. These are training-path effects, including the history of the edit: all envelopes have vanished by step 2,500. They do not isolate the instantaneous effect of introducing an edit at a late checkpoint.

Main-paper Figure 3(b) shows all twelve time-resolved paired means under the positive-benefit convention, while Table A7 preserves every envelope and window summary. They establish that early gains can leave endpoint penalties even after an edit vanishes; they do not select a remedy, optimal cutoff, or preferred late sign. The final -0.2, zero-hold, 2000-step fade is specified separately and tested in Table A8; it is not one of these twelve envelopes. The retained records do not establish a held-out schedule-selection procedure.

Table A8 tests how the selected-depth transition and early signed-X_0 correction combine in the Depth–X_0 configuration. An abrupt $12 \to 5$ switch incurs a 1.025-milli-loss target-window penalty, whereas the smooth schedule differs from the strong fixed-$q = 12$ control by only 0.033 milli-loss in paired target-window mean (paired SD 0.466). The linear transition also spends longer near the deep

Table A5: Complete static signed-X_0 board. Entries are four-seed final validation loss, mean $\pm$ sample standard deviation.

Base map	c	Final validation loss
fixed $k = 5$	0	3.19430 ± 0.00162
fixed $k = 5$	-0.05	3.19550 ± 0.00158
fixed $k = 5$	-0.10	3.19476 ± 0.00120
fixed $k = 5$	-0.20	3.19432 ± 0.00210
fixed $k = 5$	+0.05	3.19379 ± 0.00111
fixed $\beta = 0.05$	0	3.19073 ± 0.00130
fixed $\beta = 0.05$	-0.05	3.18971 ± 0.00139
fixed $\beta = 0.05$	-0.10	3.19010 ± 0.00108
fixed $\beta = 0.05$	+0.05	3.18956 ± 0.00157

Table A6: Paired step-averaged continuation benefit of plain fixed $k = 12$ relative to plain fixed $k = 5$, in milli-loss; positive means lower loss for $k = 12$.

Source step	Mean benefit $\pm$ paired SD
1000	17.879 ± 0.544
2000	6.778 ± 0.047
2500	2.484 ± 0.028
2875	0.184 ± 0.008
3010	-0.039 ± 0.010

response; continuity and depth exposure are not separately identified. This establishes the observed performance of the tested transition, not a general discontinuity penalty. The no-X_0 row further separates early benefit from endpoint gain.

This schedule-ablation board is separate from the four-seed learning-rate and crossing board. Its deep-control endpoint, $3.27890 \pm .00154$, and the crossing board's $3.27915 \pm .00181$ are therefore estimates from different runs, not duplicate summaries of one seed set. Within the schedule-ablation board, seed labels 0–3 pair separate training executions, not branches restored from a shared checkpoint. Early-window means are not identical even for nominally shared prefixes. Paired summaries therefore use seed labels without claiming identical stochastic trajectories; the retained records do not identify the source of that execution variability.

Both complete configurations retain $q(t) = 12$ through step 2,500 and interpolate to $q(t) = 5$ by step 3,010. Depth–X_0 adds an early signed-X_0 correction that vanishes by step 2,000; Depth–Fitted-Power instead ramps to a 25% blend toward the fitted $x^{0.025}$ response, holds that blend through step 1,500, and removes it by step 2,000. They are complementary constructions, and the evidence does not support ranking one above the other.

B.4 Track 3 Target-Crossing Audit

Table A9 reconstructs the stored-record crossings on Modded-NanoGPT Track 3 directly from validation grids and seed endpoints (Jordan et al., 2024a). Depth–Fitted-Power and Depth–X_0 first cross in these stored means at steps 3010 and 3020, while the matched deep control crosses at 3030; the four locally evaluated Muon-family references cross at 3175–3325 (Jordan et al., 2024b; Du & Su, 2026; Li et al., 2025; Dewulf et al., 2026). The shallow control does not cross, with endpoint $3.28529 \pm .00263$ at $\rho = 1$ in Table A13. These results separate two findings: the constructed configurations preserve the deep control's strong target-region behavior and reach the target before all four optimizer references.

First denotes the first evaluated aggregate mean in the stored records strictly below 3.28 on each method's available grid; no interpolation is used. Because the reference rows differ in run count,

Table A7: Complete temporal signed-X_0 board. Benefits relative to the paired no-X_0 control are four-seed means in milli-loss; positive means lower loss. Hold and fade are optimizer steps.

$c(0)$	Hold	Fade	$B_{125:1000}$	$B_{1000:2000}$	B_{3040}
-0.10	0	2000	+0.465	-0.543	-0.825
-0.15	0	2000	+1.618	-1.084	-1.098
-0.25	0	2000	+1.054	-1.430	-1.153
-0.30	0	2000	+0.135	-1.098	-0.615
-0.10	0	1500	+0.110	-0.684	-0.495
-0.15	0	1750	+1.454	-0.539	-0.655
-0.20	0	1500	+2.021	-0.685	-0.778
-0.20	0	1750	+2.874	-0.355	-0.683
-0.20	0	2250	+2.278	-0.900	-0.818
-0.20	0	2500	+2.305	-0.675	-0.495
-0.20	500	2000	+2.269	-0.905	-0.830
-0.25	0	2250	+1.274	-1.030	-1.023

Table A8: Depth-schedule and signed-X_0 ablation (four-seed mean $\pm$ sample standard deviation). Window columns average the evaluated checkpoints equally within each seed, including the dense target-region grid. Fixed controls use $q(t) = 5$ or 12; other rows use the transitions named in the first column. All rows except the last use $c(t) = -0.2(1 - t/2000)_+$; the last isolates the selected-depth schedule.

Schedule	Mean 125–500	Mean 125–1000
fixed $q = 5$	$4.13503 \pm .00453$	$3.93708 \pm .00377$
fixed $q = 12$	$4.07377 \pm .00337$	$3.89417 \pm .00280$
abrupt $12 \to 5$ at 2500	$4.07421 \pm .00430$	$3.89482 \pm .00275$
linear $12 \to 5$, 2000–3010	$4.07322 \pm .00386$	$3.89391 \pm .00271$
Depth–X_0: linear $12 \to 5$, 2500–3010	$4.07308 \pm .00415$	$3.89406 \pm .00275$
same depth schedule, X_0 disabled	$4.07744 \pm .00541$	$3.89621 \pm .00307$
Schedule	**Mean 2500–3040**	**Final**
fixed $q = 5$	$3.30324 \pm .00263$	$3.28536 \pm .00266$
fixed $q = 12$	$3.29631 \pm .00151$	$3.27890 \pm .00154$
abrupt $12 \to 5$ at 2500	$3.29733 \pm .00120$	$3.27938 \pm .00121$
linear $12 \to 5$, 2000–3010	$3.29670 \pm .00137$	$3.27910 \pm .00137$
Depth–X_0: linear $12 \to 5$, 2500–3010	$3.29627 \pm .00136$	$3.27882 \pm .00139$
same depth schedule, X_0 disabled	$3.29606 \pm .00095$	$3.27859 \pm .00099$

horizon, grid, and seed identity, the table reports descriptive optimizer-step checkpoints rather than paired or wall-clock comparisons. The anonymous code/data package contains the validation curves and endpoint summaries used here; its audit reconstructs every crossing. Before five-decimal rounding, the Depth–Fitted-Power and Tuned NS12 endpoint means are 3.27823875 and 3.27824000, respectively. The Fitted-Power mean at step 3010 is 3.27999917, only about 8.3×10^{-7} below target in the stored records; this crossing is sensitive to reporting precision, as detailed below. The preceding evaluated steps are 3000, 3010, 3020, 3150, 3125, 3250, and 3300 for Fitted-Power, Depth–X_0, deep control, Aurora, NorMuon, Newton–Muon, and Tuned NS12, respectively. These sampling intervals are not confidence intervals and do not exclude an unobserved earlier dip.

