THE MECHANISM OF TRAINING COLLAPSE IN ULTRA-DEEP TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultra-deep transformers can *collapse* in training: after a period of successful loss descent a run jumps, often within ten optimizer steps, onto a persistent degenerate plateau, a near-zero predictor for a flow-matching DiT and a near-unigram predictor for a language model. We study this collapse on diffusion transformers of 256–1024 residual sublayers and language models of 256–2048, pushed past their stability boundary by the learning rate, with kernel probes of the residual-stream write matrices, loss profiles along committed AdamW updates, and interventions. Our hypothesis is a condition on one optimizer step, written in output space: the step is harmful when the tangent kernel's gain along the loss gradient, times the learning rate, carries the output past the radius within which the loss still decreases. The tolerable learning rate falls with depth, as a power law over the depths tested; the gain along the gradient rises before any loss symptom and has fallen from its peak on every terminal plateau; and neither the amplified direction nor continued growth of the kernel is required for the instability. GROW, a retraction of each write matrix onto a fixed-Gram set after every AdamW step, is an intervention on this geometry, not a tuned optimizer: keeping the Post-LN forward pass and the zero-initialized write matrices, at four depths from 256 to 2048 residual sublayers it trains through a constant dose that collapses most or all AdamW seeds at the first three, and raising the scale of its fixed factor brings its sample quality to within two FID, inside the run-to-run spread, of a healthy low-rate AdamW reference at every depth.

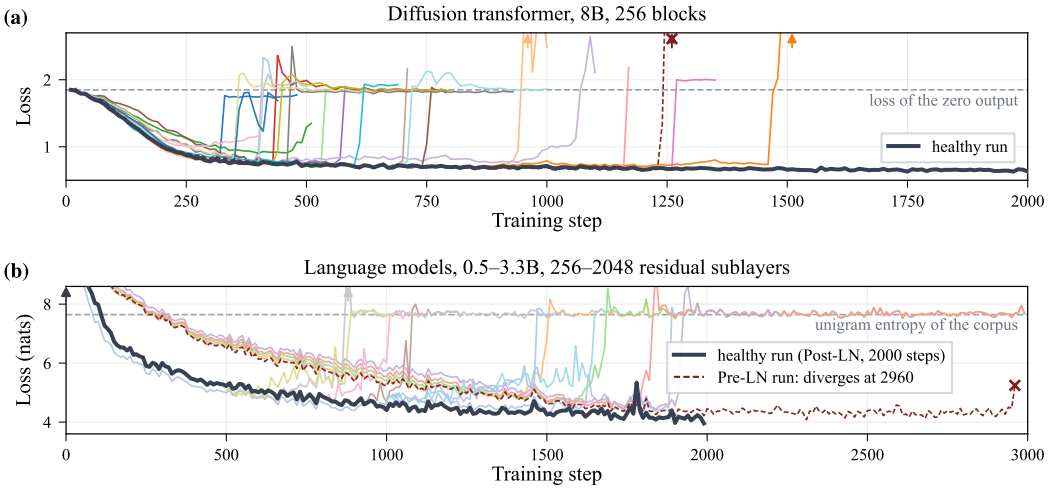


Figure 1: Training failure in ultra-deep transformers, one color per run. (a) An 8B diffusion transformer, 23 runs (the baseline recipe and attempted fixes): collapsed runs land on the loss of the zero output (dashed line) or diverge (arrows, cross). (b) Language models with 256–2048 residual sublayers: collapsed runs land on the unigram entropy (dashed line); one Pre-LN run diverges instead (cross). Grey: healthy runs.

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.

1 INTRODUCTION

Ultra-deep transformers can collapse in training (Figure 1). In the diffusion transformers and language models we train to failure here, a run that has been descending for hundreds of steps enters a persistent degenerate plateau, often within ten optimizer steps, and no later change of the learning rate that we tried — including a further tenfold increase — reverses it. The run stays finite throughout: the *collapse* is a change in the function computed, not an overflow. At the same time, large directional gain is a normal feature of healthy networks: in every released model we probed, a change of the parameters that write into the residual stream moves the output along certain structured directions 25–390$\times$ more strongly than along random ones. Any explanation of the collapse has to be consistent with both facts.

We analyze the collapse through what one optimizer step does to the output (§2). For a squared loss, a step is harmful when the change it causes in the output exceeds a radius set by the current residual, and under a gradient step that change is set by the gain of the empirical tangent kernel $K = JJ^\top$ of the write matrices along the residual. Our hypothesis: *directions of large gain are common and often useful; a step becomes harmful when the gain along the current loss gradient, times the step, exceeds the safety radius, whether because the kernel grows a direction or because the gradient rotates into one; and the terminal plateau is reached with the gain falling from its peak.* By *mechanism* we mean this one-step condition together with the links in it that we could observe and intervene on.

The paper makes three contributions. First, an output-space analysis of one optimizer step that separates the gain of the kernel from its coupling to the loss gradient, with the measurements that make the coupling observable on real AdamW runs (§2). Second, dynamic and interventional evidence on two tasks (§3): large directional gain also occurs in healthy states; the boundary falls with depth, as a power law over the depths tested, and depends on training state and residence; a Pre-LN comparison and a final-norm surgery locate part of the normalization gap in one backward factor; collapses follow task-specific trajectories to identifiable degenerate predictors; and two counterfactuals show that neither the amplified direction nor continued growth of the kernel is required for the instability. Third, GROW (§4–§5), a fixed spectral bound on the residual-stream write matrices that keeps their zero initialization and the Post-LN forward pass, used as an intervention on the geometry the first two parts implicate: at four depths it trains through a constant dose that collapses most AdamW runs at three of them, zero writer decay and a constant writer-step reduction do not reproduce this, and the scale of its fixed factor sets how much sample quality the intervention costs. The 8B production system in which the collapse was first observed is not rerun for this study; its single-run evidence is labeled as such wherever it appears (Appendix B).

2 OUTPUT-SPACE INSTABILITY IN DEEP RESIDUAL NETWORKS

This section sets up the quantities measured in §3–§5: an exact one-step identity for the squared loss, its kernel form for a linearized gradient step, and the quantities measured along real training.

2.1 ONE STEP, SEEN IN OUTPUT SPACE

Let $y = f(x;\theta) \in \mathbb{R}^m$ be the output on which the loss acts (the predicted velocity for the DiT; for the LM the final hidden state before the unembedding, which the full step also updates, so there the kernel is a sensitivity diagnostic and the loss change is measured directly; Appendix D). For a squared loss $L = \frac{1}{2}\|y-u\|^2$ with residual $e = y-u$, a step that moves the output by $\Delta y = s\,d$, $\|d\| = 1$, changes the loss by exactly

$$\Delta L(s) \;=\; e^\top \Delta y + \tfrac{1}{2}\|\Delta y\|^2 \;=\; -\,s\,\|e\|\cos\theta + \tfrac{1}{2}s^2, \qquad \cos\theta = -e^\top d/\|e\|, \tag{1}$$

a parabola in the step length (Figure 2a). Along a descent direction ($\cos\theta > 0$) the loss falls for $0 < s < r_{\text{safe}} = 2\,\|e\|\cos\theta$ and rises beyond it, however well aligned the step. The identity uses the realized output displacement, so it holds whatever the network does in between: the nonlinearity of $\theta \mapsto y$ is absorbed into Δy. Throughout we write g for the loss gradient in output space and, extending the squared-loss case where $g = e$, call it the residual (the local form for a general loss is in Appendix D).

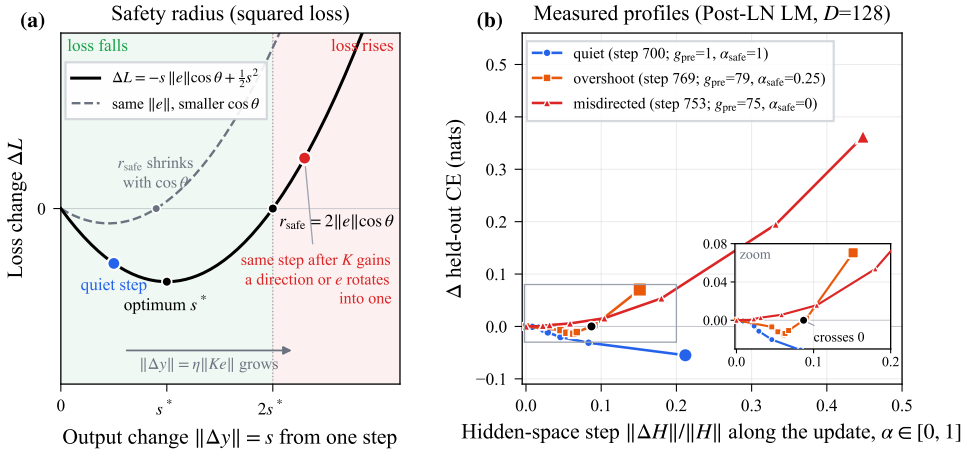


Figure 2: Output-space step safety. (a) The exact squared-loss change along a fixed output direction, Equation 1; weaker descent alignment shrinks the loss-decreasing radius (grey). (b) Held-out cross-entropy profiles along three full-parameter AdamW updates of a D=128 Post-LN LM, against the realized relative hidden-state displacement: a quiet step, an overshoot, and a misdirected step. Large markers are the committed updates.

2.2 Kernel gain and routes to instability

To first order $\Delta y = J\Delta\theta$ with $J = \partial y/\partial\theta$, restricted throughout to the *write matrices* of the residual stream — the attention output projections and MLP down projections W_l through which every sublayer l writes $u_l = W_l h_l$ into the stream (Elhage et al., 2021). In this linearized model a gradient step on the write matrices, $\Delta\theta = -\eta J^\top e$, moves the output by $\Delta y = -\eta\, Ke$ with $K := JJ^\top$, the empirical tangent kernel (Jacot et al., 2018) of the write matrices at the current parameters on a fixed probe batch, an $m \times m$ positive semidefinite matrix on output space whose eigenvalues λ we call the *gain* of the corresponding directions. Substituting into equation 1, for $\eta > 0$ and $e^\top Ke > 0$,

$$\Delta L_{\text{lin}} < 0 \iff \eta\,\Gamma(e) < 2, \qquad \Gamma(e) := \frac{\|Ke\|^2}{e^\top Ke}, \tag{2}$$

which for an eigenvector e reads $\eta\lambda < 2$, the stability condition of gradient descent on a quadratic (Cohen et al., 2021). $\lambda_{\max}$ gives the uniform bound, since $\eta\lambda_{\max} < 2$ makes every direction safe, whereas $\Gamma(e)$ depends on the current residual: a network with $\eta\lambda \gg 2$ along some direction is stable at that step as long as e has negligible weight on it, so the presence of high-gain directions does not by itself make an update harmful.

At fixed η there are three routes out of the safe region (Figure 2a). (1) **The kernel grows**: K acquires a direction of large gain that the residual overlaps. (2) **The residual rotates** into an existing high-gain direction. (3) **Alignment fails**: for a committed optimizer step, $\cos\theta \to 0$ and r_{safe} collapses (grey parabola) with no change in K. In the loss, routes 1 and 2 are one event, and they compound, because $J^\top e$ is largest along a direction of growing gain; re-measuring stored copies of the residual on later checkpoints separates growth of the kernel from rotation of the residual.

Depth enters through the responses. Since $u_{l,n} = W_l h_{l,n}$ at every token n, $K = \sum_l J_l J_l^\top$ with J_l the Jacobian of the output with respect to W_l, and J_l is built from the maps $M_{l,n} = \partial y/\partial u_{l,n}$ from sublayer l's write at token n to the output through every later sublayer and normalization (the token expansion is in Appendix D). A deeper stack has more terms and longer compositions, the normalization placement changes the maps $M_{l,n}$ (§3.2), and where the write matrices themselves enter these maps is taken up in §4. All kernel quantities are computed with exact Jacobian-vector products; the decomposition is used only to read them.

Two measurements follow. On a fixed held-out probe batch we report the residual's gain relative to random directions of the same norm, t/r (near 1 in health), and F_k, the fraction of its squared gain

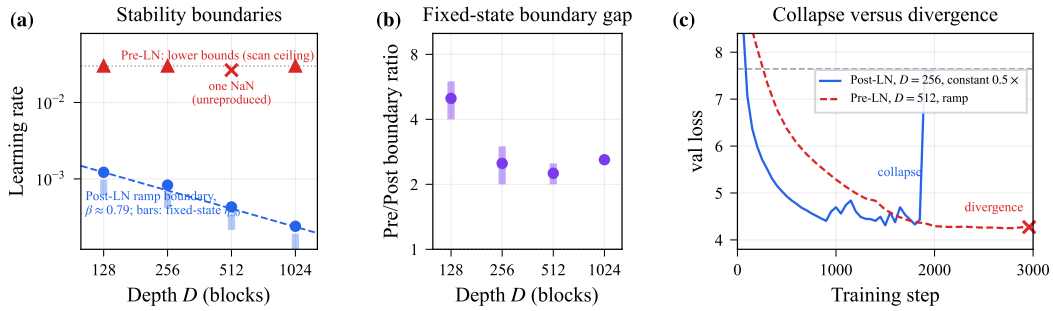


Figure 3: Pre-LN and Post-LN on the LM series (width 512). (a) Ramp boundaries against depth: Post-LN circles ($\beta \approx 0.79$) with fixed-state η_{50} bars; Pre-LN triangles are lower bounds at the scan ceiling, and the D=512 NaN (cross) was not reproduced. (b) Pre/Post boundary ratio from same-state cold jumps. (c) One run of each failure mode: a Post-LN D=256 constant-rate run reaches the unigram plateau; a Pre-LN D=512 ramp diverges (cross) with its loss still falling. Values in Table 8.

carried by the top-k eigenvectors of K; both describe the residual's coupling to high-gain directions, not Γ itself. For the committed AdamW step, whose scaling, clipping, and momentum change both its length and its direction, we measure the held-out loss along $\theta + \alpha\Delta\theta$ and record α_{safe}, the largest sampled fraction at which it has not increased (Figure 2b). Definitions, the write-matrix restriction, and its calibration are in Appendix D.

3 Collapse in diffusion transformers and language models

3.1 Platforms, probes, and labels

The DiT series has $D \in \{128, 256, 512\}$ Post-LN blocks ($L = 2D$ residual sublayers, width 768), an operator-by-operator scale-down of the 8B system of Appendix B, trained by rectified-flow velocity regression with zero-initialized write matrices. The LM series has $D \in \{128, \ldots, 1024\}$ blocks at width 512 on an educational web corpus, Post-LN as the main line and Pre-LN as the comparison. Both use AdamW with global clip 1.0 and QK-norm (full configurations in Appendix D). Kernel quantities are probed on a fixed held-out batch (§2.2); damage is labeled by validation loss, and terminal states are identified at the loss level against reference predictors. Boundaries come in three conventions — a *ramp* (the rate raised exponentially over 3000 steps until the run fails), *fixed-state* forks of a stored state at several constant rates (η_{50}), and *residence* under a constant rate below the ramp boundary — and the ramp boundary is always the most optimistic.

3.2 Pre-LN vs Post-LN: stability boundaries and failure modes

Under the ramp the failure boundary falls with depth on both tasks, consistent with a power law over the depths tested, with fitted exponents $\beta \approx 0.50$ for the DiT and $\beta \approx 0.79$ for the Post-LN LM (Figure 3a), and it is not a safety threshold: fixed-state forks of the LM fail at 0.5–$0.8\times$ the ramp boundary, and at D=256 a constant dose at $0.6\times$ the DiT's ramp boundary collapses five of six seeds after 334–1255 steps of residence, against none of three controls at $0.3\times$ (per-depth values and residence times in Appendix H).