Source precision and rounding sensitivity. The preserved Fitted-Power curve contains aggregate means to eight decimal places. Its generator averages validation values parsed from text logs; the 24 original logs referenced by the provenance records are unavailable in the current archive. The surviving trainer prints losses to five decimal places, but this does not certify the original export

Table A9: Track 3 target-crossing audit using stored aggregate means. These are original stored-record crossings; Table 1 applies the conservative rounding criterion. The grid column gives the evaluated target-region checkpoint range and stride; NorMuon lists its nonuniform checkpoints explicitly. Final loss is mean ± sample standard deviation at each row's evaluated endpoint. † Unrounded source precision is unverified for the step-3010 crossing; see the sensitivity audit in Appendix B.4.

Method	Runs	Training steps	Final loss
Depth–Fitted-Power	24	3040	$3.27824 \pm .00112$
Depth–X_0	20	3040	$3.27851 \pm .00120$
deep control	4	3040	$3.27915 \pm .00181$
Aurora	12	3200	$3.27830 \pm .00060$
NorMuon	12	3300	$3.27733 \pm .00130$
Newton–Muon	12	3300	$3.27860 \pm .00122$
Tuned NS12	12	3350	$3.27824 \pm .00065$

Method	Checkpoint grid	First mean < 3.28
Depth–Fitted-Power	3000–3040 (10)	$3010^{\dagger}$
Depth–X_0	3000–3040 (10)	3020
deep control	3000–3040 (10)	3030
Aurora	3000–3200 (25)	3175
NorMuon	3000, 3125, 3250, 3300	3250
Newton–Muon	3000–3300 (25)	3275
Tuned NS12	3000–3350 (25)	3325

precision of every run. Consequently, step 3010 identifies the first crossing of the stored mean, not a verified crossing of an unavailable unrounded mean.

If individual losses were rounded to the nearest δ, the mean's worst-case rounding error would be $\delta/2$, without division by seed count. Including the final eight-decimal export adds at most 5×10^{-9}. Under the explicit assumption $\delta = 10^{-5}$, the step-3010 mean has interval $[3.279994165, 3.280004175]$, which straddles 3.28. The next stored mean, 3.27935583 at step 3020, has interval $[3.279350825, 3.279360835]$, entirely below target. Step 3020 is the first evaluated crossing resolvable under this rounding model; it is not an independent estimate of the missing unrounded values. For the main-text comparisons and Table 1, we apply the same conservative rule to every method: the stored mean plus 5.005×10^{-6} must be strictly below 3.28. Only the Fitted-Power crossing changes; both configurations are then reported at step 3020. Table A9 retains the stored-record definition for all methods and marks the sensitive entry. These numerical bounds do not measure uncertainty across training seeds.

B.5 NanoChat Transfer without Retuning

NanoChat revision `92d63d4` (Karpathy, 2025) evaluates D16, D20, D24, and D32 models with 536.9M, 896.5M, 1.384B, and 2.819B parameters, respectively. All runs use 3,000 optimizer steps, sequence length 2,048, and paired seeds 42–45. Table A10 records the distinct batch and training-token allocations. The depth-specific D20/D32 settings use tokenizer artifact B; all other settings use artifact A. Their full SHA-256 identities are listed in the anonymous code/data package README, and every comparison remains within depth and protocol.

All three methods share model, data, evaluation, and auxiliary optimizer settings within each protocol. Their matrix-update pipelines differ. Tuned NS12 applies the fixed 12-step map with direct weight decay; Depth–X_0 and Depth–Fitted-Power retain NanoChat's row equilibration, Muon+ renormalization, NorMuon variance reduction, and cautious weight decay. Thus the comparison includes wrapper differences as well as spectral paths. The ChromaMuon coefficients and schedule breakpoints are fixed in fractional training progress: host step t of a T-step run uses source step

Table A10: Training allocation for the larger-model NanoChat benchmarks. Fixed-budget rows share the listed batch and token count across their three depths.

Protocol	Depth(s)	Batch	Train tokens
Depth-specific	D16	524,288	1.573B
	D20	1,048,576	3.146B
	D24	1,048,576	3.146B
	D32	2,097,152	6.291B
196.6M fixed	D16/20/24	65,536	196.6M
786.4M fixed	D16/20/24	262,144	786.4M

Table A11: Paired final-validation summary for the larger-model NanoChat benchmarks. Tuned NS12 is the absolute four-seed mean $\pm$ marginal sample SD in BPB. Each Δ column reports $10^3(\mathrm{BPB}_{\mathrm{Tuned\ NS12}} - \mathrm{BPB}_{\mathrm{configuration}})$ as paired mean $\pm$ paired sample SD, so positive values favor the named configuration. Each configuration wins all four paired seeds in every row. Protocols and depths are not pooled.

Depth	Tuned NS12 BPB	Δ Depth–X_0	Δ Fitted-Power
Depth-specific batch			
D16	0.807495 ± 0.002982	9.545 ± 2.489	9.606 ± 2.673
D20	0.757374 ± 0.003771	7.794 ± 3.428	7.953 ± 3.623
D24	0.739630 ± 0.002614	5.824 ± 2.465	5.818 ± 2.573
D32	0.691141 ± 0.000433	8.385 ± 0.744	8.310 ± 0.642
196.6M-token fixed budget			
D16	0.934764 ± 0.001254	6.498 ± 0.944	6.106 ± 1.543
D20	0.922076 ± 0.002569	8.318 ± 0.913	6.643 ± 1.780
D24	0.912347 ± 0.003294	8.791 ± 1.724	8.022 ± 2.563
786.4M-token fixed budget			
D16	0.841571 ± 0.002846	4.058 ± 2.519	4.009 ± 2.671
D20	0.827423 ± 0.000375	7.659 ± 0.197	7.902 ± 0.230
D24	0.813645 ± 0.002531	7.061 ± 2.327	7.087 ± 2.611

$t_{\mathrm{src}} = 3040t/T$, with $T = 3000$ here. The comparison tests frozen portability of complete configurations against the transferred Tuned NS12 map, without isolating coefficients or re-testing each phase preference at larger scale.

Every final endpoint uses an 8,388,608-token evaluation at step 3,000. The two fixed-budget protocols additionally use a separate 1M-token routine estimator through step 2,900, so their detached final markers are not connected to the routine curves. Table A11 reports absolute Tuned NS12 BPB and each configuration's paired mean improvement, paired sample SD, and win count. Seed-level endpoints and routine/final validation curves are included in the anonymous code/data package; its audit reconstructs paired differences and win counts.

Across the ten depth–protocol settings, each ChromaMuon configuration finishes below Tuned NS12 on all 40 paired seed–depth endpoints. On the depth-specific-batch protocol, both aggregate means also remain below Tuned NS12 at all 30 post-initial checkpoints for every depth from D16 through D32 (Table A12). Main-paper Figure 4 shows those depth-specific trajectories, while Supplementary Figure A3 gives the complete fixed-budget trajectories. Together, the three protocols support portability in the tested settings without retuning; the close configuration means do not support a ranking between the two configurations.

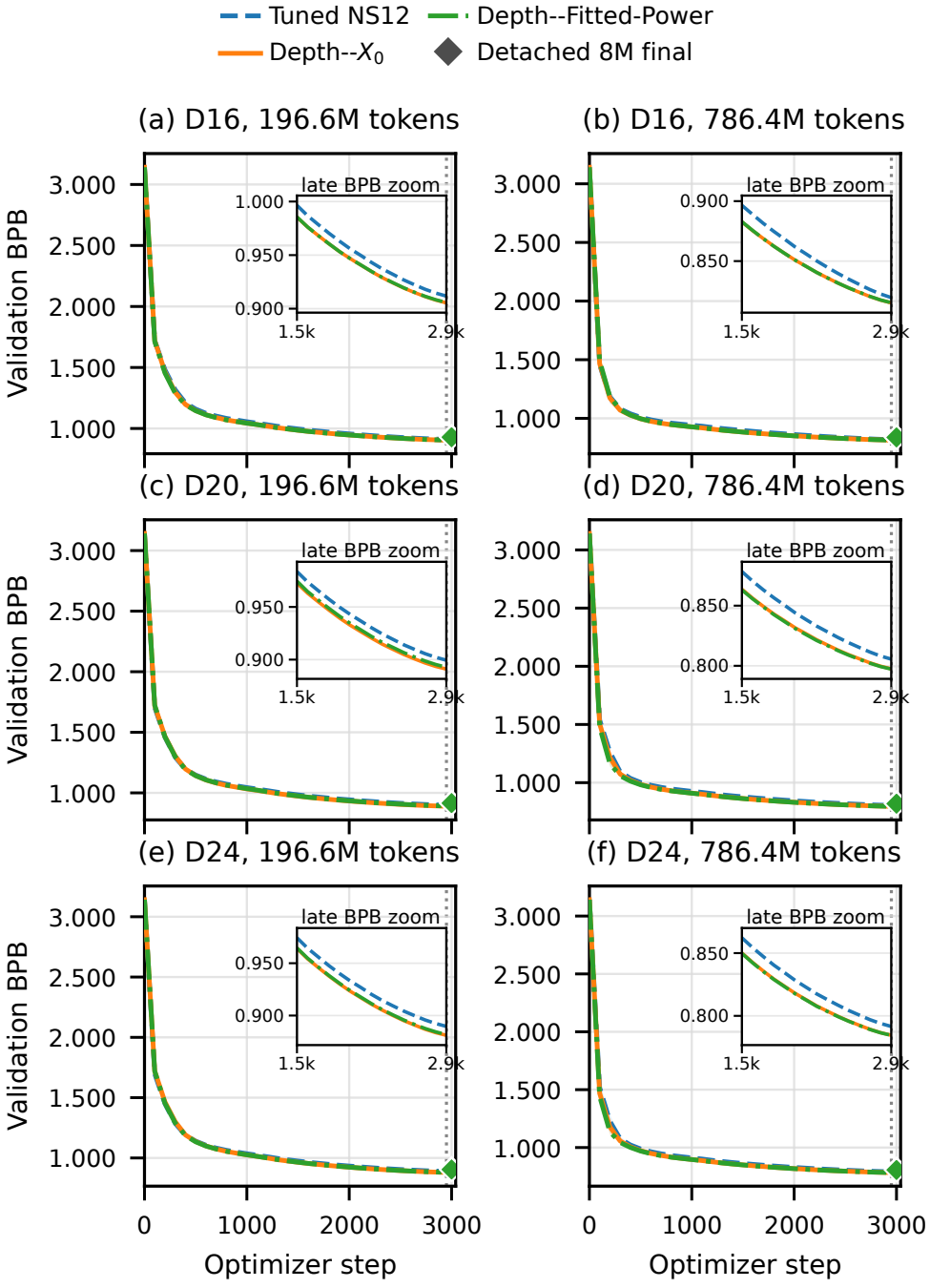

Figure A3: Fixed-budget NanoChat validation. Rows are D16, D20, and D24; columns use 196.6M and 786.4M training tokens. Lines are four-seed routine means through step 2,900; bands show one sample standard deviation across seeds. Diamonds show the detached 8M-token final evaluations at step 3,000; their bars likewise show one sample standard deviation. Insets show aggregate-mean BPB over steps 1,500–2,900 without bands.

Table A12: Depth-specific-batch validation checkpoint dominance. The 30 post-initial checkpoints are steps 100–3,000. Minimum advantage is the smallest aggregate-mean $\mathrm{BPB}_{\text{Tuned NS12}} - \mathrm{BPB}_{\text{config}}$ and its step.

Depth	Configuration	Checkpoints below Tuned NS12	Min. advantage (BPB)	Step
D16	Depth–X_0	30/30	0.006833	100
D16	Depth–Fitted-Power	30/30	0.005795	100
D20	Depth–X_0	30/30	0.007794	3000
D20	Depth–Fitted-Power	30/30	0.007953	3000
D24	Depth–X_0	30/30	0.005823	3000
D24	Depth–Fitted-Power	30/30	0.005818	3000
D32	Depth–X_0	30/30	0.008385	3000
D32	Depth–Fitted-Power	30/30	0.008310	3000