Normalization shifts the boundary and the failure mode. Pre-LN tolerates larger learning rates than Post-LN in our LM series, with a 2–6$\times$ gap under fixed-state forks that shrinks with depth (Figure 3b), at no lower loss at each placement's own dose (Appendix G). Its observed failures are divergences rather than unigram-loss plateaus (Figure 3c). A backward-only intervention implicates one factor in part of the gap: at a healthy D=512 Pre-LN checkpoint, removing the final norm's backward $1/\mathrm{RMS}$ scaling, with the forward computation and the AdamW state unchanged, turns a rate at which the control trains cleanly into a divergence. This fixed-state intervention does not separate the gradient rescaling from its mismatch with the inherited moments (Appendix G). We

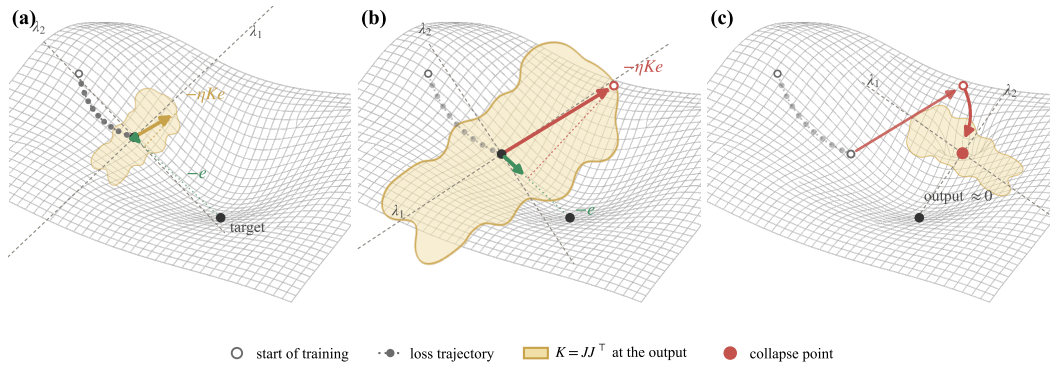


Figure 4: One DiT collapse (D=256, seed 3407) in output space, on an illustrative loss landscape whose minimum is the target. (a) Healthy, step 1850: the step $-\eta Ke$ stays low on the slope. (b) The collapse step, 1875: λ_1 has grown $\times 85$ in 25 steps, the top mode carries 0.87 of the residual's squared sensitivity $\|J^\top e\|^2$, and the step runs up the slope. (c) Dead, 1900–2700: the output norm has fallen to 0.05 and no longer moves, while K stays intact with λ_1 along the spatial mean. Directions and axis ratios follow the probes; lengths are compressed for drawing; the surface and the pit are schematic, and the amber contours depict kernel gain schematically rather than literal set images.

focus on Post-LN because it exhibits the persistent collapse studied here, and ask in §4 whether its boundary can be moved without changing its forward normalization.

3.3 Collapse dynamics

Large directional gain also occurs in healthy models. On 13 released checkpoints of 11 open models (1.7B–1.6T) and 52 training checkpoints from three families, every model amplifies a mean-channel direction 25–390$\times$ over white noise; in the base models the residual's gain is unremarkable ($t/r = 0.94$–1.11), post-training changes it by model, and along training trajectories the residual decouples from the high-gain directions early and stays decoupled while the gain regrows late (Appendix A). The controlled DiT shows the same with the very direction that will carry its failure: through the healthy segment of the D=256 ramp, t/r stays 1.5–1.9 while the spatial-DC direction is amplified 3.2–6$\times$ throughout.

The transition differs between the tasks. In the DiT it falls within one probe interval (Figure 4): between two probes 25 steps apart the top eigenvalue rises $\times 85$ and the top mode's share of the residual's squared sensitivity $\|J^\top e\|^2$ (its F_1) goes $0.21 \to 0.87$, the training loss doubles within the next ten steps, and the output norm falls to 0.05 with the top eigenvector the spatial mean. In the LM the transition takes longer and follows one order at $D = 256/512/1024$: t/r leaves its healthy band, a validation symptom follows within 20 steps, then bursts and a grinding phase of 100–450 steps in which transient single-sample amplifiers form and dissolve (Appendix C; the event timing of the fifteen collapses of the timing set, nine DiT and six LM, is in Appendix H). In that pool no damage followed a snapshot at which every stored direction's gain was shrinking; as an alarm, the plain pre-clip gradient norm is the best single predictor (AUC 0.72–0.77, Appendix H).

Terminal states. The DiT plateau lies at the zero predictor's loss (val 1.9367 against 1.9357) and the LM plateau at the unigram entropy (7.57–7.79 nats around 7.6444), consistent at the loss level with a near-zero and an approximately context-independent predictor. The gain that carried the run there has fallen from its event peak: the dead DiT's top eigenvector is pure spatial DC while its output carries no DC, and its top eigenvalue is 0.13–0.18 of its pre-collapse value after the ramp collapses and 0.7–4.5$\times$ it after the constant-dose collapses; the dead LM's gain along every stored direction is 0.2–0.4 of healthy (Appendix H).

Interventions distinguish the instability from its terminal state. Projecting every DiT write free of its spatial mean removes the zero-predictor plateau under the ramp but moves the boundary by only 8%: the kernel regrows its gain in another subspace and the run diverges instead. A tangent model frozen at a pre-collapse checkpoint also destabilizes within 100 steps, yet forms no plateau

within 2000 steps, where normal training from the same state completes its collapse in 52 steps. Neither the amplified direction nor continued growth of the kernel is therefore required for the instability in these systems; the experiments do not establish what is required for the plateau to form. Protocols and trajectories are in Appendix H.

4 GROW: SPECTRALLY BOUNDED RESIDUAL WRITES

The write matrices enter the maps M_l of §2.2 through their own sublayer. With $h_{l+1} = N_l\big(h_l + W_l\phi_l(h_l)\big)$ — ϕ_l the sublayer's read side, N_l the normalization after the add in Post-LN — one sublayer's Jacobian with respect to the stream is $A_l = N_l'(I + W_l\Phi_l)$ with $\Phi_l = \partial\phi_l/\partial h_l$, every $M_{l'}$ with $l' < l$ is a product of such factors, and the quadratic part of the Gram of $I + W_l\Phi_l$ satisfies

$$W_l\Phi_l\Phi_l^\top W_l^\top \;\preceq\; \|\Phi_l\|^2\, W_l W_l^\top \tag{3}$$

(full expansion in Appendix E): what one write matrix can add to its own sublayer's Jacobian, and hence to every map upstream of it, is controlled by its spectrum. At failure the write matrices carry exactly this signature. Across the five steps in which the 8B production run fails, the attention-output write matrices' largest singular value rises $2.95\times$ while their stable rank falls $18.5 \to 8.9$; on the controlled platform the collapsed $D{=}256$ AdamW DiT carries three times the write energy $\mathrm{tr}(WW^\top)$ per matrix of the healthy control; and holding $WW^\top$ fixed from before the failure step prevents the failure for the whole observed window, $\sim 20\times$ the failure step, on a small replica of the 8B system and on the system itself (single historical runs; Appendix B).

GROW (Gram-Retracted Optimization of Writes) holds every write matrix inside a fixed spectral bound $W_lW_l^\top \preceq 4C_0$ at every step, up to the logged Gram error, with the forward pass kept strictly Post-LN, so that by Equation 3 the quadratic term of every sublayer's Jacobian is bounded by a fixed-Gram envelope scaled by $\|\Phi_l\|^2$, namely $4\|\Phi_l\|^2C_0$: it is the optimizer step, not the normalization, that is asked to move the boundary.

Parameterization. Each write matrix $W \in \mathbb{R}^{d\times m}$ ($m \geq d$) is the difference of a trainable factor and a fixed one,

$$W \;=\; P - P_0, \qquad PP^\top \;\equiv\; P_0P_0^\top =: C_0 \quad \text{at every step,} \qquad \|P_0\|_F = \alpha, \tag{4}$$

with P_0 drawn once per write matrix from a seed. Two fixed factors are used: a normalized Gaussian matrix, $P_0 = \alpha R/\|R\|_F$, in the $\alpha{=}1$ evaluation and the scale sweep of §5, and a row-orthogonal one, $P_0 = (\alpha/\sqrt{d})\,Q_0$ with $Q_0Q_0^\top = I$, so that $C_0 = (\alpha^2/d)I$, in the final configuration. $P(0) = P_0$ gives $W(0) = 0$, so the zero initialization of the recipe is preserved exactly. Writing $P_0 = U_0S_0V_0^\top$, the constraint forces $P = U_0S_0V^\top$ for some V with orthonormal columns, so $\|W^\top x\| \leq 2\|P_0^\top x\|$ for every x: $WW^\top \preceq 4C_0$, $\sigma_i(W) \leq 2\sigma_i(P_0)$ for every i, and $\mathrm{tr}(WW^\top) \leq 4\alpha^2$, a fixed energy budget (Appendix E); for the row-orthogonal factor the envelope is isotropic, $\sigma_i(W) \leq 2\alpha/\sqrt{d}$. P_0 has full row rank, so its column space is the whole output space and the constraint carries no subspace restriction; what remains free is V, which input features feed which output directions, so new dominant directions can form inside the envelope, only not beyond it (Figure 5a). The bound is on each write matrix alone: it constrains neither the read side Φ_l, the direction of the cross terms of the Gram (their norm only through $\|W_l\|_2$), nor the composition across sublayers. It puts no floor on the stable rank independent of the energy, but a conditional one, $\mathrm{sr}(W) = \|W\|_F^2/\|W\|_2^2 \geq \mathrm{tr}(WW^\top)/4\lambda_{\max}(C_0)$, which for the row-orthogonal factor reads $\mathrm{sr}(W) \geq d\,\mathrm{tr}(WW^\top)/4\alpha^2$: a write matrix that spends energy cannot concentrate it in a few directions (Appendix E).

What the retraction removes. The constraint set $\{P : PP^\top = C_0\}$ has tangent space $T_P = \{Z : ZP^\top + PZ^\top = 0\}$ and, in the Frobenius metric, normal space $N_P = \{SP : S = S^\top\}$, of dimensions $dm - d(d+1)/2$ and $d(d+1)/2$: to first order the retraction removes the component of each candidate update along N_P, half of the parameters of a square write matrix and a third of the wide MLP one's. For the row-orthogonal factor the normal projection is explicit, $\Pi_N(Z) = (d/\alpha^2)\,\mathrm{sym}(ZP^\top)\,P$, and depends on α only through the scale of P: the direction filter does not weaken as α grows, only the envelope widens, so a large scale does not return GROW to AdamW. Along any finite displacement the identity $WP_0^\top + P_0W^\top = -WW^\top$ holds. GROW therefore acts in two places, a fixed envelope on the realized write and a direction selection on each candidate update; the runs of §5 do not separate which of the two carries the stability they show (Appendix E).

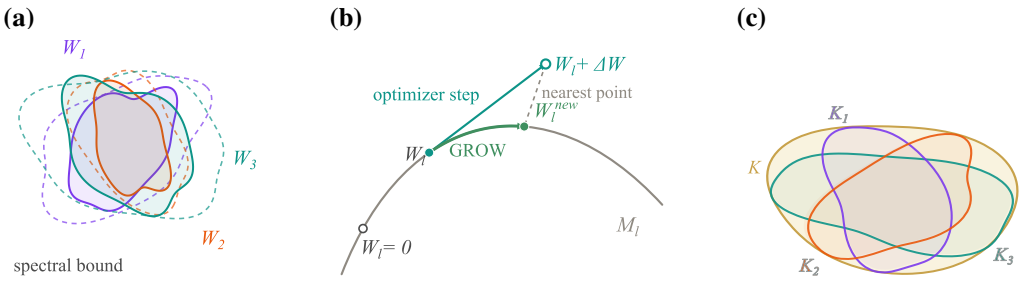


Figure 5: GROW places a fixed spectral envelope on the residual writes. (a) Weight space: the unit-ball images of three write matrices W_l (one color each) stay inside their own fixed dashed envelopes because $W_l W_l^\top \preceq 4C_{0,l}$. (b) One step in the same space: GROW retracts the AdamW candidate $W_l + \Delta W$ to the Frobenius-nearest point W_l^{new} on the shifted fixed-Gram set $\mathcal{M}_l = \{W : (W + P_{0,l})(W + P_{0,l})^\top = C_{0,l}\}$, which contains $W_l = 0$; Adam's moments are unchanged. (c) Output space: each write enters the kernel $K = \sum_l J_l J_l^\top$ through its sublayer's Jacobian factor $A_l = N_l'(I + W_l \Phi_l)$, and Equation 3 bounds the corresponding quadratic term by $4\|\Phi_l\|^2 C_{0,l}$ without imposing a fixed bound on K itself; the colored contours are the layer-wise terms $K_l = J_l J_l^\top$, drawn in the schematic gain geometry of Figure 4. Geometry is schematic.

The step. Every parameter takes the AdamW candidate step of the production trainer unmodified; weight decay is zero on the write matrices (decay applied to W before the retraction would keep the set and add a pull toward P_0; it is not used). GROW changes only the committed update of the write matrices: after the candidate step, each factor is returned to the constraint set by the nearest-point (polar) retraction computed from the fixed SVD of P_0,

$$P_{\text{cand}} = W_{\text{cand}} + P_0, \qquad P \leftarrow U_0 S_0 \,\text{polar}\big(P_{\text{cand}}^\top U_0 S_0\big)^\top, \qquad W \leftarrow P - P_0, \tag{5}$$

the Euclidean nearest point on the feasible set to the AdamW candidate W_{cand} (Figure 5b); Adam's moments are left untouched. For the row-orthogonal factor $S_0 = (\alpha/\sqrt{d})I$ and the retraction reduces to $P \leftarrow (\alpha/\sqrt{d})\,\text{polar}(P_{\text{cand}})$, which three to four Newton–Schulz iterations in fp32 deliver; the Gaussian factor needs an fp64 eigendecomposition, its polar problem having condition number $\text{cond}(P_0)^2$. The realized constraint error $\|PP^\top - C_0\|_F / \|C_0\|_F$ stays at 10^{-6} (1.0–1.2×10^{-6} Gaussian, at most 8.6×10^{-6} row-orthogonal) at every logged step of every run in this paper. Pseudocode, numerics, cost, the relation of the feasible set to the Stiefel manifold, the tangent-kernel split that would connect the direction selection to the gain of §2, and the three stages through which the form was reached are in Appendix E.

5 Results

GROW is evaluated as an intervention on the write geometry, not as a tuned training recipe. Every run uses the DiT of §3 at $D \in \{128, 256, 512\}$, extended to D=1024, with the production schedule (1000-step warmup, then constant), batch 128, and data streams identical step for step to the AdamW runs. Relative to AdamW, GROW also shortens the committed writer step (to a median 68% of the candidate norm at α=1) and drops weight decay on the write matrices; an AdamW control for each follows; a same-state counterfactual that re-adds the retracted component at matched norm, and the fixed-initial-tangent control of Appendix E, were not run. No comparison below is at equal compute: GROW arms run 4000 steps against 3500 for the references, plus the retraction.