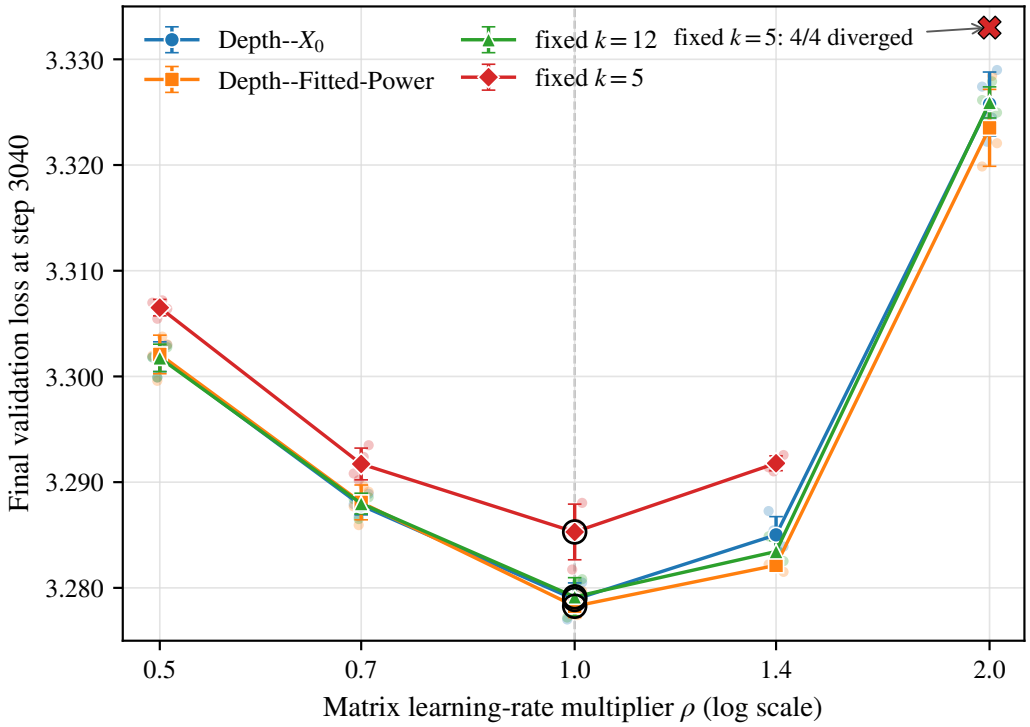

Figure A4: Matrix-rate robustness. Markers and bars are four-seed means and one sample standard deviation; faint points are individual seeds. Depth–X_0 and Depth–Fitted-Power are the two configurations; fixed $k = 12$ and $k = 5$ are the deep and shallow controls under the shared fading envelope. A run is marked diverged when validation loss exceeds 5.0 at or after step 2,500; fixed $k = 5$ at $\rho = 2$ met this criterion in all four seeds.

B.6 Learning-Rate Robustness

The robustness board applies a common multiplier $\rho \in \{0.5, 0.7, 1.0, 1.4, 2.0\}$ to both implementation-level matrix-rate controls; $\rho = 1$ is the reference setting with rate 0.0375. Auxiliary AdamW rates, wrappers, schedule shape, horizon, and seeds remain fixed. Table A13 gives every endpoint, and Figure A4 visualizes the basin and individual seeds.

Table A13: Matrix-rate board: final validation loss at step 3,040 (four paired seeds, mean ± sample standard deviation). Deep and shallow controls fix $q(t) = 12$ and $q(t) = 5$ under the shared fading X_0 envelope.

ρ	Depth–X_0	Depth–Fitted-Power	deep control	shallow control
0.5	$3.30187 \pm .00141$	$3.30208 \pm .00182$	$3.30176 \pm .00130$	$3.30651 \pm .00078$
0.7	$3.28779 \pm .00089$	$3.28808 \pm .00165$	$3.28798 \pm .00098$	$3.29172 \pm .00150$
1.0	$3.27891 \pm .00157$	$3.27826 \pm .00106$	$3.27915 \pm .00181$	$3.28529 \pm .00263$
1.4	$3.28501 \pm .00172$	$3.28213 \pm .00044$	$3.28344 \pm .00102$	$3.29179 \pm .00070$
2.0	$3.32577 \pm .00303$	$3.32352 \pm .00364$	$3.32592 \pm .00147$	diverged (4/4)

Both ChromaMuon configurations have their lowest tested mean at $\rho = 1$ and worsen on either side. Both configurations beat the shallow control at every finite comparison, while remaining close to the strong deep control across the full grid. For Depth–X_0, paired benefits relative to the deep control are -0.11, $+0.18$, $+0.25$, -1.57, and $+0.15$ milli-loss as ρ increases, with descriptive 95% Student-t intervals spanning zero. Thus the reference rate lies in a reproducible local basin across the tested seeds.

C Protocols and Reproducibility

C.1 Small-Model Protocol

The controlled experiments use the Modded-NanoGPT-style FineWeb10B setup (Jordan et al., 2024a): a 12-layer decoder-only Transformer with model dimension 768, vocabulary size 50,304, and sequence length 1,024. The global training batch is 524,288 tokens per optimizer step, and every validation evaluates 10,485,760 tokens. Matrix parameters inside Transformer blocks use the tested Muon or ChromaMuon map. Embeddings, the output projection, and one-dimensional parameters use common AdamW groups (Loshchilov & Hutter, 2019) with learning rates 0.3, $1/320$, and 0.01, respectively, betas $(0.8, 0.95)$, epsilon 10^{-10}, and zero weight decay. Matrix momentum is 0.95 and the reference matrix learning rate is 0.0375.

Static-map experiments use 6,000 optimizer steps and a horizon-adjusted cooldown. They validate every 500 steps and additionally at steps

$$0, 250, 500, 750, 1000, 1500, 2000, 2500,$$
$$3000, 3040, 3500, 4000, 4500, 5000, 5500, 6000.$$

Complete short-horizon runs stop at step 3,040 under a schedule whose rates reach zero at 3,105. These runs validate every 125 steps and additionally at 3,000, 3,010, 3,020, 3,030, and 3,040. Checkpoint probes instead execute 300 continuation steps from each source checkpoint, including any zero-rate steps beyond the schedule horizon (Appendix B.3). Their saved evaluation offsets are 0, 50, 100, 150, 200, 250, and 300. Unless stated otherwise, a setting uses four seeds. Error bars on independent settings are sample standard deviations across seeds; matched contrasts use sample standard deviations of seed-level differences.

For continuation records (t_j, L_j), the reported step-averaged loss is the normalized trapezoidal estimate

$$\overline{L}_{\mathrm{cont}} = \frac{1}{t_J - t_0} \sum_{j=0}^{J-1} (t_{j+1} - t_j) \frac{L_j + L_{j+1}}{2}.$$

Validation loss is the small-model endpoint metric. NanoChat uses validation bits per byte (BPB), defined by its fixed evaluation implementation.

D THEORY AND PROOFS

This section first establishes the finite retained-coordinate representation, coefficient functionals, contribution identity, signed-X_0 non-equivalence, and recurrence stability. It then relates those raw-transform statements to norm geometry and accumulated preconditioning.

D.1 RETAINED-COORDINATE THEORY AND PROOFS

The analysis uses the standard polar-factor and matrix-function framework (Higham, 1986; 2008), specialized to a finite retained trajectory.

Throughout, nonzero $Z \in \mathbb{R}^{m\times n}$ denotes the matrix presented to the recurrence after upstream preprocessing, and $G = \nabla_W F(W)$ is the current gradient. Write

$$Z = U\,\mathrm{diag}(\sigma_i)V^\top, \qquad x_i = \sigma_i/s(Z),$$
$$s(Z) = \left\|ZZ^\top\right\|_F^{1/2} = \|Z\|_{S_4}.$$

Because $s(Z) \geq \|Z\|_2$, every $x_i \in [0,1]$. We work on the active singular subspace and assign zero response to null-space modes. All spectral identities below concern the exact-arithmetic raw direction; finite precision, norm matching, and non-scalar wrappers are addressed separately.

D.1.1 SCALAR BASIS AND CONTRIBUTION IDENTITIES

Theorem A1 (Retained-iterate spectral coordinates)**.** *Let $X_0 = U\,\mathrm{diag}(x_i)V^\top$, and define*

$$X_{k+1} = aX_k + b(X_kX_k^\top)X_k + c(X_kX_k^\top)^2X_k.$$

Then, in exact arithmetic,

$$X_k = U\,\mathrm{diag}(h_k(x_i))V^\top,$$

where $h_0(x) = x$ and

$$h_{k+1}(x) = ah_k(x) + bh_k(x)^3 + ch_k(x)^5.$$

For the reported coefficients $(a,b,c) = (2,-3/2,1/2)$, $\deg h_k = 5^k$. Hence $h_0,\ldots,h_K$ are linearly independent scalar functions and form a basis for their finite span.

Proof. Induct on k. If $X_k = UD_kV^\top$, then

$$X_kX_k^\top = UD_k^2U^\top, \quad (X_kX_k^\top)X_k = UD_k^3V^\top,$$

and $(X_kX_k^\top)^2X_k = UD_k^5V^\top$. Substitution gives $X_{k+1} = U(aD_k + bD_k^3 + cD_k^5)V^\top$, proving the scalar recurrence. For the reported recurrence, the nonzero quintic term makes the leading degree five times the preceding degree. Since $\deg h_0 = 1$, induction gives $\deg h_k = 5^k$; polynomials with distinct degrees are linearly independent. □

Exact functional independence does not imply numerically well-conditioned coordinates on a finite fitting interval, nor does it guarantee that the evaluated matrices $X_k(Z)$ are independent at one finite spectrum. Repeated singular values also make U and V non-unique, but a shared rotation inside a repeated-value subspace leaves $U\,\mathrm{diag}(f(x_i))V^\top$ unchanged.

Proposition A2 (Signal-relative and gradient-aligned contributions)**.** *For a response f, define*

$$D_f^{\mathrm{raw}}(Z) = U\,\mathrm{diag}(f(x_i))V^\top.$$

Then

$$\left\langle G, D_f^{\mathrm{raw}}(Z)\right\rangle_F = \sum_i (u_i^\top Gv_i)f(x_i).$$

If $Z = G$, then

$$\left\langle G, D_f^{\mathrm{raw}}(G)\right\rangle_F = s(G)\sum_i x_i f(x_i).$$

Proof. Expanding the Frobenius inner product gives

$$\operatorname{tr}\big(G^\top U \operatorname{diag}(f(x_i))V^\top\big) = \sum_i (u_i^\top G v_i) f(x_i).$$

When $Z = G$, $u_i^\top G v_i = \sigma_i = s(G)x_i$. □

Thus $C_f(x) = xf(x)$ is the signal-relative contribution profile. It is an exact modal first-order decomposition up to $s(G)$ only when $Z = G$. When Z is preprocessed momentum, the signed modal term is $(u_i^\top G v_i)f(x_i)$, which may be negative. Under a common top-end normalization $f(1) = 1$, greater contribution at a fixed weak mode relative to the top endpoint requires greater weak-mode response.

Proposition A3 (Power response plus a retained X_0 edit)**.** *Let $I \subset (0, \infty)$ be a nontrivial interval and $a \neq 0$. If*

$$x^\beta + cx = ax^{\beta'} \qquad \textit{for every } x \in I,$$

then $c = 0$ or $\beta = 1$. For $\beta < 1$, a nonzero edit cannot be represented exactly by another scaled power on I; it preserves the leading weak-tail exponent while changing the raw endpoint at $x = 1$ from 1 to $1 + c$ and the endpoint slope from β to $\beta + c$.

Proof. Distinct real monomials are linearly independent on a positive interval. If β, 1, and β' were all distinct, the displayed identity would contradict that independence. If $\beta' = \beta$, then $cx = (a - 1)x^\beta$, which requires $c = 0$ or $\beta = 1$. If $\beta' = 1$, then $x^\beta = (a - c)x$, which requires $\beta = 1$. The endpoint and slope statements follow by direct evaluation of $x^\beta + cx$. □

D.1.2 Local Recurrence Structure and Stability

For the reported recurrence, let

$$p(x) = 2x - \frac{3}{2}x^3 + \frac{1}{2}x^5,$$
$$h_0(x) = x, \qquad h_{k+1} = p \circ h_k.$$

Lemma A4 (Local structure)**.** *For all $k \geq 0$,*

$$h_k(0) = 0, \qquad h_k(1) = 1, \qquad h_k'(0) = 2^k.$$

Moreover, $h_0'(1) = 1$, $h_k'(1) = 0$ for $k \geq 1$, and

$$h_k(x) = 2^k x + O(x^3), \qquad x h_k(x) = 2^k x^2 + O(x^4)$$

as $x \downarrow 0$.

Proof. The claims follow from

$$p(0) = 0, \quad p(1) = 1, \quad p'(0) = 2, \quad p'(1) = 0$$

and the chain rule. The expansion follows inductively from $p(y) = 2y + O(y^3)$ near zero. □

Consequently, every retained response satisfies

$$f_\alpha(1) = \sum_{k=0}^{K} \alpha_k, \qquad f_\alpha'(0) = \sum_{k=0}^{K} 2^k \alpha_k, \qquad f_\alpha'(1) = \alpha_0.$$

These are distinct linear functionals: a signed X_0 edit changes both top value and top slope. The derivative at zero is an exact local statement; how far equalization extends over an observed spectrum is empirical.

Let $x_\star = 2/\sqrt{5}$ and

$$B = p(x_\star) = \frac{56}{25\sqrt{5}} \approx 1.00175845.$$

Proposition A5 (Stability of the reported scalar recurrence)**.** *The interval* $[0, B]$ *is invariant under* p*. For every* $x \in (0, 1]$*,* $h_k(x) \to 1$*, and*

$$0 \le h_k(x) \le B \qquad (x \in [0, 1],\ k \ge 0).$$

Proof. The factorizations

$$p'(x) = \frac{1}{2}(5x^2 - 4)(x^2 - 1),$$
$$p(x) - x = \frac{1}{2}x(x^2 - 1)(x^2 - 2)$$

give the required interval behavior. On $[0, x_\star]$, p is increasing; on $[x_\star, 1]$, it is decreasing. Hence $p([0, 1]) \subseteq [0, B]$. Since $1 < B < \sqrt{2}$, for $x \in (1, B]$ one has $1 < p(x) < x \le B$, so $p([1, B]) \subseteq [1, B]$. This proves invariance.

For $x \in (0, 1)$, $p(x) > x$. If all iterates remain below one, they form an increasing bounded sequence whose limit is a fixed point; the only positive fixed point in $[0, 1]$ is one. If an iterate enters $(1, B]$, all later iterates remain in that interval and decrease strictly to a fixed point. The only fixed point in $[1, B]$ is again one. The case $x = 1$ is immediate. □

The exact statement applies because the reported normalization gives $x_i \in [0, 1]$. Bfloat16 rounding and epsilon clamping are implementation perturbations outside this exact-arithmetic proposition.

D.1.3 Fitted-Map Interval and Validity Diagnostics

Every finite retained blend satisfies $f_\alpha(0) = 0$. It therefore cannot uniformly approximate on an interval reaching zero a continuous target with a nonzero right limit there. Reported power and floored-target fits are defined only on $[10^{-4}, 1]$, not at $x = 0$.