Constant-dose stability at four depths. At $0.6\times$ each depth's AdamW ramp boundary, AdamW collapses after 1325–1675 steps of residence at D=128 (both seeds), 334–1255 at D=256 (five of six), and 250–480 at D=512 (both); at D=1024 the same dose is not lethal, all three seeds training to the end of the window. GROW runs the full 4000-step window without collapse in every arm at every depth: eleven runs across $\alpha^2 \in \{1, 2, 4, 8, 16\}$ and both factors at D=256 and, at α^2=16, two at each of $D = 128$ and 512 and three at D=1024 (Table 1), with whole-run g_{pre} peaks below 1.9 at α=1 against 241 for the violent AdamW arm (Appendix B). The dose is 0.4–$0.5\times$ GROW's

Table 1: Constant dose $0.6\times$ each depth's AdamW ramp boundary for 4000 steps (AdamW: 3500), and the AdamW references at $0.3\times$ for 3500 steps. Residence counts steps after the 1000-step warmup; FID-10k and IS under the frozen sampler protocol, per run; the D=256 reference counts two runs of seed 1234. Every scale-sweep arm of Table 4 also survived the window.

D	arm	dose	runs	collapsed (residence)	val	FID-10k	IS
128	AdamW	$0.6\times$	2	2/2 (1325, 1675)	1.914	—	—
128	GROW, α^2=16	$0.6\times$	2	0/2	0.875, 0.875	57.8, 56.9	18.3, 19.3
128	AdamW reference	$0.3\times$	2	0/2	0.874, 0.874	52.8, 58.1	19.6, 18.1
256	AdamW	$0.6\times$	6	5/6 (334–1255)	1.914; 0.876 alive	—	—
256	AdamW, writer decay 0	$0.6\times$	2	2/2	1.916, 1.916	—	—
256	AdamW, writer step $\times 0.68$	$0.6\times$	3	1/3	1.918; 0.862, 0.862 alive	—	—
256	AdamW	$0.45\times$	1	1/1	1.915	—	—
256	GROW, α=1, Gaussian	$0.6\times$	3	0/3	0.893–0.895	67.6	16.8
256	GROW, α^2=16	$0.6\times$	3	0/3	0.873, 0.872, 0.874	58.6, 54.6, 54.3	19.3, 19.7, 19.6
256	AdamW reference	$0.3\times$	5	0/5	0.880–0.882	53.3, 58.0, 57.6	19.8, 18.1, 19.6
512	AdamW	$0.6\times$	2	2/2 (250, 480)	1.914	—	—
512	GROW, α^2=16	$0.6\times$	2	0/2	0.877, 0.876	59.6, 56.1	18.5, 19.9
512	AdamW reference	$0.3\times$	2	0/2	0.886, 0.885	57.6, 56.0	18.8, 19.1
1024	AdamW	$0.6\times$	3	0/3	0.889, 0.889, 0.890	59.5, 60.4, 58.0	18.2, 18.0, 17.9
1024	GROW, α^2=16	$0.6\times$	3	0/3	0.887, 0.887, 0.888	63.3, 59.8, 54.8	18.2, 19.0, 19.6
1024	AdamW reference	$0.3\times$	2	0/2	0.900, 0.901	62.0, 58.5	18.3, 18.7

own ramp boundary; where GROW's long-residence boundary lies was not located. AdamW's, at D=256, lies between $0.3\times$ and $0.45\times$: one seed at $0.45\times$ (1.66×10^{-4}) collapses in the same window.

Neither control reproduces the survival across seeds. Two AdamW arms at the D=256 dose isolate these two changes, on the violently and a quietly collapsing AdamW seed. Zero weight decay on the write matrices alone leaves both collapsing (2/2 collapsed). Scaling the write matrices' rate by a constant 0.68, the median step-norm ratio of the α=1 retraction (not a per-step match, nor the ratio at α^2=16), with every other parameter at the full rate, rescues the quiet seed and not the violent one; a third seed, alive under AdamW too, survives (1/3 collapsed). Neither control gives GROW's survival across seeds (Appendix B).

The ramp boundary moves, by less as the scale grows. Under the ramp that locates the AdamW boundary at 3.72×10^{-4} at D=256, the α=1 run crosses that step without incident ($g_{\mathrm{pre}} = 0.48$ against 22.6) and collapses at 2.7–3.4$\times$ the boundary; at α^2=16 with the row-orthogonal factor the boundary is $1.63\times$ AdamW's at D=128, 1.43–1.51$\times$ at D=256, 1.12–1.26$\times$ at D=512, and $1.50\times$ at D=1024 (AdamW: 5.63×10^{-4}, 3.72×10^{-4}, 2.82×10^{-4}, 1.83×10^{-4}; Appendix H). The two ends of this comparison differ in the scale and in the factor, so the drop is consistent with the scale, not attributed to it. The ramp reads the rate tolerated over a few dozen steps per rate, and for AdamW it overstates the rate tolerated in residence at $D \leq 512$ (§3.2) while at D=1024 its $0.6\times$ dose is tolerated for the whole window; the constant-dose window above, not the ramp ratio, is the comparison the intervention is judged on.

Sample quality follows the scale of the fixed factor. As α^2 rises from 1 to 16 (Table 4, Appendix B) the median write energy $\mathrm{tr}(WW^\top)$ per matrix rises from 1.70 to 10.0, val falls from 0.893 to 0.873, and FID-10k falls from 67.6 to 58.6, against 57.6 for the healthy AdamW reference at $0.3\times$ and 3500 steps (seed 1234, energy 7.95). The energy grows more slowly than the budget, from $1.7\alpha^2$ to $0.63\alpha^2$: the committed steps move P a bounded distance in a fixed window, so the angle it turns from P_0 shrinks as the sphere grows (Equation 7). At α^2=16 the write energy exceeds the healthy reference's and the run still holds a dose that collapses AdamW: the stability does not require the energy to be held below the healthy level. These runs do not determine whether it depends on the envelope, the scale, or the direction selection of §4 (a Gaussian factor at the raw initialization scale collapses with the Gram exact; Appendix B). The row-orthogonal factor matches the Gaussian one within 0.1 FID at α^2=8 and 16 (one seed each), so in these runs the shape of the fixed Gram is not the main source of the remaining gap. In means over runs (Table 1), the final configuration ends 1.9 FID above the $0.3\times$ reference at D=128 (57.3 against 55.5), 0.4 below at

D=256 (55.8 against 56.3), 1.0 above at D=512 (57.8 against 56.8), and 0.9 below at D=1024 (59.3 against 60.2), where the dose is not lethal and AdamW at the same dose also reaches 59.3 over three seeds. The reference's own runs differ by 1.6–5.3 FID and GROW's by 0.9–8.5, so every gap above is inside the run-to-run spread, a descriptive statement rather than an equivalence test; a comparison against the reference's better run alone would read 0.8–4.5 FID. Val and FID rank the arms differently: at $D \geq 256$ the GROW arms end 0.007–0.013 below the reference in val without a matching FID advantage, and at D=128 the two are level in val. What these runs establish is that the intervention does not buy its stability by preventing learning; equal-compute equivalence with a tuned AdamW is not established.

Design conclusions. Three readings from historical single runs on the 8B system and its small replica fixed the form (Appendix B): a bound on the singular values alone, whether an isotropic cap, an isotropic manifold with free radius, or a random P_0 at the raw initialization scale, did not prevent the collapse; a normalized random P_0 matched the checkpoint-derived one; and keeping only the inequality without the equality on a factor kept the stability and lost learning quality.

6 Related work

Training instability and amplification. $K = JJ^\top$ is the empirical tangent kernel (Jacot et al., 2018) restricted to the write matrices and evaluated at the current parameters; the condition $\eta\lambda < 2$ of Equation 2 organizes the edge-of-stability literature (Cohen et al., 2021; 2022; Damian et al., 2023) and the catapult and break-even analyses of early training (Lewkowycz et al., 2020; Jastrzebski et al., 2020); our object replaces the sharpness by the gain along the residual on the committed step of an adaptive optimizer, and the outcome studied is a persistent degenerate plateau rather than oscillation at the edge. Loss spikes at scale have been attributed to attention-logit growth (Dehghani et al., 2023; Zhai et al., 2023; Wortsman et al., 2024), to Adam's second-moment estimate (Molybog et al., 2023), and to the embedding and normalization layers (Takase et al., 2025), and are often handled by restarting (Chowdhery et al., 2023). Liu et al. (2020) showed that too strong an amplification through the residual branch is unstable and too weak a one limits the model; what this paper adds to that trade-off is the direction of the amplification relative to the loss gradient, its evolution during training, and the degenerate terminal states. Attention sinks and massive activations (Xiao et al., 2024; Sun et al., 2024) are phenomenon-level counterparts of the fingerprints of §3.3.

Depth and normalization. DeepNet (Wang et al., 2022) bounds the model update at initialization and reports Post-LN stacks outperforming Pre-LN at a thousand layers once trainable; Mix-LN (Li et al., 2025), Peri-LN (Kim et al., 2025), and hybrid placements in production models (Gemma Team, 2025) show that placement is not a settled binary, deep Pre-LN layers have been reported to approach identity maps (Sun et al., 2025), and Xiong et al. (2020) give the gradient analysis of Pre-LN's advantage at initialization. Signal-propagation analyses and depth-scaled residual branches (Noci et al., 2022; Hayou et al., 2021; Zhang et al., 2019; Bachlechner et al., 2021; Touvron et al., 2021) address the forward pass at initialization; the profiles of Appendix G are measured on trained states. Hyperparameter transfer across width (Yang et al., 2021) and depth (Yang et al., 2024; Bordelon et al., 2024) prescribes how the learning rate should scale to preserve feature learning; our β is a failure-boundary exponent at fixed width, not an optimal-rate exponent.

Manifold- and Gram-constrained optimization. Constraining weights to a manifold and retracting after each step is classical (Absil et al., 2008). Manifold Muon (Bernstein, 2025) holds each weight matrix semi-orthogonal and takes the tangent direction of steepest linearized descent under a spectral-norm budget; Gram-space Manifold Muon (Keigwin et al., 2025) generalizes the constraint to a class of Gram matrices, of which a fixed factor Gram is one case; Yoshihara (2026) assign different manifolds to attention and MLP weights and report singular-value growth in attention weights as a source of instability; Cayley ADAM (Li et al., 2020) and Muon (Jordan et al., 2024; Liu et al., 2025) are the Adam-with-retraction and orthogonalized-update precedents (Appendix E). GROW's feasible set is a fixed linear image of the Stiefel manifold translated to pass through $W = 0$ (Appendix E); what it adds is the anchored difference parameterization that keeps the write matrices zero-initialized, the W-space nearest-point projection of the AdamW candidate (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) in place of a tangent steepest-descent step, and the residual-stream write matrices as the constrained family; a direct comparison with these optimizers on the same feasible set has not been made (Appendix B).

REPRODUCIBILITY STATEMENT

All controlled runs use deterministic data streams (seeded, rank-sharded, replayable sample for sample), instrumentation gated by configuration and validated against frozen numerical regression baselines (LM: 224 scalars; DiT: 76 scalars; worst relative difference 0.000), and fixed held-out probe batches. GROW is configuration-gated (default off reproduces the production trainer byte for byte); its reference matrices are derived from a single integer seed; constraint error, retraction fallbacks, energy, and step ratio are logged at fixed cadence in every run. Replication is protocol-matched rather than bit-exact: floating-point chaos shifts event timing by 150–380 steps without changing event types or order (Appendix D). All DiT runs share code, configurations, data streams, and the FID protocol, on four-GPU B200 and B300 nodes. Appendix F lists run families and protocols; boundary conventions are defined in §3.1. Code, configurations, and run ledgers accompany the submission.

AI USE STATEMENT

Generative AI tools were used to draft and polish the manuscript from the authors' experiment logs and fact archives, to write analysis and plotting code, to locate and check related work, and, on the 8B production system, to propose candidate interventions that the authors selected, implemented, and evaluated. Quantitative claims were checked by the authors against the run records and evaluation outputs (a number-to-source index is maintained alongside the manuscript), and all AI-assisted code was reviewed and validated against the frozen regression baselines above. The grey surface and the amber shape in Figure 4 were generated and cleaned by image models from text prompts and composited by the authors; every point, axis, and arrow drawn on them is computed from the logged measurements. Generative AI was not used to generate, select, or filter experimental results. The authors take responsibility for the final content of this work.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021)*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Jeremy Bernstein. Modular manifolds. Thinking Machines Lab blog, September 2025. URL https://thinkingmachines.ai/blog/modular-manifolds/.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2309.16620.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability, 2022. URL https://arxiv.org/abs/2207.14484.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://arxiv.org/abs/2209.15594.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, et al. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, December 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Gemma Team. Gemma 3 technical report, 2025. arXiv:2503.19786.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable ResNet. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1324–1332. PMLR, 2021. URL https://proceedings.mlr.press/v130/hayou21a.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Stanisław Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof J. Geras. The break-even point on optimization trajectories of deep neural networks. In *The Eighth International Conference on Learning Representations (ICLR 2020)*, 2020. URL https://openreview.net/forum?id=r1g87C4KwB.

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

Ben Keigwin, Dhruv Pai, and Nathan Chen. Gram-space manifold Muon. Tilde Research blog, October 2025. URL https://blog.tilderesearch.com/vignettes/gram-space.

Jeonghoon Kim, Byeongchan Lee, Cheonbok Park, Yeontaek Oh, Beomjun Kim, Taehwan Yoo, Seongjin Shin, Dongyoon Han, Jinwoo Shin, and Kang Min Yoo. Peri-LN: Revisiting normalization layer in the transformer architecture. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, Proceedings of Machine Learning Research. PMLR, 2025. URL https://arxiv.org/abs/2502.02732.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism, 2020. URL https://arxiv.org/abs/2003.02218.

Jun Li, Li Fuxin, and Sinisa Todorovic. Efficient Riemannian optimization on the Stiefel manifold via the Cayley transform. In *International Conference on Learning Representations (ICLR)*, 2020.

Pengxiang Li, Lu Yin, and Shiwei Liu. Mix-LN: Unleashing the power of deeper layers by combining Pre-LN and Post-LN. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2412.13795.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, et al. Muon is scalable for LLM training, 2025. URL https://arxiv.org/abs/2502.16982.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

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

Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, Binh Tang, Diana Liskovich, Puxin Xu, Yuchen Zhang, Melanie Kambadur, Stephen Roller, and Susan Zhang. A theory on Adam instability in large-scale machine learning, 2023. URL https://arxiv.org/abs/2304.09871.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ae0cba715b60c4052359b3d52a2cff7f-Abstract-Conference.html.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2013.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM 2024)*, 2024. URL https://arxiv.org/abs/2402.17762.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. URL https://arxiv.org/abs/2502.05795.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models. In *Second Conference on Language Modeling (COLM 2025)*, 2025. URL https://arxiv.org/abs/2312.16903.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pp. 32–42, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Touvron_Going_Deeper_With_Image_Transformers_ICCV_2021_paper.html.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*, 2022.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2309.14322.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2309.17453.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, 2020.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. URL https://arxiv.org/abs/2203.03466. arXiv version titled "Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer".

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs VI: Feature learning in infinite-depth neural networks. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2310.02244.

Kirato Yoshihara. Different layers, different manifolds: Module-wise weight-space geometry in transformer optimization, 2026. URL https://arxiv.org/abs/2606.13276. Accepted at the WSS workshop, ICML 2026.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *The Seventh International Conference on Learning Representations (ICLR 2019)*, 2019. URL https://openreview.net/forum?id=H1gsz30cKX.

A Released-model panel

The panel measures K on the final hidden state of 13 released checkpoints spanning 11 open models (1.7B–1.6T; dense and mixture-of-experts; two controlled base/post-trained pairs) and on 52 training checkpoints along three public training trajectories (Pythia-1.4B, 25 checkpoints; Pythia-6.9B, 12; OLMo-2, 15; two teams, two corpora). Table 2 lists the checkpoints. Only dimensionless quantities are compared across models: gains of structured directions relative to random cotangents of the same norm. For MoE models the kernel is the conditional one with recorded routing replayed bit-exactly (Appendix D).