Proposition A6 (Bounds for the floored contribution target)**.** *For*

$$f_{\gamma,\tau}(x) = \left(\frac{x + \tau}{1 + \tau}\right)^{\gamma - 1}, \quad 0 < \gamma \le 1, \quad \tau > 0,$$

and $x \in [0, 1]$*,*

$$1 \le f_{\gamma,\tau}(x) \le A_{\gamma,\tau} := \left(\frac{1 + \tau}{\tau}\right)^{1 - \gamma}.$$

Its contribution profile $C_{\gamma,\tau}(x) = x f_{\gamma,\tau}(x)$ *satisfies*

$$C'_{\gamma,\tau}(x) = f_{\gamma,\tau}(x)\frac{\tau + \gamma x}{\tau + x} > 0.$$

Proof. The value bound follows from $(x+\tau)/(1+\tau) \in [\tau/(1+\tau), 1]$ and $\gamma - 1 \le 0$. Differentiating $x f_{\gamma,\tau}(x)$ gives the displayed positive derivative. □

The analytic target is monotone, but its maximum response grows as $\tau^{-(1-\gamma)}$. The fitted map is an unconstrained least-squares approximation and does not inherit the target's value or derivative bounds. Table A14 audits the actual fitted maps. The fit RMSE uses the 512 fitting points. The audit RMSE, response extrema, and minimum $C'_\alpha(x) = f_\alpha(x) + x f'_\alpha(x)$ use 16,384 geometric midpoints over $[10^{-4}, 1]$, disjoint from the fitting grid; derivatives are evaluated analytically through the polynomial recurrence.

The independent-grid RMSE closely tracks the fitting-grid RMSE, while aggressive targets require larger coefficient norms and exhibit larger approximation error. The contribution derivative is positive at every audit point, but this dense-grid check is not an interval-wide analytic certificate and monotonicity is not an inherited property of unconstrained fitting.

Table A14: Numerical validity audit for the 12 fits to bounded weak-tail targets. All quantities use float64 scalar recurrence evaluation. The positive audited minimum of C'_α shows increasing fitted contribution at every audit point, but several fits do not preserve $f \geq 1$. All twelve exceed the analytic upper bound $A_{\gamma,\tau} = ((1+\tau)/\tau)^{1-\gamma}$. No claim is made below 10^{-4} or for exact BF16 matrix realization.

Metric	$\tau = .100$	$\tau = .030$	$\tau = .010$	$\tau = .003$
$\gamma = 0.9$				
Fit RMSE	.1024	.1147	.1278	.1440
Audit RMSE	.1012	.1134	.1264	.1424
$\min f_\alpha$	.7670	.8591	.9559	1.0000
$\max f_\alpha$	1.4936	1.6725	1.8596	2.0833
$A_{\gamma,\tau}$	1.2710	1.4242	1.5865	1.7882
$\max f_\alpha / A_{\gamma,\tau}$	1.1752	1.1743	1.1721	1.1650
$\min C'_\alpha$	.6079	.6776	.7439	.8007
$\|\alpha\|_2$	5.634	6.314	7.033	7.926
$\gamma = 0.8$				
Fit RMSE	.1301	.1634	.2027	.2575
Audit RMSE	.1287	.1616	.2005	.2546
$\min f_\alpha$	.9749	1.0001	1.0001	1.0001
$\max f_\alpha$	1.8983	2.3801	2.9421	3.6921
$A_{\gamma,\tau}$	1.6154	2.0284	2.5169	3.1977
$\max f_\alpha / A_{\gamma,\tau}$	1.1751	1.1734	1.1689	1.1546
$\min C'_\alpha$	.7710	.7876	.7850	.7864
$\|\alpha\|_2$	7.167	8.999	11.165	14.178
$\gamma = 0.7$				
Fit RMSE	.1653	.2326	.3216	.4604
Audit RMSE	.1635	.2301	.3180	.4553
$\min f_\alpha$	1.0001	1.0001	1.0001	1.0001
$\max f_\alpha$	2.4126	3.3867	4.6534	6.5397
$A_{\gamma,\tau}$	2.0531	2.8888	3.9930	5.7181
$\max f_\alpha / A_{\gamma,\tau}$	1.1751	1.1724	1.1654	1.1437
$\min C'_\alpha$	.6972	.6776	.6754	.6809
$\|\alpha\|_2$	9.116	12.824	17.717	25.341

D.2 Current-Spectrum Filtering versus Accumulated Preconditioning

On the active singular subspace,

$$\begin{aligned}\operatorname{msgn}(Z) &= UV^\top = Z(Z^\top Z)^{-1/2},\\ D_f^{\mathrm{raw}}(Z) &= U\operatorname{diag}(f(x_i))V^\top = Zq_{f,s}(Z^\top Z),\\ q_{f,s}(\lambda) &= \frac{f(\sqrt{\lambda}/s(Z))}{\sqrt{\lambda}}, \qquad \lambda > 0.\end{aligned}$$

This identity characterizes the retained transform as a filter of its current input that preserves that input's singular subspaces. K-FAC, Shampoo, and SOAP use accumulated information to rotate or rescale directions (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2024). The local host uses SOAP preprocessing on selected matrix groups before the retained transform; the identity applies to the resulting Z, not necessarily to raw momentum.

D.3 Norm-Geometry Bridge and Scope

This background connects the paper's response maps to familiar matrix steepest-descent geometries. For a matrix gradient $G = U\operatorname{diag}(\sigma_i)V^\top$, minimizing the local linear model under a unit Schatten-r budget, $1 < r < \infty$, gives, up to scale,

$$\Delta_r^\star \propto -U\operatorname{diag}\left(\sigma_i^{1/(r-1)}\right)V^\top.$$

The $r = 2$ endpoint is matrix SGD, while $r = \infty$ is the polar factor $-UV^\top$ associated with ideal Muon (Jordan et al., 2024b; Pethick et al., 2025; Massena et al., 2026). This explains why SGD-like $f(x) = x$ and Muon-like $f(x) = 1$ are useful response endpoints. The contribution profile $xf(x)$ answers a different question: how a response allocates the current signal across singular modes.

ChromaMuon's retained-coordinate map is not, by itself, a new global norm geometry. Its scalar interpretation applies to the raw transform of the matrix signal presented to the recurrence. Upstream momentum or preprocessing can make that signal differ from the current gradient; positive norm matching preserves relative spectral shape but changes overall scale; and subsequent non-scalar wrappers can alter singular directions. These boundaries separate the response diagnostic from claims about the complete applied optimizer.

E Implementation Details

E.1 Optimizer Pipeline and Definitions

The paper experiments use the fixed stable quintic $(a, b, c) = (2, -1.5, 0.5)$ within Muon's Newton–Schulz construction (Jordan et al., 2024b). The retained-coordinate construction itself can be applied to another finite polynomial recurrence; the reported ChromaMuon coefficients and matched retained-depth controls use this one. Algorithm 1 constructs the retained dictionary, and Algorithm 2 records the complete application order used by the reported configurations.

We write this map as $\textsc{NSCoordinates}(H, K, a, b, c, \epsilon)$. Algorithm 2 invokes it with $K = 12$ and the reported recurrence coefficients.