Table 2: Released checkpoints of the panel (all public). Precision is that of the probe forward pass; the kernel is measured on the final hidden state with the write-matrix Jacobian, frozen-route conditional for MoE models. Probe configurations and per-checkpoint numbers are in the panel ledger.

checkpoint	**type**	**precision**	**role**
Qwen3-1.7B-Base	dense	fp32	base
Qwen3-8B-Base	dense	fp32, bf16	base
Qwen3-30B-A3B-Base	MoE	bf16	base; controlled pair
Qwen3-30B-A3B (Instruct)	MoE	bf16	post-trained; controlled pair
DeepSeek-V2-Lite	MoE	fp32	base
DeepSeek-V2 (236B)	MoE	bf16	base
DeepSeek-V3-Base (671B)	MoE	bf16, streamed	base
DeepSeek-V3-0324 (671B)	MoE	bf16, streamed	post-trained
DeepSeek-V4-Flash-Base (284B)	MoE	bf16	base
DeepSeek-V4-Pro-Base (1.6T)	MoE	bf16, streamed	base
GLM-4.7-Flash	MoE	bf16	post-trained
gpt-oss-20b	MoE	bf16	post-trained
gpt-oss-120b	MoE	bf16	post-trained
Pythia-1.4B (25 checkpoints, steps 0–143k)	dense	fp32	trajectory
Pythia-6.9B (12 checkpoints, steps 0–143k)	dense	fp32	trajectory
OLMo-2 stage 1 (15 checkpoints, 1B–2.2T tokens)	dense	bf16	trajectory

Structured gain is universal and architecture-specific. Every checkpoint carries structured directions at the top of K. The mean-channel direction is amplified 25–390$\times$ over white noise in every model; token-mean composite directions reach 9–36$\times$ in most models and are suppressed to 2–6$\times$ in the family with the heaviest attention-sink machinery. Models without an explicit sink mechanism amplify the position-0 direction 3–10$\times$; models with one suppress it to 0.04–0.8$\times$. The load sits on always-on write matrices: in MoE models the shared expert's share of the residual's squared gain grows with scale (48% $\rightarrow$ 77% $\rightarrow$ 87–89% from 16B to 671B) while the routed experts holding $\sim$85% of the write parameters carry 6–12%; one 671B model concentrates 84–87% in a single deepest-layer shared expert. One family built around non-expansive residual mixing shows the lowest structured gains on the panel at both 284B and 1.6T. We rank no vendors and draw no quality conclusions.

The residual is decoupled in base models; post-training varies. For every base model $t/r = 0.94$–1.11, with overlap onto the top eigenvector of order 10^{-6}–10^{-9}. On the one controlled base/instruct pair (same architecture, same base) the instruct model moves t/r from 0.94 to 2.44 with cross-entropy unchanged; one heavily post-trained family reads 11.2/23.1 (with a known raw-text distribution-shift confound); one 671B post-trained checkpoint reads 1.05 against its base's 1.03. Re-alignment depends on what the post-training does, not on the label.

Trajectories. At random initialization the high-gain cone already exists (token-mean 32$\times$) and the residual is inside it (t/r 5.3–5.7); early training keeps them aligned (t/r 2.2–4.0) while the loss falls fast; between 1k and 16k steps the residual decouples to $t/r \approx 1$ (top-eigenvector token-mean energy $0.92 \rightarrow 0.05$); late in training the gain machinery regrows (mean-channel $20 \rightarrow 138\times$; position-0 $1 \rightarrow$ 3–4$\times$) while the residual stays decoupled. All three trajectories reproduce this shape. The panel is static: it shows existence and decoupling, not causation.

B THE 8B PRODUCTION SYSTEM AND ITS SMALL REPLICA: SINGLE-RUN EVIDENCE

The phenomenon was first observed on an 8B, 256-block, width-2048 Post-LN DiT trained with flow-matching MSE, which collapsed spontaneously at step $\sim$400 of warmup (effective learning rate $\sim 9 \times 10^{-5}$), after the same stack had repeatedly hit a similar wall under adversarial training. No new 8B runs were performed for this paper; every controlled claim rests on the series of §3. Two systems appear below. The *8B system* is the production run, with forks of it and from-scratch reruns of its recipe (the unconstrained run collapses at step 400–405 across data streams). The *small replica* is a 64-block, width-64, 4-head Post-LN DiT of $\sim$2M parameters on the same recipe (16 tokens, batch 16, learning rate 16$\times$ the 8B base rate), whose unconstrained run collapses at step 360–408 with the same sequence of healthy learning, stable-rank climb, burst, and plateau; it was used to screen constraint variants over 8000 steps, about 20$\times$ its failure step, before any 8B run. A replica result does not transfer automatically: the $WW^\top$ lock applied at 75% of the failure step ran without a burst on the replica and met a non-lethal burst at 1.8$\times$ the failure step on the 8B system. Every reading below is one run per arm, and every window is stated as a multiple of the failure step.

A frozen near-death kernel diverges by rotation. A tangent model with the Jacobian frozen at the near-death state diverges monotonically with zero kernel growth by construction; a step-size-matched control (factor 0.6041, leaving the matched arm's function step 19% smaller than the true network's) still diverges nearly identically. Freezing a spectrally flat kernel instead pins the rotation gain at $\times 1.0$–1.18 and prevents the divergence, at an extreme learning cost; freezing a spread-spectrum kernel does not (rotation gain $\times 5.5$). The controlled three-state result of §3.3 is the trained-system version of this dichotomy.

The lesion family. Per-layer spatial-DC quotients of the stream remove the classic collapse but leave a high-pass learning plateau; restricting the quotient to the writes while giving DC a legitimate shallow path was the first stable variant; and Jacobian-source surgeries show the gain regrowing through other channels when one is cut. §3.3 reproduces the pattern at 1.13B.

Design ablations: full readings. Table 3 summarizes; the readings behind each row follow. *Isotropic cap* (replica, from scratch): $\sigma_i(W) \le \tau$ enforced by clamping, $\tau = k\, s_{\mathrm{ref}}$ at $k \in \{1.5, 2.25, 3.5\}$. The k=1.5 arm has the cap active from step 302 for 149 steps, clipping 44 singular values per step and pinning $\sigma_{\max}$ at 1.74 while $\mathrm{tr}(WW^\top)$ rises $1502 \to 2359$; it collapses at step 350, on schedule. The k=2.25 cap first activates 55 steps after the collapse, and k=3.5 never activates. The singular-value rise at failure (§4) is an expression of the collapse, not something a cap on it prevents. *Isotropic manifold with free radius* (replica, from scratch): $WW^\top = \rho^2 I$ with ρ trainable, tangent projection and an energy-preserving polar retraction, Adam's state kept. It collapses at step 5232 (13×): in the 200 steps before, the energy rises $1264 \to 3357$ and every singular value rises together ($0.38 \to 0.61$) while the stable rank stays at 64.00 and the isotropy error at 4×10^{-8}. A control that replays its per-step norms with plain Adam directions collapses at 3472 (8.7×), so most of the delay is the smaller step; the isotropic projection itself contributes ×1.5. *Difference without the Gram* (replica, from scratch): $W = P - P_0$ with P free collapses at 384 against 408 unconstrained. *Random P_0 at the raw initialization scale* (8B, from scratch): $P_0 \sim \mathcal{N}(0, 0.02^2)$ entrywise, full rank, with a budget $4\sum_l \|P_{0,l}\|_F^2 \approx 4 \times 10^6$ against a total write energy of $\sim 1.2 \times 10^3$ at the unconstrained failure. The run crosses the failure window (steps 336–470) without a trigger and leads every constrained variant in loss at step 500, then bursts at step 587 (g_{pre} $4.9 \to 12.0 \to 44.4$, the magnitude of the unconstrained failure) and lands on the plateau at ∼600 (1.48×) with no recovery to step 1950. The factor Gram stays exact (1.9×10^{-6}); before the burst W's energy is 1724, 1.41× the energy at the unconstrained failure, and within the burst the attention-output write matrices' $\sigma_{\max}$ rises 43% and their stable rank falls $20.7 \to 11.0$: the signature of §4 inside a legal motion on the factor manifold. *Healthy-checkpoint P_0 at $\alpha^2 \in \{1, 2, 4\}$* (replica, from scratch; $P_0 = \frac{1}{2}W_{300}$ of the same stream): all three arms survive 8000 steps (∼22×); realized energy $496 \to 851 \to 1374$ (the budget binds and each increment is used), step ratio $0.61 \to 0.67$, minimum loss 1.125–1.136 in all three. An 8B arm at α^2=2 crossed the failure window without a trigger and was stopped inside a burst at step 637, unjudged. *Unit random P_0* (replica, from scratch; the form of §4): survives 8000 steps (20×) with no burst, minimum loss 1.1485 against 1.1389 for the checkpoint frame, realized energy 43% of the budget, step ratio 0.656, constraint error 3×10^{-8}. An 8B run of this form was launched and not judged within the campaign; it is not used. *Inequality only* (8B, from scratch): $WW^\top \preceq C$ with C a healthy checkpoint's Gram, enforced through an auxiliary matrix $A = (C - WW^\top)^{1/2}$ with the interior AdamW step preserved to first order. It crosses the failure window and the burst window of the fixed-Gram run (steps 810 ± 150) without a burst (g_{pre} 0.58–0.74 against 10.7–11.0), riding the bound from step ∼405, but its loss lags and its samples under-converge, and it was stopped at ∼2250. At step 1000 its write matrices have stable rank 72 against 24 under the fixed-Gram factor, $\sigma_{\max}$ 0.060 against 0.132, and 73% of their energy in tail directions against ∼50% in the top eight: the bound is a ceiling, and the equality on the factor is what steers the energy into a few strong directions.

Two controls, one now run. A centered fixed-radius control, $C_0 = (\alpha^2/d)\, I$ with the same zero initialization, write matrices, AdamW candidate, and nearest-point projection, differs from the Gaussian factor only in the shape of the Gram; it is the row-orthogonal factor of §4, run at $\alpha^2 \in \{8, 16\}$ at D=256 and at α^2=16 at D=512 (§5). At equal α its realized energy is 7.54 against 6.83 and 10.05 against 9.21, and survival and FID are indistinguishable at single-seed resolution (Table 4). The tangent steepest-descent update of Manifold Muon (Bernstein, 2025), adapted to the anchored Gram and labeled as an adaptation, would compare the two update rules on the same feasible set. Neither the isotropic manifold with free radius nor the isotropic cap of Table 3 is the centered control.

Per-seed readings of the D=256 evaluation. At the constant dose 2.21×10^{-4}, the AdamW seeds paired with GROW collapsed at step 2255 (seed 1234, quiet, whole-run g_{pre} peak 2.6) and at step 1334 (seed 4242, burst, g_{pre} peak 241), and survived (seed 7777, val 0.876). The GROW seeds 1234/4242/7777 end at val 0.8928/0.8946/0.8937 with whole-run g_{pre} peaks 1.07/1.84/1.04 and median write energy $\mathrm{tr}(WW^\top)$ per matrix 1.69/1.70/1.69 (budget 4, near the $2\alpha^2$ point of Equation 7); the GROW ramp run's energy is 1.70–1.73 across the ramp's 12× span. The healthy AdamW control at 0.3× reaches val 0.8795–0.8817 at step 3500 at 7.95 per matrix (4.77 at step 2000); the collapsed AdamW arm at 0.6× sits at 23.3.

Readings from the scale sweep. Scale sweep at D=256, dose 2.21×10^{-4}, seed 1234, 4000 steps (Table 4): val 0.8857/0.8803/0.8761/0.8733 and median energy 2.91/4.65/6.83/9.21 for the

Table 3: Design ablations behind GROW, one run per row. "Failure step" is the step at which that system's unconstrained run collapses; replica: the small replica of the 8B recipe; 8B: the production system, from scratch unless a fork is stated. Full readings in Appendix B.

what is held	system	outcome
Bounding the spectrum without a fixed-Gram factor		
isotropic cap $\sigma_i(W) \leq \tau$, three τ	replica	collapses at the failure step at every τ; at the tightest the cap is active every step and $\sigma_{\max}$ is pinned while the energy rises $1502 \to 2359$
isotropic manifold $WW^\top = \rho^2 I$, ρ free	replica	collapses at $13\times$ the failure step by energy runaway ($\times 2.7$ in 200 steps); a control replaying its step norms with plain Adam directions collapses at $8.7\times$
$W = P - P_0$ with P unconstrained	replica	collapses at the failure step (384 against 408)
The factor's content and scale ($W = P - P_0$, $PP^\top = P_0 P_0^\top$)		
random P_0 at the raw initialization scale (budget $>10^3\times$ the failure energy)	8B	collapses at $1.48\times$ the failure step; factor Gram exact throughout; W's energy at onset $1.41\times$ the unconstrained failure energy
P_0 from a healthy checkpoint, $\alpha^2 \in \{1, 2, 4\}$	replica	all survive 8000 steps ($\sim 22\times$); realized energy $496 \to 851 \to 1374$
$P_0 = R/\|R\|_F$ random, unit norm (GROW)	replica	survives 8000 steps ($20\times$); loss within 0.01 of the checkpoint-P_0 arm
Bound only against fixed Gram; the update rule		
$WW^\top \preceq C$ only (dilation; no factor)	8B	no collapse and no burst to 2250 steps; worse loss; stable rank 72 against 24 under the fixed-Gram factor
$WW^\top = C_0$ from a healthy checkpoint, applied before / after the failure step	replica; 8B	before: no failure in 8000 steps ($\sim 20\times$; 8B: $19.8\times$, one non-lethal burst); after: delayed $17.5\times$, then collapses (replica)
five-arm replay from one $WW^\top = C_0$ state (fork at step 500 through a burst window to 1000)	8B	Adam state kept: survives; moments reset: collapses ~ 790; scale released (a free): collapses ~ 800; unconstrained, step norms matched: collapses in 8 steps; first moment reset only: survives, milder burst
update applied in the factor space against W-space retraction (25-step replays)	8B	the factor-space update delivers 21–27% of the allowed descent; the W-space retraction 99.6%

Gaussian factor at $\alpha^2 = 2/4/8/16$, FID-10k 65.69/62.83/62.52/58.73, IS 17.66/18.06/18.26/18.91; row-orthogonal factor at $\alpha^2 = 8/16$: val 0.8770/0.8729, energy 7.54/10.05, FID 62.48/58.62, IS 18.35/19.31. Ramps with the row-orthogonal factor at α^2=16: D=256 last healthy probe at 5.32×10^{-4}, first collapsed at 5.63×10^{-4}; D=512 collapsed at $\approx 3.55 \times 10^{-4}$ (the run stopped at step 2480 when the polar iteration's residual exceeded its guard inside the collapsed state). D=512 at 1.69×10^{-4} ($0.6\times$), 4000 steps: seeds 1234/4242 end at val 0.8767/0.8758, FID 59.61/56.09, IS 18.50/19.90; the AdamW reference at 8.5×10^{-5} ($0.3\times$), 3500 steps, seed 1234, at val 0.8864, FID 57.61, IS 18.79, and seed 4242 at val 0.8852, FID 55.99, IS 19.13 (mean 56.80). The AdamW D=512 arms at $0.6\times$ collapse 250 and 480 steps into residence by the event convention of Appendix D (275 and 500 by the val criterion); the D=128 arms at 3.4×10^{-4} collapse 1325 and 1675 steps into residence.

Later batch (Table 1). Same protocol, row-orthogonal factor at α^2=16 unless AdamW, probes thinned to every 100 steps (heavy every 1000, two modes, no residual or response profile; the thinning changes no update). D=128: GROW ramp boundary 9.19×10^{-4}; GROW at 3.4×10^{-4} seeds 1234/4242 end at val 0.8749/0.8747, FID 57.79/56.90, IS 18.26/19.31, median energy 15.2 (seed 1234); AdamW at 1.7×10^{-4}, 3500 steps, seeds 1234/4242, val 0.8740/0.8742, FID 52.81/58.14, IS 19.64/18.07 (mean 55.47). D=256: AdamW at 1.1×10^{-4}, 3500 steps, seeds 1234/4242, val 0.8804/0.8800, FID 53.33/57.99, IS 19.78/18.05; with the seed-1234 run of the scale sweep's reference (FID 57.56) the three-run mean is 56.29; GROW at 2.21×10^{-4} seeds 4242/7777 val 0.8724/0.8735, FID 54.62/54.29, IS 19.72/19.65; AdamW with writer weight de-

Table 4: Scale sweep at D=256, constant dose 2.21×10^{-4}, 4000 steps, one seed each (the α=1 row is seed 1234 of the three-seed evaluation). Energy is the median $\mathrm{tr}(WW^\top)$ per write matrix at the end of the window; G: Gaussian fixed factor, R: row-orthogonal.

α^2	factor	budget $4\alpha^2$	energy	val	FID-10k	IS
1	G	4	1.70	0.893	67.6	16.8
2	G	8	2.91	0.886	65.7	17.7
4	G	16	4.65	0.880	62.8	18.1
8	G	32	6.83	0.876	62.5	18.3
8	R	32	7.54	0.877	62.5	18.4
16	G	64	9.21	0.873	58.7	18.9
16	R	64	10.05	0.873	58.6	19.3
AdamW reference, 0.3×, 3500 steps, seed 1234			7.95	0.880	57.6	19.6

cay 0, seeds 4242/1234, val 1.9163/1.9159 (both collapsed); AdamW with the write matrices' rate scaled by 0.68, seeds 4242/1234/7777, val 1.9177 (collapsed) / 0.8616 / 0.8618 (alive); AdamW at $0.45\times$ (1.66×10^{-4}), seed 1234, val 1.9152 (collapsed). D=1024 (4.5B): AdamW ramp boundary 1.83×10^{-4} (last healthy probe 1.78×10^{-4} at step 1250, first collapsed 1.89×10^{-4} at 1275); GROW ramp boundary 2.74×10^{-4}; AdamW at 1.1×10^{-4} ($0.6\times$), 3500 steps, seeds 1234/4242/7777, val 0.8890/0.8885/0.8902, FID 59.51/60.35/57.98 (mean 59.28, sample s.d. 1.20), IS 18.19/18.03/17.87; GROW at 1.1×10^{-4}, 4000 steps, seeds 1234/4242/7777, val 0.8868/0.8867/0.8877, FID 63.31/59.83/54.78 (mean 59.31), IS 18.24/19.01/19.61, median energy 1.30 at step 1300 of seed 1234; AdamW at 5.5×10^{-5} ($0.3\times$), 3500 steps, seeds 1234/4242, val 0.8995/0.9009, FID 61.99/58.48, IS 18.28/18.67. Peak memory at D=1024 was 127 GiB per GPU without activation checkpointing.

When the constraint is applied, and what it discards. Holding $WW^\top$ at a healthy checkpoint's Gram from a step at $\leq$75% of the failure step prevents the collapse for the whole observed window (replica: 8000 steps, $\sim 20\times$, with learning; 8B: 8000 steps, $19.8\times$, with one non-lethal burst); holding it from just after that step delays the collapse $17.5\times$ on the replica, after which the run collapses with the constraint still exact. Under the constraint the gradient that would have produced the failure still arrives on the 8B system (pre-clip gradients on the write matrices $\times 12.5$, 94% of the burst's energy in the attention-output writes) and is removed by the retraction; three such bursts were survived. In a five-arm replay from one state of that run (fork at step 500 through the burst window to step 1000, same data stream): the constraint with Adam's accumulated state left untouched survived (burst peak 44.9, self-healed); resetting both moments, with the gradient projected before the clip, collapsed at $\sim$790; releasing the scale, $WW^\top = a^2 C_0$ with a free, collapsed at $\sim$800 with a frozen at 3.4 by the burst; the unconstrained arm with its step norms matched to the released arm collapsed within 8 steps (g_{pre} 56.4); and resetting only the first moment survived with a milder burst (loss peak 1.21 against 1.97) and a longer tail. Pulling the gradient back to the factor space instead of retracting in W-space delivers only 21–27% of the descent the constraint allows, against 99.6% for the W-space retraction (25-step same-batch replays). These readings are on the historical $WW^\top$-locked form; their transfer to the $PP^\top$ form of §4 is assumed, not tested.

C LM AMPLIFIER ANATOMY

The finest picture of what grows was captured on an earlier bare configuration of the LM series (D=128, no QK-norm, scaled initialization, constant LR 6×10^{-4}; probes every 25/100 steps); §3.1 records that the main line reproduces its boundary and event order.

Onset. After $\sim$600 steps of healthy training ($t/r \sim 1.4$; top eigenvector a deep-layer token-mean direction), the top eigenvalue jumps $\times 22$ in 100 steps, and the new top eigenvector belongs to no preregistered structured family: $\sim$100% of its energy sits in one sample of the probe batch, $\sim$85% within token positions 627–633 of that sample (0.595/0.127/0.060 at positions 627/628/633), channel-diffuse, cross-sample coherence 0.06, hosted in the shallowest quarter of the stack with the MLP down-projections carrying 0.87. The positions straddle an unremarkable text span, and the

per-position energy of the residual is unchanged from healthy runs (top-5 position energy 0.001–0.002 in both; masked-gain ratios 0.9–1.4): the amplified direction lives in the channel dimension *at* positions, not on rare tokens.

Reload. Within 100 steps t/r runs $1.4 \to 8.6 \to 36.3$ and the residual's depth profile flips from deep ($q_3 = 0.83$) to the eigenvector's shallow profile ($q_0{+}q_1 = 0.95$). The gain is closed quantitatively: $\text{overlap} \times \sqrt{\lambda_0} = 0.031 \times 5.78 \times 10^4 \approx 1793$ against 1809 measured (0.9%). Stored directions confirm growth of the kernel rather than rotation alone: every historical residual direction gains $+3.1$ to $+5.2$ in log squared gain at this snapshot.

Transience. The stored top direction's gain falls from 5.78×10^4 to 336 within 11 steps (a burst with $g_{\text{pre}} = 127.7$ and a clip intervene), to 281 at $+50$ steps, and to 123 on the plateau, against a healthy baseline of 171. Grinding is therefore a cycle: transient amplifiers form (lifetime $\lesssim$10 steps), the residual's small overlap with them produces a burst, burst and clip partly destroy the amplifier, and it rebuilds, each cycle ratcheting the loss upward ($4.6 \to 7.7$). Committed-step profiles lock α_{safe} to this phase: quiet steps 1.0 with $\cos(g, d) \approx -0.23$ to -0.25; damaging bursts 0–0.25 even though the directional derivative is usually still negative (in three burst steps the committed direction is uphill at the pre-step parameters); repair bursts 1.0, recovering up to 1.59 nats in one step. The linearity error is large in every burst-zone step regardless of sign, and its explosion precedes the first large burst by $\sim$80 steps.

Not a probe artifact. At an amplifier-period snapshot, 16 of 16 independent held-out batches show the giant local direction (median top eigenvalue 9.7×10^{10}, median max-sample concentration 0.990, write matrices' share of the full-parameter squared gain 0.42–0.43). Cross-batch transfer is far above random (parameter-space $\cos(g_A, g_B)$ 0.12–0.17 against $\sim 10^{-4}$ at this dimension; up to $|\cos|$ 0.89–0.94 at a later state of a second seed), while each batch's own expression is 5–40$\times$ stronger than imported directions: one parameter-space substrate, per-input token-local faces.

The loop is cooperative. Masked-update forks from a pre-onset state (full gradient and clip computed; the commit masked for one parameter family) show that only the full update explodes: masking the MLP down-projections (20% of parameters), the attention output projections (7.2%), all write matrices (27%), or all non-write parameters (73%) each blocks the onset in a 400-step window, and every masked arm trains better than the full run (val 4.38–4.63 against 4.72). Static replay attribution mis-ranks these families (attention-out alone shows no effect off-trajectory, and no single family reproduces more than a sliver of the full $\times 1475$ eigenvalue growth): the runaway is a joint loop between the write matrices and the features they read, not a bad parameter block.

D Measurement framework and QA

Platforms. *DiT.* $D \in \{128, 256, 512\}$ blocks, width 768, MLP hidden 1152, 12 heads with GQA 2:1, $\approx$0.57B/1.13B/2.27B parameters. Post-LN residual blocks (add, then RMSNorm), non-learnable QK-norm, SwiGLU, 2D RoPE on image tokens, write matrices and final projection zero-initialized (so at $t{=}0$ each block reduces to its normalization, which is the identity on an already-normalized stream), no output norm, no explicit timestep input (t enters only through x_t). Rectified-flow velocity regression, $x_t = (1-t)x_{\text{data}} + t x_{\text{noise}}$, target $x_{\text{data}} - x_{\text{noise}} = -\mathrm{d}x_t/\mathrm{d}t$ (the velocity toward the data), $t \sim$ sigmoid with α-shift 2, on frozen-autoencoder latents ($[128, 16, 16]$ at 256 px, patch size 1, 256 image tokens) with caption embeddings from a frozen 0.6B text encoder projected and concatenated as $\leq$64 prefix tokens (condition dropout 0.1). AdamW $\beta = (0.9, 0.999)$, weight decay 0.1, global clip 1.0, global batch 128. *LM.* Causal LMs at fixed width 512, $D \in \{128, 256, 512, 1024\}$ blocks, 0.46B/0.87B/1.70B/3.34B parameters, head dimension 64, sequence length 1024, GPT-2 BPE on an educational web corpus (unigram entropy 7.6444 nats). Post-LN main line with QK-norm and depth-scaled initialization of the write matrices (standard deviation $\propto 1/\sqrt{2L}$; a zero-initialized control gives the same $D{=}128$ boundary, 1.19 against 1.22×10^{-3}, and the same terminal state); Pre-LN comparison axis; AdamW with global clip 1.0. The fine-grained amplifier anatomy of Appendix C was captured on an earlier bare configuration ($D{=}128$, no QK-norm, constant LR 6×10^{-4}) whose $D{=}128$ boundary (1.22–1.29$\times 10^{-3}$) coincides with the main line's (1.22–1.30$\times 10^{-3}$).

Output space, write matrices, depth. For the LM, y is the final hidden state before the unembedding (bit-identical to the logits' input); for the DiT, the predicted velocity. The write matrices are the attention output projection and the MLP down-projection of every block (two residual sublayers per block). D counts blocks and $L = 2D$ residual sublayers; the boundary exponents β are fitted against D (fitting against L changes nothing). Ramp protocols: $10^{-5} \to 10^{-2}$ over 3000 steps for the DiT and $10^{-5} \to 3 \times 10^{-2}$ over 3000 steps for the LM, so the LM ceiling is 3×10^{-2}.

General losses, and the token expansion of K. For a general loss ℓ the analogue of Equation 1 is local: $\Delta\ell(s) = s\, g^\top d + \frac{1}{2} s^2\, d^\top H_\ell d + O(s^3)$ with $g = \nabla_y \ell$ and H_ℓ the output-space Hessian, so, for $g^\top d < 0$ and $d^\top H_\ell d > 0$, the second-order radius along d is $2|g^\top d|/(d^\top H_\ell d)$, an approximation rather than the exact radius of the squared loss; for cross-entropy on logits z, $H_\ell = \mathrm{diag}(p) - pp^\top$ per token, not the identity. For the LM, y is the hidden state before the unembedding and the map from y to the logits is itself updated, so the profiles of Figure 2b are measured directly rather than computed from this expansion. Since $u_{l,n} = W_l h_{l,n}$ at every token n, the Jacobian of the output with respect to the shared matrix W_l is $J_l = \sum_n M_{l,n}(h_{l,n}^\top \otimes I)$, hence

$$K \;=\; \sum_l J_l J_l^\top \;=\; \sum_l \sum_{n,n'} \langle h_{l,n}, h_{l,n'} \rangle\, M_{l,n} M_{l,n'}^\top; \tag{6}$$

the diagonal terms $n = n'$ read as the squared input a write matrix sees times the response $\rho_{l,n}(q) = \|M_{l,n}^\top q\|$ of the output to a write at that depth, and the cross terms couple tokens whose inputs to W_l are correlated. Equation 6 is used only to read the measured kernel, never to estimate it.

Kernel quantities. $J = \partial y / \partial \theta_W$ is evaluated at the current parameters on a fixed held-out probe batch; $K = JJ^\top$. The gain of an output direction q is $\|J^\top q\|$ (one vector–Jacobian product). The residual direction is the normalized loss gradient with respect to y (teacher-forced cross-entropy for the LM; the MSE residual for the DiT). t/r divides its gain by the mean gain of random directions of the same norm (white noise; per-token norm-matched; spectrum-matched variants in the panel); pooled values are not comparable across architectures, and only within-run changes are ever interpreted. Top eigenpairs of K are computed by deflated power iteration with exact forward-mode JVPs, guarded by the residual $\|Ku - \lambda u\|/\lambda$; each eigenvector is decomposed onto interpretable families (token-mean, batch-coherent, position-0, low-DCT, mean-channel for the LM; spatial DC over image tokens for the DiT), with concentration statistics and depth-quartile attribution. $F_k = \sum_{m \le k} \lambda_m \,\mathrm{align}(u_m, q)^2 / \|J^\top q\|^2$ (in an exact, consistent decomposition $F_k \in [0, 1]$; in degenerate states the normalization can fail and F_k can exceed 1 — one fork arm read 464 — and such values are unreliable). Residual directions are stored at probe steps and re-measured at later checkpoints; the log ratio of the squared gain then to the squared gain when stored separates growth of the kernel from rotation of the residual. The response profile is $\rho_l(q) = \|M_l^\top q\|$ across all sublayers (one backward per direction); for Pre-LN the raw profile before the final norm is recorded as well.

One-step diagnostics. Adam's step is not $-\eta J^\top e$ (Kingma & Ba, 2015): per-coordinate scaling, global clipping (Pascanu et al., 2013), and momentum change its length and direction, and the training update comes from training batches and from all parameters, whereas K is restricted to the write matrices and evaluated on a held-out batch (calibrated in Table 5). The kernel diagnostics t/r and F_k therefore characterize the coupling of the residual to the kernel on the probe batch and are not measurements of Γ in Equation 2. Along a committed update $\Delta\theta$ of all parameters we record: the held-out and training line profiles $L(\theta + \alpha\Delta\theta)$ for $\alpha \in \{0, 2^{-7}, \ldots, 1\}$; α_{safe}, the largest grid α at which the held-out loss has not increased (tolerance 10^{-4}); the linearity error of the first-order loss prediction; the hidden-space step $\|J\Delta\theta\|/\|y\|$ and its nonlinearity $\|y(\theta + \Delta\theta) - y(\theta) - J\Delta\theta\|/\|J\Delta\theta\|$ (118% on the burst step of Figure 2b, which is why the profiles are plotted against the realized displacement rather than α). The profiles separate steps that a training log does not: the overshoot and the misdirected step of Figure 2b are the same spike of the pre-clip gradient norm g_{pre}.