Here $\mathcal{W}$ may also depend on the current gradient and persistent host-optimizer state; these arguments are suppressed to keep the transform ordering readable.

E.1.1 Raw, Norm-Matched, and Fully Applied Updates

The exact spectral identities concern

$$D_\alpha^{\mathrm{raw}}(Z) = \sum_k \alpha_k X_k(Z).$$

The implementation matches the direction to its base before the signed edit, $D_{\mathrm{base}} = (1 - w)D_{\mathrm{depth}} + wD_{\mathrm{fit}}$:

$$\widehat{D}_\alpha(Z) = \frac{\max\{\|D_{\mathrm{base}}\|_{S_4}, \epsilon\}}{\max\{\|D_\alpha^{\mathrm{raw}}\|_{S_4}, \epsilon\}} D_\alpha^{\mathrm{raw}}(Z).$$

Algorithm 1 Retained Newton–Schulz coordinate construction

Require: Host-preprocessed matrix H, depth K, recurrence (a, b, c), clamp ϵ
1: $Z \leftarrow H^\top$ if H is taller than wide, else $Z \leftarrow H$; record orientation $\mathcal{O}$
2: $X_0 \leftarrow Z / \max\{\|Z\|_{S_4}, \epsilon\}$
3: **for** $k = 0, \ldots, K-1$ **do**
4: $\quad S_k \leftarrow X_k X_k^\top$
5: $\quad X_{k+1} \leftarrow aX_k + bS_k X_k + cS_k^2 X_k$
6: **end for**
7: **return** $X_0, \ldots, X_K$ and $\mathcal{O}$

Algorithm 2 Complete ChromaMuon optimizer application order

Require: Matrix blocks $\mathcal{M}$, rate η, momentum μ, host preprocessing $\mathcal{P}$, host wrappers $\mathcal{W}$, reported schedule functions $q(t), w(t), c(t)$, $\alpha^{(0.025)}$, and clamp ϵ
1: Initialize host momentum state $B_W \leftarrow 0$ for $W \in \mathcal{M}$
2: **for** step $t = 0, \ldots, T-1$ **do**
3: $\quad$ Compute stochastic gradients G_W
4: $\quad$ **for** $W \in \mathcal{M}$ **do**
5: $\quad\quad B_W, M_W \leftarrow \textsc{HostMomentum}(B_W, G_W; \mu)$
6: $\quad\quad H_W \leftarrow \mathcal{P}(M_W)$
7: $\quad\quad \mathcal{C}_W \leftarrow \textsc{NSCoordinates}(H_W, 12, 2, -1.5, 0.5, \epsilon)$
8: $\quad\quad$ Unpack $\mathcal{C}_W$ as $X_0, \ldots, X_{12}, \mathcal{O}$
9: $\quad\quad D_{\text{depth}} \leftarrow \sum_{k=0}^{12} d_k(q(t)) X_k$
10: $\quad\quad D_{\text{fit}} \leftarrow \sum_{k=0}^{12} \alpha_k^{(0.025)} X_k$
11: $\quad\quad D_{\text{base}} \leftarrow (1 - w(t)) D_{\text{depth}} + w(t) D_{\text{fit}}$
12: $\quad\quad D_{\text{pre}} \leftarrow D_{\text{base}} + c(t) X_0$
13: $\quad\quad \widehat{D} \leftarrow \dfrac{\max\{\|D_{\text{base}}\|_{S_4}, \epsilon\}}{\max\{\|D_{\text{pre}}\|_{S_4}, \epsilon\}} D_{\text{pre}}$
14: $\quad\quad D_W^{\text{applied}} \leftarrow \mathcal{W}(\mathcal{O}^{-1}(\widehat{D}), W, t)$
15: $\quad\quad W \leftarrow W - \eta D_W^{\text{applied}}$
16: $\quad$ **end for**
17: $\quad$ Update non-matrix parameters with the specified auxiliary optimizer
18: **end for**

For Depth–X_0, this is the depth-only base. For Depth–Fitted-Power, the base already contains the fitted pulse and $c = 0$, so this ratio is one; the pulse is not separately matched to the depth-only norm. Positive scalar matching preserves relative spectral shape but can make dependence on edited coefficients nonlinear. Matching one raw Schatten-4 norm does not hold every update norm or the fully applied step size fixed. We reserve D^{applied} for the direction after the configured host wrappers. Non-scalar row/column wrappers may change singular vectors, so the final parameter update must not be identified with the raw map.

E.1.2 Reported Matrix Pipeline

Algorithm 1 is host-independent: H denotes matrix momentum after host-side preprocessing, and $Z = \mathcal{O}(H)$ is its wide-oriented recurrence input. Both integrations construct lookahead momentum. The local host uses coefficient 0.95; the pinned NanoChat trainer ramps it from 0.85 to 0.97 over 400 steps, holds it through step 1,050, then decreases it linearly toward 0.90 at step 3,000, identically for all three methods. The local host then applies SOAP preprocessing to both MLP matrices and the attention-value projection; other matrix groups use the unpreconditioned lookahead signal. These host settings are shared within the controlled comparisons. The local shell casts H to bfloat16; NanoChat instead applies MuonEq row equilibration before the recurrence. In both cases, $X_0 =$

$Z/s(Z)$, where

$$\begin{aligned} s(Z) &= \left\| ZZ^\top \right\|_F^{1/2} \\ &= \left(\sum_i \sigma_i(Z)^4 \right)^{1/4} \\ &= \|Z\|_{S_4} . \end{aligned}$$

Thus active normalized singular values lie in $[0, 1]$ in exact arithmetic. The denominator is epsilon-clamped, and $\mathcal{O}^{-1}$ restores parameter orientation before host wrappers. Bfloat16 rounding and clamping are implementation perturbations outside the exact scalar identities.

After blending, the local implementation matches to the base before the signed edit as defined above. Its shared Track 3 shell then applies rectangular scaling, row/column variance normalization, and an update-to-weight floor. In NanoChat, the two ChromaMuon configurations retain MuonEq row equilibration before the spectral transform, followed by Muon+ norm rescaling, NorMuon variance reduction, and cautious weight decay. The Tuned NS12 comparator instead feeds lookahead momentum directly to the terminal NS12 map and applies ordinary weight decay; it omits those four operations. All methods share the model, data, auxiliary optimizer, and outer matrix-rate scaling, but their complete matrix updates differ. Consequently, the transfer comparison does not isolate the spectral path from the wrappers. Exact singular-vector statements apply to the raw transform, before non-scalar wrappers.

E.2 Coefficient Arrays and Schedules

The power-response and bounded-tail training maps use ordinary least squares in $h_0, \ldots, h_{12}$ on 512 logarithmically spaced points in $[10^{-4}, 1]$, followed by coefficient-sum normalization. The independent logistic-window continuation probe instead uses the ridge objective specified in Appendix B.3. Fits may contain negative coefficients and are not clipped. Both procedures enforce $f_\alpha(1) = 1$ but impose no target value or derivative bounds; Table A14 audits the bounded-tail fits. Figure A1's separate approximation illustration uses the stated 900-point grid.