Event conventions. Two conventions are used and give the same order in every recorded collapse: the offline scoring below, used for the event counts and lead times quoted in the text, and the threshold detectors of Appendix H, used for the figures. Offline scoring: a geometry event is

a top-eigenvalue growth of $\geq 4\times$ between heavy probes with a change of the eigenvector's family, or a stored direction's squared gain rising by more than a factor e; an alignment event is t/r above twice the run's healthy median on consecutive probes with the residual's depth profile migrating into the top eigenvector's; a damage event is a positive 100-step windowed loss slope or a persistent validation excursion above the prefix minimum (the label used for the alarm comparison of §3.3 is this loss-based one: validation loss above its prefix minimum by 0.15 without recovery, so the gradient-norm alarm is not scored against a gradient-norm label); a persistent gradient-norm excursion is recorded separately as a burst event. Windowed averages only; no single-point calls. Same-protocol twins reproduce event types and order while timing drifts by 150–380 steps under floating-point chaos; every replication claim in the paper is protocol-matched, never bit-exact.

Table 5: Measurement QA checks on the probe machinery.

check	**result**
JVP↔VJP identity	rel. error 2.8×10^{-8} (DiT, fp32); 0.999999–1.000000 (panel, fp32); 0.949–1.032 (bf16)
finite-difference cross-checks	response-profile FD injection, worst $\lvert\text{ratio} - 1\rvert = 0.000$; FD-JVP valid in fp32 only; the FD step must scale with the stream RMS in degenerate states
tangent model	JVP linear to machine precision; gradient vs. central difference 2.1×10^{-6}
frozen-route replay (MoE)	$\max \lvert\Delta y\rvert = 0$ across six architectures
write matrices vs. full parameters	released healthy: 0.83 of the residual's full gain (0.68–0.69 of the Rayleigh quotient); LM pathological states: 0.38–0.43 of the full squared gain, stable across 16 held-out batches
probe-batch independence	giant local directions on 16/16 independent batches (two seeds)
determinism and regression	double forward bit-identical; instrumentation gated by frozen baselines (LM 224 scalars, DiT 76 scalars; worst rel. diff 0.000)

Labels. A collapsed LM is called dead when val exceeds its prefix minimum by a persistent margin; a DiT when val enters the band $[1.914, 1.94]$ for three consecutive evaluation windows; a non-finite streak is labeled a divergence. A conditioning-use score (loss with the caption or the first half of the context shuffled, minus the normal loss) was logged alongside and is not used in this paper. Kernel quantities are never death criteria.

E METHOD DETAILS

The energy budget. Taking traces in $W = P - P_0$ with $\|P\|_F = \|P_0\|_F = \alpha$,

$$\operatorname{tr}(WW^\top) \;=\; 2\alpha^2 - 2\langle P, P_0\rangle \;\in\; [0,\, 4\alpha^2] \tag{7}$$

by Cauchy–Schwarz: the energy a write matrix can spend is a fixed budget, with $2\alpha^2$ its value once P has rotated away from P_0. The ordered singular values of every write matrix are bounded by twice those of the fixed factor: quarter-circle distributed for a square Gaussian P_0 with the largest $\approx 2\alpha/\sqrt{d}$, and all equal to $\alpha/\sqrt{d}$ for the row-orthogonal one, so no single direction can acquire a large singular value. Algorithm 1 gives one training step: lines 1–3 are the AdamW step, lines 4–10 the retraction.

The sublayer Gram. With $A_l = N_l'(I + W_l\Phi_l)$ as in §4,

$$(I + W_l\Phi_l)(I + W_l\Phi_l)^\top \;=\; I + W_l\Phi_l + \Phi_l^\top W_l^\top + W_l\Phi_l\Phi_l^\top W_l^\top; \tag{8}$$

the last term is bounded by $\|\Phi_l\|^2\, W_l W_l^\top$ (Equation 3) and has its range in the column space of W_l, while the cross terms are bounded in norm by $2\sigma_{\max}(W_l)\|\Phi_l\|$ and are not so confined. A new strong output direction u of W_l with singular value σ adds a term of order $\sigma^2\|\Phi_l\|^2 uu^\top$ to the quadratic part, which the maps $M_{l'}$ upstream inherit. Under GROW, $W_l W_l^\top \preceq 4C_0$ gives $W_l\Phi_l\Phi_l^\top W_l^\top \preceq 4\|\Phi_l\|^2 C_0$, a bound on one term of one sublayer's Jacobian: it constrains neither the read side Φ_l, the direction of the cross terms, nor the composition across sublayers. For the square or wide Gaussian P_0 of §4 the column space of P_0 is the whole output space, so the bound

Algorithm 1 One step of GROW with AdamW

Require: parameters $\theta = (\{W_l\}_{l \le L}, \varphi)$ with write matrices $W_l \in \mathbb{R}^{d \times m}$; AdamW state; fixed factors $P_{0,l} = U_l S_l V_l^\top$, drawn once from a seed, Gaussian ($\alpha R_l / \|R_l\|_F$) or row-orthogonal ($(\alpha/\sqrt{d})Q_l$, $S_l = (\alpha/\sqrt{d})I$) ($W_l(0) = 0$, so $P_l(0) = P_{0,l}$); rate η, weight decay λ

1: $g \leftarrow \nabla_\theta \mathcal{L}$ on the batch; apply the global clip
2: update the AdamW moments from g {unmodified}
3: $\theta \leftarrow \theta - \eta\, \hat{m}/(\sqrt{\hat{v}} + \epsilon)$; decay φ by $\eta\lambda$; no decay on $\{W_l\}$
4: **for** each write matrix $l = 1, \dots, L$ {sharded across ranks} **do**
5: $\quad P_{\mathrm{cand}} \leftarrow W_l + P_{0,l}$
6: $\quad$ **if** Gaussian factor **then**
7: $\quad\quad B \leftarrow P_{\mathrm{cand}}^\top U_l S_l$ $\{m \times d\}$
8: $\quad\quad Q \leftarrow \mathrm{polar}(B)$ {nearest matrix with orthonormal columns; fp64 Gram-eigh + Newton–Schulz, fp64 SVD fallback}
9: $\quad\quad P_l \leftarrow U_l S_l Q^\top$ $\{P_l P_l^\top = U_l S_l^2 U_l^\top = C_{0,l}\}$
10: $\quad$ **else**
11: $\quad\quad P_l \leftarrow (\alpha/\sqrt{d})\,\mathrm{polar}(P_{\mathrm{cand}})$ {row-orthogonal; fp32 Newton–Schulz}
12: $\quad$ **end if**
13: $\quad W_l \leftarrow P_l - P_{0,l}$ $\{\sigma_i(W_l) \le 2\sigma_i(P_{0,l}),\ \mathrm{tr}(W_l W_l^\top) \le 4\alpha^2\}$
14: **end for**
15: log $\|P_l P_l^\top - C_{0,l}\|_F / \|C_{0,l}\|_F$, $\mathrm{tr}(W_l W_l^\top)$, $\|\Delta W_l\| / \|\Delta W_l^{\mathrm{cand}}\|$

constrains the spectral profile of W_l, how large and how concentrated its singular values can be, not the subspace it writes into; it is an upper bound on the spectrum and the energy and implies no lower bound on the stable rank independent of the energy (the conditional bound is below). For the row-orthogonal factor $C_0 = (\alpha^2/d)I$ and the envelope is isotropic.

Relation to the Stiefel manifold. For the full-row-rank P_0 of §4, $C_0 = P_0 P_0^\top \succ 0$, and with $Q = C_0^{-1/2} P$, $Q_0 = C_0^{-1/2} P_0$ the constraint $PP^\top = C_0$ becomes $QQ^\top = I$ (the Stiefel manifold; the orthogonal group for square P_0) and the write is $W = C_0^{1/2}(Q - Q_0)$. The feasible set of GROW is therefore $\{C_0^{1/2}(Q - Q_0) : QQ^\top = I\}$, a fixed linear image of the Stiefel manifold translated to pass through $W = 0$; in the Gram-space family of Keigwin et al. (2025) the factor constraint is the case $\mathcal{P} = \mathrm{Id}$ with the single Gram C_0. The linear map is not an isometry: $\|\Delta W\|_F^2 = \mathrm{tr}(\Delta Q^\top C_0\, \Delta Q)$, which equals $\|\Delta Q\|_F^2$ only for $C_0 = I$ and is proportional to it for $C_0 \propto I$. For the Gaussian factor the W-space nearest point of Equation 5 is therefore not the nearest point in the whitened coordinates, and the two coordinate systems give different tangent steepest-descent steps as well; for the row-orthogonal factor, $C_0 = (\alpha^2/d)I$, they coincide. Equation 5 is the exact nearest point: for $P = U_0 S_0 V^\top$ with $V^\top V = I$, $\|P_{\mathrm{cand}} - U_0 S_0 V^\top\|_F^2 = \|P_{\mathrm{cand}}\|_F^2 - 2\langle P_{\mathrm{cand}}^\top U_0 S_0, V\rangle + \mathrm{tr}(S_0^2)$, and the inner product is maximized over V by the polar factor of $P_{\mathrm{cand}}^\top U_0 S_0$. The projection of $W + \Delta$ onto a smooth set retains the tangential part of Δ to first order; what it does not do is choose the tangent direction by steepest linearized descent under a spectral-norm budget, as Manifold Muon does (Bernstein, 2025). Combining Adam's moments with a Stiefel retraction has precedent in Cayley ADAM (Li et al., 2020).

Numerics of the retraction. Near zero initialization $B := P_{\mathrm{cand}}^\top U_0 S_0 \approx V_0\, \mathrm{diag}(S_0^2)$, so $\mathrm{cond}(B) = \mathrm{cond}(P_0)^2$, about 1.2×10^9 for square Gaussian P_0. Batched fp32 SVD drivers truncate the tail of square inputs silently (orthogonality error up to 0.83 with no exception raised). The implementation therefore dispatches by shape: an approximate SVD for rectangular write matrices, and for square ones an fp64 ridged Gram-eigendecomposition followed by Newton–Schulz polishing (the iterate $B(B^\top B + \delta)^{-1/2}$ shares B's singular vectors with positive coefficients, so the polish converges to the exact polar factor). The Gram must be formed in fp64: in fp32 the σ^4 tail underflows to signed noise whose clamped inverse square root throws Newton–Schulz out of its basin. Every path is guarded by an explicit orthogonality check with an fp64 SVD backstop (orthogonality errors 1.7×10^{-14} for Gram-eigh with Newton–Schulz; 6.6×10^{-13} for fp64 SVD). Under data parallelism the retraction is sharded round-robin across ranks and re-gathered (weights and all-reduced

gradients are rank-identical, so the sharding is exact); three successive implementations produce bit-identical training losses at steps 20/30/40.

Cost. On live shapes (a family of 256 write matrices, one GPU) the dispatched route costs 1.15 s (rectangular) + 5.2 s (square) per family against 33–35 s for an all-fp64 or SVD-ladder route. Sharded across four ranks: 0.6 s per step net, 2.62 s per step against an estimated $\sim$2.0 for the unconstrained trainer at D=256 with activation checkpointing, a 25–30% overhead. The work is per write matrix and perfectly sharded, so the relative overhead is expected to stay flat in depth; multi-node scaling was not measured. On B300 GPUs, which deliver about one thirtieth of the B200's fp64 throughput, the Gaussian route's fp64 eigendecompositions dominated the step (7.5 s per step at D=256 against about 3 s on B200); the row-orthogonal factor's fp32 polar iteration, three to four Newton–Schulz sweeps, removes the fp64 work and returned the step to 2.5$\times$ the Gaussian route's throughput there, with constraint error at most 8.6×10^{-6} over a 4000-step run.

Tangent and normal spaces. For full-row-rank P the set $\{P : PP^\top = C_0\}$ is a smooth manifold with tangent space $T_P = \{Z : ZP^\top + PZ^\top = 0\}$ and, in the Frobenius metric, normal space $N_P = \{SP : S = S^\top\}$; their dimensions are $dm - d(d+1)/2$ and $d(d+1)/2$ (for the 768×768 attention write 50.1% of the parameters are normal, for the 768×1152 MLP write 33.4%). For the row-orthogonal factor, $PP^\top = c^2 I$ with $c = \alpha/\sqrt{d}$, the two projections are explicit: $\Pi_N(Z) = c^{-2}\,\mathrm{sym}(ZP^\top)\,P$ and $\Pi_T = \mathrm{Id} - \Pi_N$. In coordinates where $P_0 = c[I\ \ 0]$ and $Z = [A\ \ B]$, the tangent condition at initialization is $A + A^\top = 0$ with B free: what is removed at $W = 0$ is the symmetric part of the update in the frame of P_0, not every update with a shared output direction. The projection contains α only through the scale of P, so the direction selection does not weaken as α grows; in the limit of large α at bounded displacement GROW approaches AdamW restricted to the initial tangent space, not unconstrained AdamW. Along a finite displacement $WP_0^\top + P_0 W^\top = -WW^\top$ exactly, so the normal component of a finite W is second order. A tangent update can still add a positive-semidefinite first-order term to the sublayer Gram: for square $P_0 = cI$ and a skew-symmetric tangent update $Z = \Omega$, a read side $\Phi = -\Omega$ gives $Z\Phi = -\Omega^2 \succeq 0$. The selection is therefore not by itself a bound on the kernel. The nearest-point retraction retains the tangential part of a candidate to first order; the 0.68 step ratio of the α=1 runs is close to the $\sqrt{1/2} \approx 0.71$ that an isotropic candidate would retain on the square family, so the ratio alone does not show that the removed component is the harmful one; that requires the same-state counterfactual named in §5, and a control that projects onto the initial tangent space T_{P_0} without any retraction would separate the direction selection from the envelope.

A conditional stable-rank bound. From $WW^\top \preceq 4C_0$, $\|W\|_2^2 \le 4\lambda_{\max}(C_0)$ and hence $\mathrm{sr}(W) = \|W\|_F^2/\|W\|_2^2 \ge \mathrm{tr}(WW^\top)/4\lambda_{\max}(C_0)$; for the row-orthogonal factor $\lambda_{\max}(C_0) = \alpha^2/d$ and $\mathrm{sr}(W) \ge \max\{1,\, d\,\mathrm{tr}(WW^\top)/4\alpha^2\}$. At $d = 768$, $\alpha^2 = 16$: $\|W\|_2 \le 0.289$ and $\mathrm{sr}(W) \ge 12\,\mathrm{tr}(WW^\top)$, so a write matrix at the median energy of the final configuration, 10.05, has stable rank at least 121. The bound is conditional on the realized energy: $P = c\,\mathrm{diag}(R_\theta, I_{d-2})$ with R_θ a plane rotation satisfies the constraint with $\mathrm{sr}(W) = 2$, so there is no floor independent of the energy. GROW does not force a flat spectrum on W; it prevents a write matrix that spends energy from concentrating it in a few directions. The $18.5 \to 8.9$ reading of §4 is from the 8B system and is not comparable with this platform's energies.

The tangent-kernel split. Vectorizing the write matrices and writing Π_T, Π_N for the block-diagonal projections onto their tangent and normal spaces at the current state, the write-restricted kernel of §2 splits exactly, $K = J\Pi_T J^\top + J\Pi_N J^\top =: K_T + K_N$ with $0 \preceq K_T \preceq K$, and for an output direction u the fraction $\rho_N(u) = \|\Pi_N J^\top u\|^2/\|J^\top u\|^2$ is the part of u's parameter sensitivity that lies in the directions GROW removes to first order; for the top eigenvector, $u^\top K_T u = (1 - \rho_N(u))\lambda_{\max}$. Measuring ρ_N at the amplified direction and along the loss gradient across the collapse window of §3.3 is the measurement that would connect the direction selection to the gain condition of §2; it was not made here. Two cautions apply: K_T is the natural kernel of projected gradient descent, whereas GROW retracts an AdamW candidate, and the preconditioner does not commute with the projection; and the sensitivity being normal does not by itself make the update harmful.

The GROW dead state. After its death at 2.7–3.4× the boundary, the GROW ramp run rides to a hard NaN cliff at 12.6× ($\approx 4.7 \times 10^{-3}$; $g_{\mathrm{pre}} = 421$ at autopsy), where the dead AdamW state coasts quietly to 10^{-2}: the GROW dead state is not a low-gain absorbing state. The constraint error stays at 1.05×10^{-6} through crossing, bursts, death, and cliff.

Sample-quality protocol. FID and IS use 25-step Euler ODE sampling with classifier-free guidance 3.0 and timestep shift 4.0 against 49,996 validation images preprocessed by the training pipeline; the sampler was calibrated once on a healthy checkpoint over a 14-cell grid (the optimum is interior: guidance 5 degrades, 50 steps do not beat 25) and frozen for the campaign.

How the form was reached. Table 6 records the three stages of the constraint, all single-run on the small replica or the 8B system (Appendix B); only the last is used in §5.

Table 6: Three stages of the constraint (single runs; small replica and 8B system as marked).

what is held	needs	reading
$WW^\top$ itself equal to a healthy checkpoint's Gram	a healthy checkpoint; not from scratch ($W(0) = 0$ has zero Gram)	no failure within $\sim$20× the failure step when applied before it (replica and 8B); delayed 17.5× when applied after (replica)
$W = P - P_0$ with P_0 from a healthy checkpoint, $PP^\top = P_0 P_0^\top$	a checkpoint for P_0	from scratch; zero initialization exact; interior freedom improves learning over the equality on W (replica: minimum loss 1.139 against 1.159); dose-responsive in $\alpha^2 \leq 4$ (replica); an 8B from-scratch run crossed the failure window and was stopped healthy at 2760
$W = P - P_0$ with $P_0 = \alpha R/\|R\|_F$ random (§4)	a seed	no failure within 20×, loss within 0.01 of the checkpoint frame (replica); the same random frame at the raw scale collapses (8B); used in §5

F RUN LEDGER

Table 7 lists the run families behind the numbers in this paper. Every arm's configuration, seed, and logged scalars are in the run directories that accompany the submission; the index that maps each number in the text to its source is maintained alongside the manuscript.

G NORMALIZATION INTERVENTIONS: DETAILS

Table 8 gives the values plotted in Figure 3(a–c).

This appendix carries the final norm's Jacobian, the depth profiles of the two placements, the full readings of the two interventions of §3.2, and the depth-quartile table behind Figure 6.

The final norm's Jacobian. For RMSNorm with gain γ the final norm's Jacobian at token n is, with its ϵ neglected, $B_n = \frac{\sqrt{d}}{\|h_{L,n}\|} \operatorname{diag}(\gamma)\big(I - \hat{h}_n \hat{h}_n^\top\big)$, so $K_{\mathrm{norm}} = B K_{\mathrm{raw}} B^\top$: every layer's contribution to the output is rescaled by $1/\mathrm{RMS}(h_{L,n})$ per token and its radial component is removed; backward, the same factor divides the gradient into the stack by $\mathrm{RMS}(h_L)$, which is the ×23 of the surgery at one state. $\mathrm{RMS}(h_L)$ grows with depth as $L^{1.3}$–$L^{1.5}$ on the fixed-width grid and through training, and the rescaling applies to every layer equally.

Response across depth. How response is distributed across depth is a separate measurement (Figure 6): in a healthy Pre-LN checkpoint the stream grows as $l^{0.57}$ and the response of the output to a unit write falls from 5.2 at the shallow end to 1.9 at the deep end; in a Post-LN checkpoint the stream is pinned at 1 and the response is largest at the tail (0.64 against 0.36). At this early Pre-LN

Table 7: Run families and protocols.

family	system	protocol	used in
ramp	DiT $D = 128/256/512$; LM Post and Pre, $D = 128$–1024	LR from 10^{-5} to 10^{-2} (DiT) or 3×10^{-2} (LM) over 3000 steps, one seed	§3.2, Table 8
healthy control	DiT D=256 at 1.1×10^{-4} ($0.3\times$)	3 seeds, 3500 steps	§3.3, §5
constant lethal dose	DiT D=256 at 2.21×10^{-4} after 1000-step warmup; $D = 128/512$ at their own $0.6\times$	5 seeds + pilot; 2 + 2 seeds	§3.2–§3.3
LM dense collapse	LM Post $D = 256/512/1024$ near the boundary, dense event windows	2 seeds each + 1	§3.3
fixed-state forks	LM Post and Pre, $D = 128$–1024, from stored anchors	400-step arms, 2–3 seeds per dose	§3.2, Table 8
final-norm surgery	LM Pre D=512 from one healthy anchor	backward-only edits at two strengths, $\times 2$ and $\times 4$ the anchor LR	§3.2
depth profile γ_l	LM D=512: Post with p=+0.5, Pre with p=−0.5	fresh ramps	§3.2
lesion	DiT D=256	ramp; constant 2.21×10^{-4} (2 seeds)	§3.3
frozen kernel	DiT D=256; anchors at step 2000 (healthy) and 1831 (loaded)	2000 steps at 2.21×10^{-4} and 1.1×10^{-4}	§3.3
GROW, α=1	DiT D=256, Gaussian factor	ramp; constant 2.21×10^{-4}, 3 seeds, 4000 steps	§5
GROW scale sweep	DiT D=256, $\alpha^2 \in \{2, 4, 8, 16\}$, Gaussian and row-orthogonal factors	constant 2.21×10^{-4}, 1 seed, 4000 steps; ramps at α^2=16	Table 4
GROW $D = 128/512/1024$	DiT, α^2=16, row-orthogonal	ramp at each depth; constant $0.6\times$ for 4000 steps, 2 seeds (3 at D=1024); AdamW reference at $0.3\times$, 3500 steps, 2 seeds; at D=1024 AdamW ramp and $0.6\times$ (3 seeds) as well	Table 1
AdamW controls	DiT D=256 at 2.21×10^{-4}	writer weight decay 0, 2 seeds; writer rate $\times 0.68$, 3 seeds; $0.45\times$ dose, 1 seed	Table 1
design ablations	8B system and its small replica	one historical run per arm, from scratch or forked as stated	Table 3, Appendix B
FID-10k	healthy references at 3500; GROW arms at 4000	frozen sampler protocol	§5

checkpoint $\mathrm{RMS}(h_L) < 1$, so the response after the norm still exceeds the raw response; the suppression appears as $\mathrm{RMS}(h_L)$ grows. The two checkpoints are not matched in training state, and the response is a local sensitivity to a unit write, not a measure of what a layer contributes to the function; gating the writes, $u_l \mapsto a_l u_l$, or ablating tail layers would measure that and was not done. Neither placement is thereby shown to use its depth more effectively. Pre-LN also passes through states in which the residual loads onto the kernel's top directions (F_4 0.15–0.4, the signature that precedes Post-LN damage) without a change in loss.

(A) Final-norm backward surgery. The two surgical modes are *detach-denominator*, dropping the derivative through the RMS denominator (a 6.5% gradient change on a reference model), and *unscaled*, removing the whole $1/\mathrm{RMS}$ factor (a $\times 23$ gradient change). Forked from a healthy

Table 8: Post-LN and Pre-LN on the controlled LM series (width 512, $D \in \{128, \ldots, 1024\}$ blocks; ramp boundaries single-seed, fixed-state boundaries 2–3 seeds; the Pre-LN ramp NaN at D=512 did not reproduce in twins).

	Post-LN	**Pre-LN**
ramp boundary	$1.22/0.83/0.43/0.24 \times 10^{-3}$ at D=128 $\rightarrow$ 1024; $\beta \approx 0.79$	above the scan ceiling ($\approx 3 \times 10^{-2}$) at three depths; one unreproduced NaN at D=512
fixed-state boundary	η_{50} at 0.5–0.8$\times$ the ramp value	cold-jump divergence at 2–8$\times$ the anchor LR
same-state Pre/Post	4–6$\times$ / 2–3$\times$ / 2–2.5$\times$ / 2.6$\times$ at D=128 $\rightarrow$ 1024	
how it fails	burst cluster and grinding for 100–450 steps	escalating bursts to non-finite values
terminal state	plateau at 7.57–7.79 nats (unigram entropy 7.6444)	non-finite

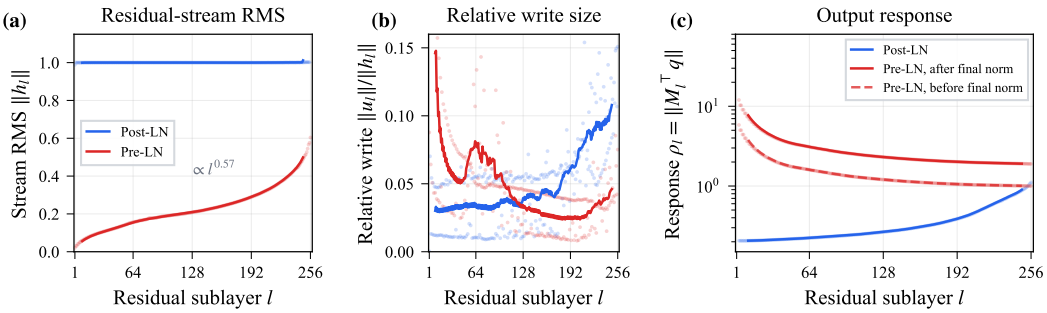


Figure 6: Stream RMS (a), relative write size (b), and the output's response to a unit write (c) across the 256 residual sublayers of two healthy D=128 checkpoints (Pre-LN at ramp step 1000, Post-LN at constant-rate step 500; training states not matched). In (c) the early Pre-LN checkpoint has $\mathrm{RMS}(h_L) < 1$, so the final norm raises rather than suppresses its response; the profiles are local sensitivities, not effective layer use.

D=512 Pre-LN anchor at $\times 2$ and $\times 4$ its LR against surgery-free controls (Table 9), the effect is dose-monotone in surgery strength. The unscaled arm turns an LR at which normal Pre-LN learns cleanly into a hard divergence: the Pre-LN cliff drops from $(4.4, 8.8] \times 10^{-4}$ to below 4.4×10^{-4}, within 1.26$\times$ of the Post-LN η_{50} upper bound at this depth. The form does not change: the surgical deaths are Pre-style cliffs (escalating bursts, no plateau). Probes through an operated backward graph acquire a definition offset (the same anchor reads its top eigenvalue $\times 60$ across surgery modes), so cross-arm comparisons use loss and label dynamics only.

Table 9: Final-norm backward surgery (D=512 Pre-LN forks from one healthy anchor; val at window close).

LR	**normal**	**detach-denominator**	**unscaled**
4.4×10^{-4} (anchor $\times 2$)	healthy (4.43)	stagnation (5.03) + amplifier episodes	**NaN @ step 976**
8.8×10^{-4} (anchor $\times 4$)	edge damage (5.17)	damage deepened (5.30) + one burst	**NaN in 14 steps**

Readings behind §3.2. Pre-LN stacks of the same width and depth reach the LM scan ceiling at three depths and produced one unreproduced NaN at D=512; under the ramp the boundaries differ by 24–120$\times$, but most of that separation is adaptation along the training path, since under same-state cold jumps the gap is 4–6$\times$ at D=128 and 2.6$\times$ at D=1024 (Table 8). The larger

boundary buys no lower loss: at each architecture's own constant dose the healthy arms reach val 3.97 (Pre) against 4.03 (Post) at D=512 and 4.18 against 4.16 at D=1024. $\mathrm{RMS}(h_L)$ grows with depth as $L^{1.3}$–$L^{1.5}$ on the fixed-width grid and through training; the surgery removes a $\times 23$ factor at the anchor state, and its divergence threshold lies below 4.4×10^{-4} against a Post-LN η_{50} upper bound of 3.5×10^{-4} at this depth, two bounds rather than two located thresholds. The Adam-state control that would separate the backward factor from the mismatch with the inherited moments, rescaling them at the fork, was not run. Pre-LN also passes through states in which the residual loads onto the kernel's top directions (F_4 0.15–0.4, the signature that precedes Post-LN damage) without a change in loss. One reading of the two failure modes is that they are the two places an instability can go: in a Pre-LN stack the stream's amplitude is free and the final norm makes the output invariant to it, so an instability can express itself as amplitude up to the divergence, whereas a Post-LN stack renormalizes after every sublayer, so the instability expresses itself in the function, which reorganizes into the degenerate solution. This is an interpretation; the DC lesion of §3.3 marks its limit, since a Post-LN system whose degenerate solution has been removed also diverges, and the native Pre-LN failures were not probed with the loss profiles and kernel measurements used on Post-LN, so whether the two placements share the same onset is untested. A second untested reading concerns residence: for gradient descent the condition of §2.2 is scale-free in the residual, since $\Delta y = -\eta K e$ and r_{safe} both shrink with it, whereas Adam's committed output step, normalized by the second-moment estimate, may not shrink in proportion to the loss-decreasing radius as the residual shrinks; the output-space step along training was not measured.

(B) Depth-decay profile. The second lever multiplies each sublayer's write by a fixed scalar $\gamma_l = (l+1)^{-p}$ (non-learnable, forward-side; fresh 3000-step ramps). With p=+0.5 at D=512, Post-LN crosses its original boundary (4.33×10^{-4}) without incident — with the LR already $1.4\times$ past it, val still falls and the top eigenvalue is actively suppressed ($5.69 \times 10^6 \to 2.92 \times 10^5$) — and the amplifier returns only near 1.68×10^{-3}: the boundary moves $3.9\times$, and the run then falls onto the same unigram plateau (7.57–7.71). With p=−0.5, i.e. $\gamma_l = \sqrt{l+1}$ amplifying the deepest layers $\times 22.6$, Pre-LN suffers a severe early pathology (continuous bursts from step 130, top eigenvalue 2×10^{12}, val near 7.4), then completely self-heals by step $\sim$2250 and rides to the scan ceiling (2.93×10^{-2}) without failing. This is a deliberately violent probe, not a calibrated inversion of Pre-LN's measured $L^{1.3}$–$L^{1.5}$ profile.

Growth of $\mathrm{RMS}(h_L)$ through training. In one Pre-LN run $\mathrm{RMS}(h_L)$ grows $0.60 \to 20 \to 6.3 \times 10^4$ across three training stages while the tail-layer response falls $1.94 \to 0.073 \to 10^{-5}$; the raw tail-layer response before the final norm stays $O(1)$ (1.02, 1.04, 3.32), and the fitted exponents of stream growth and response decay agree to a few percent (1.27/1.27, 1.51/1.54, 1.46/1.45).

Depth quartiles. Table 10 gives the quartile means behind Figure 6, whose D=128 checkpoints are from the earlier bare configuration of the series (Appendix D); the plotted lines are 17-sublayer moving means, and panel (c) omits the first two sublayers, an embedding-scale artefact. In the degenerate Pre-LN state the shallow/deep response ratio reaches $2000\times$; the L-exponents quoted above come from the fixed-width grid, and the profiles are one depth and one checkpoint per architecture.

Table 10: Depth-quartile means (shallow → deep) for the checkpoints of Figure 6.