For the Depth–Fitted-Power configuration's $x^{0.025}$ target,

$$\begin{aligned} \alpha^{(0.025)} = (&0.031533, 0.009853, 0.032520, -0.011684, \\ &0.065050, -0.072113, 0.171908, -0.265051, \\ &0.513819, -0.863558, 1.493178, -2.143654, \\ &2.038200). \end{aligned}$$

The compact code/data supplement includes fitting utilities and the complete configuration coefficients. It also includes the saved Figure 3(a) continuation curves, a reconstructed full-precision vector for the independent bounded target, and a standard-library audit of the paired continuation means. Additional records provide seed-level learning-rate and basis-depth endpoints, schedule-ablation validations, and the stored per-spectrum fit errors. It does not archive every secondary fitted array or original checkpoint.

E.2.1 Depth–X_0 Configuration

The schedules below use source-step units. In Track 3, t is the optimizer step; NanoChat evaluates these functions at $3040t/T$, preserving their fractional-progress locations in a T-step run. The selected-depth schedule is

$$q(t) = \begin{cases} 12, & t \le 2500, \\ 12 + \dfrac{t - 2500}{3010 - 2500}(5 - 12), & 2500 < t < 3010, \\ 5, & t \ge 3010. \end{cases}$$

For $j = \lfloor q \rfloor$, the depth coefficients are $d_j(q) = j + 1 - q$, $d_{j+1}(q) = q - j$, and zero otherwise; at integer q, $d_q(q) = 1$. The early signed correction is

$$c(t) = \begin{cases} -0.2(1 - t/2000), & 0 \le t < 2000, \\ 0, & t \ge 2000. \end{cases}$$

For Depth–X_0, $w(t) = 0$. Before norm matching,

$$D_t^{\mathrm{pre}} = D_{\mathrm{depth}}^{\mathrm{raw}}(Z_t; q(t)) + c(t)X_0(Z_t).$$

The norm-matched direction is $r(Z_t, t)D_t^{\mathrm{pre}}$, where the positive scalar matches the uncorrected depth-update norm. The common non-scalar wrappers then produce D_t^{applied}.

E.2.2 Depth–Fitted-Power Configuration

Depth–Fitted-Power uses the same $q(t)$, disables the signed X_0 correction, and blends the all-matrix depth direction toward $\alpha^{(0.025)}$ with

$$w(t) = \begin{cases} 0, & t < 500, \\ 0.25(t - 500)/500, & 500 \le t < 1000, \\ 0.25, & 1000 \le t \le 1500, \\ 0.25(2000 - t)/500, & 1500 < t < 2000, \\ 0, & t \ge 2000. \end{cases}$$

Thus $D_t^{\mathrm{pre}} = (1 - w(t))D_{\mathrm{depth}}^{\mathrm{raw}} + w(t)D_{\mathrm{fit}}^{\mathrm{raw}}$. This is also the norm-matching reference, so its scalar matching factor is one before the host wrappers.

The implementation evaluates the complete dictionary through $K = 12$ throughout training, including after selected depth reaches $q(t) = 5$ in either configuration.

E.3 Configuration Manifest and Provenance Coverage

Table A15 maps the reported configurations to the anonymous package. The manifest distinguishes inspectable implementations from historical results whose full launch records are unavailable. The package's `data/track3/configuration_manifest.json` records source hashes, settings, seed coverage, and unrecovered fields. Its command generator defaults to four seeds; this default does not reproduce the larger 20- and 24-run aggregates in Table 1.

Table A15: Configuration mapping. Paths are relative to the anonymous package. The coefficient schedules are specified in Appendix E.2. Historical external Track 3 settings cannot be inferred from the current packaged implementation.

Reported row	Implementation and comparison boundary
Depth–X_0	`src/chromamuon.py`, method `depth_x0`: scheduled depth, fading signed edit, no fitted pulse.
Depth–Fitted-Power	Same core, `depth_fitted_power`: scheduled depth and fitted pulse, no signed edit.
Internal deep control	Same core, `fixed_k12_shared_x0`: $q = 12$, sharing Depth–X_0's signed fade and local host shell.
Tuned NS12 (NanoChat)	`adapters/nanochat_92d63d4_three_method.patch`: terminal 12-step map with direct updates and ordinary decay; omits the ChromaMuon wrappers.
Tuned NS12 (Track 3)	Stored curves and 12-run endpoint summaries are preserved; exact historical recurrence, preprocessing, wrappers, rates, and tuning record are not recovered.

Shared retained core and local host. The packaged paths and internal controls use twelve iterations with fixed coefficients $(2, -3/2, 1/2)$, wide orientation (transpose if rows exceed columns), and Schatten-4 input normalization clamped at 10^{-7}. Matrix arithmetic is bfloat16 with float32 norm evaluation. Signed edits match the pre-edit base's Schatten-4 norm with clamp 10^{-10}. The local adapter, `adapters/track3_chromamuon_train.py`, uses lookahead momentum 0.95. SOAP preprocessing has second-moment factor 0.90, basis-update frequency 10, MLP blend 1, value-projection blend 0.95, and a value denominator floor ratio of 0.55. After the spectral transform, the host applies rectangular scaling, axis-wise variance normalization with factor 0.95 and

norm restoration, and a Frobenius update-to-weight ratio floor of 0.3825. The packaged matrix step has no separate decay term; an unused configured decay value must not be interpreted as an applied update.

The local matrix rate is $\min(0.0375, 3.3169534699576625\times10^{-6}(3105-t)_+^{1.2})$. Auxiliary AdamW uses initial rates $(0.3, 1/320, 0.01)$ for embeddings, output projection, and remaining parameters, with betas $(0.8, 0.95)$, epsilon 10^{-10}, and zero decay. Each rate follows the same capped power cooldown; its coefficient is listed in the machine-readable manifest. The complete runs use 3,040 steps, sequence length 1,024, global batch 524,288 tokens, and 10,485,760 validation tokens. Evaluations occur every 125 steps and at 3,000, 3,010, 3,020, 3,030, and 3,040.

Selection and historical coverage. The Depth–X_0 aggregate uses seeds 0–19, choosing the lexicographically first complete log per seed. The Fitted-Power aggregate uses the 24 completed runs for the stated pulse: seeds 4–7, 12–23, and 80–87. Internal controlled comparisons use seeds 0–3. These retained selections do not establish an exhaustive tuning budget or held-out schedule selection. The external Aurora, NorMuon, Newton–Muon, and Tuned NS12 rows have twelve runs each, with their final steps and evaluation grids in Appendix B.4. Their complete launch configurations and tuning histories are unavailable; their names and saved curves do not establish a common host pipeline. Consequently, the external-reference gaps do not isolate coefficient-path effects. The internal deep control provides the shared-shell comparison.

NanoChat mapping. The supplied patch targets NanoChat revision `92d63d4e8bb4df75c3b71618f31ddde2378b2bcd`. Its ChromaMuon and Tuned NS12 branches share the fixed recurrence and normalization described above, but use the distinct wrapper pipelines in Appendix E.1. Their outer training settings are paired within protocol; the command generator and Appendix B.5 specify batches, validation, and tokenizer provenance. This inspectable NanoChat implementation does not certify the unrecovered historical Track 3 Tuned NS12 settings.