	Pre-LN	**Post-LN**
stream RMS $\|h_l\|$	0.10 / 0.19 / 0.24 / 0.39	1.00 / 1.00 / 1.00 / 1.00
relative write $\|u_l\|/\|h_l\|$	0.085 / 0.055 / 0.028 / 0.034	0.033 / 0.035 / 0.045 / 0.089
response ρ_l (random cotangent)	5.17 / 2.62 / 2.14 / 1.94	0.36 / 0.24 / 0.31 / 0.64

H COLLAPSE EVENT TIMING AND DETECTOR CONVENTIONS

Figure 7 summarizes the timing of the detected events in the fifteen collapses of the timing set (nine DiT, six LM) on one axis, Figure 8 shows the DiT collapse of Figure 4 and an LM collapse as time series, and Table 11 sets the two realizations side by side. The detectors follow one convention across both tasks. t=0 is the first validation excursion above the prefix minimum by 0.15 that

persists; the alignment event is the first probe at which t/r exceeds twice the healthy median on two consecutive probes; the eigenvalue jump is a $\geq 4\times$ rise between heavy probes to more than ten times the healthy median; the F_k event is the first heavy probe at which $F_k \geq 0.3$ persists; the first burst is the first training step with $g_{\text{pre}} > 50$; the grinding phase runs from the first burst to the last step with $g_{\text{pre}} > 20$. For the DiT the kernel-side detectors are masked during the 1000-step warmup.

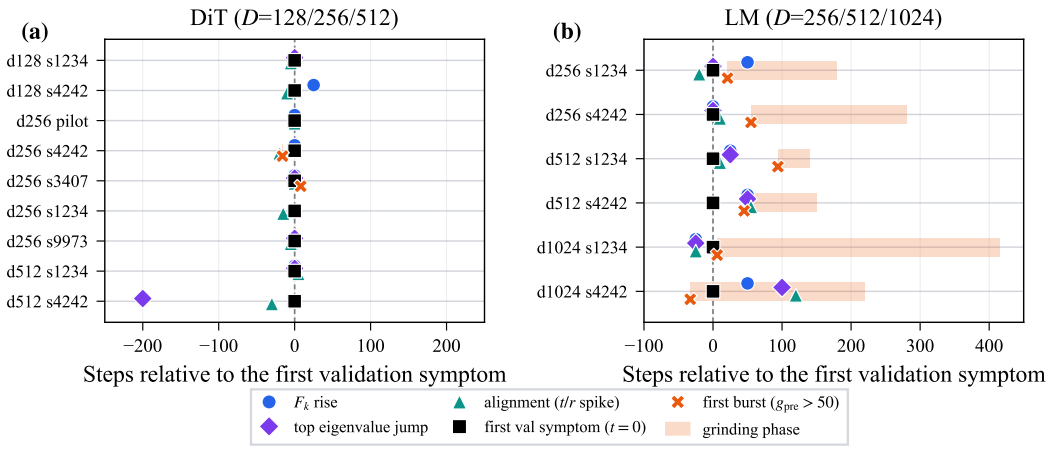


Figure 7: Event timing in the fifteen collapses of the timing set, aligned at the first validation symptom (t=0). (a) DiT, nine runs across three depths: symptom and burst, where there is one, fall within a probe interval or two of each other; kernel-side events sometimes lead by hundreds of steps without immediate damage. (b) LM, six runs: alignment near the symptom, the first burst after it, then a grinding phase (bar) of 100–450 steps. Detector conventions above.

Table 11: Two realizations of the collapse.

	DiT series	**LM series**
amplified direction	spatial DC; global, standing	token-local, single sample, shallow; recurrent transients
transition	single-step; quiet to violent; no grinding	burst clusters and grinding for 100–450 steps
terminal state	at the zero predictor's loss (1.914–1.94)	at the unigram entropy (7.57–7.79 nats around 7.6444)
gain at the end	top eigenvalue 0.13–0.18× pre-collapse (ramp); 0.7–4.5× (constant dose)	stored-direction gain 0.2–0.4×; shallow response $< 10^{-5}$

The two counterfactuals. The write-restricted kernel is a diagnostic of the state, not the operator that generates the update: in pathological LM states the write matrices carry 0.38–0.43 of the full-parameter squared gain along the residual, and in masked-update forks from a pre-onset state, withholding the commit of any single parameter family — the MLP down-projections, the attention output projections, all write matrices, or all non-write parameters — blocks the onset within a 400-step window (Appendix C). The runaway is a joint loop between the write matrices and the features they read, so both counterfactuals act on the DiT's kernel as a whole. *Deleting the amplified direction.* Does removing the direction that carries every DiT collapse restore trainability? The lesion projects every sublayer's write free of its spatial mean over image tokens. Under the ramp the plateau at the zero predictor's loss never appears, but the run is not made trainable: val bottoms at $\sim$1.15 where the unlesioned run reads 0.95, chronic bursts appear from step 910, the top eigenvalue grows $8.5 \times 10^5 \to 3.2 \times 10^{13}$ with the residual loading into the new non-DC direction (F_k 0.06 $\to$ 0.94), and after escalating bursts the run diverges to NaN at step 1627. The boundary moves from 3.72×10^{-4} to $\approx 4.0 \times 10^{-4}$, $+8\%$; at the constant $0.6\times$ dose the lesioned system has

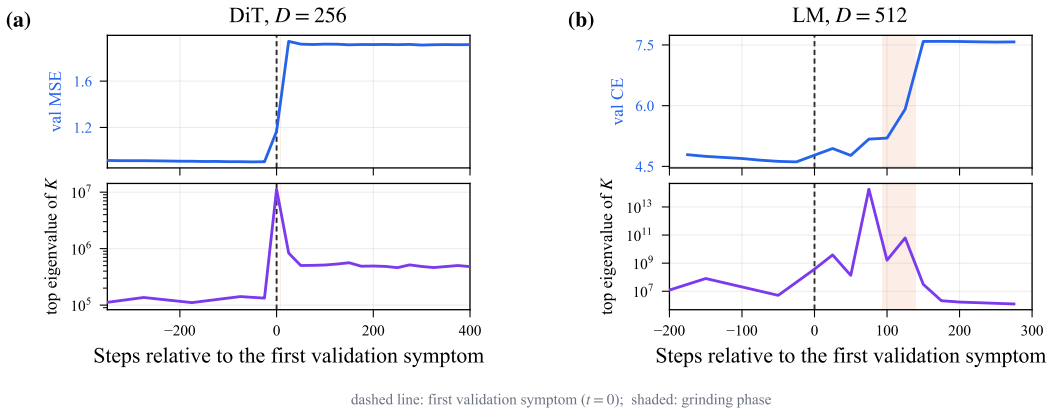


Figure 8: The DiT collapse of Figure 4 and an LM collapse as time series, aligned at the first validation symptom (t=0, dashed): validation loss above and the top eigenvalue of K below; shaded: grinding phase.

0/2 collapses, at a cost in loss (0.9606/0.9246 against 0.876–0.882). Deleting the amplified direction turns the collapse into a divergence and degrades learning; the kernel regrows its gain in another subspace. *Freezing the kernel.* Is growth of the kernel necessary for the instability? The tangent model $v(x;\delta) = f(x;\bar{\theta}) + J_{\bar{\theta}}(x)\,\delta$ trains only δ, so its kernel is frozen at the anchor $\bar{\theta}$ and route 1 is closed, while rotation of the residual (route 2), Adam's momentum and clipping, and alignment failure (route 3) all remain (exact fp32 JVP; AdamW on δ with weight decay 0, a declared deviation). From a *healthy* anchor (step 2000 of a control) at the $0.6\times$ dose, 2000 tangent steps — twice the median residence — produce neither collapse nor net learning. From a *loaded* anchor (step 1831 of the violent seed, inside its pre-collapse window) both doses destabilize within 100 steps (probe loss $1.23 \to 1.97$), and what follows is 2000 steps of relaxation oscillation (val swinging across 0.908–3.192) that reaches neither the plateau nor a NaN, where normal training from the same state completes its collapse in 52 steps. Instability therefore does not require kernel growth, and the anchor state sets the violence of the response; since the loaded anchor differs in its base prediction as well as in its kernel, the comparison isolates the anchor state, not the kernel alone. That no plateau formed does not show that the gain must fall for the collapse to complete: the linearized model also changes the set of reachable functions, and Adam's state is reinitialized at the fork. One seed per arm, one dose pair, a 2000-step window.

Secondary readings. DiT ramp boundaries: 5.63×10^{-4} / 3.72×10^{-4} / 2.82×10^{-4} at $D = 128/256/512$ (AdamW); 1.0×10^{-3}–1.26×10^{-3} at D=256 for GROW at α=1, and 9.19×10^{-4} / 5.32×10^{-4}–5.63×10^{-4} / $\approx 3.55 \times 10^{-4}$ / 2.74×10^{-4} at $D = 128/256/512/1024$ for GROW at α^2=16 with the row-orthogonal factor, and 1.83×10^{-4} for AdamW at D=1024 ($\beta \approx 0.50$ holds through D=1024: $2.82 \times 10^{-4} \to 1.83 \times 10^{-4}$ is a factor 0.65 for a doubling; $2^{-0.5} = 0.71$). At the D=256 constant dose the five collapses occur at post-warmup residences of 334/875/883/1190/1255 steps, one seed survives past 2500, and the median residence at each depth's own $0.6\times$ is $\sim$1500/1035/365 for $D = 128/256/512$. Through the healthy segment of the D=256 ramp val falls $1.79 \to 0.955$; at the transition (step 1570, $g_{\text{pre}} = 22.6$) the probe at 1550 is healthy (t/r 1.90) and the probe at 1575 is on the plateau (loss 2.26). The constant-dose collapses span quiet (g_{pre} peak $\leq$2.9), intermediate (7–10.8), and violent ($g_{\text{pre}} = 241$; one seed 1948) events and land at val 1.914–1.94; in the violent seed the training loss jumps $0.94 \to 1.94$ within ten steps of the transition and the output norm falls to 0.05 by step 1950. Terminal top eigenvalues relative to the pre-collapse value: 0.13/0.16/0.18 for the three ramp collapses ($D = 128/256/512$) and 0.69–4.47 for the nine constant-dose collapses, from event peaks of 2.4–25,000$\times$ (§3.3). The dense LM collapses are two seeds each at $D = 256/512/1024$ plus a third run. The alarm pool of 553 heavy-probe snapshots from 75 runs is dominated by cold-jump forks in which damage arrives before the first post-jump heavy probe, so it measures immediate alarm rather than lead time; its damage label is the loss-based one of Appendix D, and against it the plain pre-clip gradient norm is the best single predictor (AUC 0.72–0.77). The panel's residual overlap onto the top eigenvector is 10^{-6}–

10^{-9}. In the lesioned ramp the NaN at step 1627 falls at LR 4.0–4.2×10^{-4}, and at the constant dose the lesioned arms show chronic bursts and transient non-DC amplifiers. In the healthy-anchor frozen-kernel arm val never falls below its starting 0.896 through $\sim$8 episodes in which the residual loads onto high-gain directions; in the loaded-anchor arms the loss spikes to 4–60 and relaxes under clipping. Ramp transitions at the other depths: D=128 at $\sim$1750 (LR 5.63×10^{-4}; the probe lands mid-transition and reads t/r 60–66; the nearest training step has $g_{\text{pre}} = 23.5$, a single step, no cluster); D=512 at $\sim$1445–1450 (LR $\approx 2.8 \times 10^{-4}$; an isolated $g_{\text{pre}} = 2.1$). At D=256 the dead run rides the plateau for 1430 further steps with no second event. Across width, the D=256 DiT at width 768 has its boundary at $\approx 3.7 \times 10^{-4}$ while the 8B system at the same depth and width 2048 failed at $\approx 9 \times 10^{-5}$. The constant dose 2.21×10^{-4} of the main DiT experiment is the 8B production base rate. At every LM depth the boundary-edge fork closes its window with the top eigenvalue already excursed (e.g. 2.8×10^{15} with concentration 1.0 at D=128, $0.5\times$) while val is intact. In the dense LM collapses t/r leaves its healthy band as, e.g., $1.5 \to 5.8 \to 21.5$ across three probes five steps apart; three super-critical forks of one state land at 7.6159/7.6162/7.6164 nats. At the D=256 ramp transition the probe at 1575 reads t/r 15–16, and the dead run's terminal g_{pre} is 0.45–0.57; in the record violent death t/r runs $2.3 \to 19.5$ within $\sim$15 steps. The dead DiT's per-timestep-bin losses flatten from the healthy $[1.24, 1.10, 0.91, 0.89]$ to $[1.89, 1.91, 1.92, 1.92]$. Along the released-model training trajectories the residual is born inside the high-gain cone (t/r 5.3–5.7 at initialization) and decouples to $t/r \approx 1$ between 1k and 16k steps. The surviving $0.6\times$ AdamW DiT seed ends at val 0.876; the LM at $0.5\times$ its ramp boundary collapses after $\sim$1500 steps at D=128 and $\sim$1900 at D=256. The LM onset closure: $0.031 \times 5.78 \times 10^4 \approx 1793$ against 1809 measured. In the LM series, QK-norm on/off leaves boundary and event order unchanged and only damps in-burst loss peaks (5.3–7.8 with, 9.0–17.8 without); kernel quantities are probed in fp32 every 25 (light) and 100 (heavy) steps, 5/25 inside event windows. The lead time of the kernel signal to the first validation symptom is 15–20 steps. In the lesioned ramp the top eigenvector's DC share stays ≤ 0.015 and the bursts escalate $103 \to 199 \to 684 \to 3564$ before the NaN; in the healthy-anchor frozen-kernel arm val wanders 0.900–0.919 with t/r cycling $1.4 \leftrightarrow 2.4$, and in the loaded-anchor arm loss relaxes to $\sim$1.0 between spikes. The GROW D=256 runs descend monotonically with the timestep-bin structure of val intact. In the dense LM collapses the top eigenvalue peaks at 2.5×10^{12} / 1.8×10^{14} / 2.8×10^{14} for $D = 256/512/1024$; the dead LM's shallowest quartile responds below 10^{-5}. In the lesioned ramp val bottoms at $\sim$1.15 at step 800 and rises $1.17 \to 1.29 \to 1.75$. In the loaded-anchor frozen-kernel arms t/r runs $2.8 \to 5.2$ within 100 steps and g_{pre} peaks at 779 (lethal dose) and 355 (half dose). The lesion's residual DC content in the writes measures 10^{-8}; the lesioned ramp shows 34 bursts with $g_{\text{pre}} > 5$ from step 910. In the frozen-kernel arms the gradient agrees with central differences to 2.1×10^{-6}, and the loaded anchor shows 70 spikes above 5 on a 10-step grid at the lethal dose and 35 at half dose. The fixed-state LR fork of one LM state on the same data: no event in 400 steps at $\leq 4.5 \times 10^{-4}$; onset at $\sim$280 steps at 6×10^{-4}, $\sim$80 at 8×10^{-4}, 20–50 at $\geq 10^{-3}$, pinning that state's critical rate into $(4.5, 6] \times 10^{-4}$. The simple-predictor panel (same batch, same noise): zero predictor 1.9357, global mean 1.9328, per-channel mean 1.9346, oracle per-sample spatial mean 1.8597; the spatial-DC direction carries 0.039 of the target's energy. The dead LM's tail quartile responds at 1/28 of healthy, and its terminal residual direction already reads gains of 123/96 on early healthy checkpoints, higher than on the plateau's own geometry. The healthy DiT controls at $0.3\times$ reach val 0.8795–0.8817 at step 3500.