Optimizer-Induced Low-Dimensional Drift and Transverse Dynamics in Transformer Training

Yongzhong Xu[*]

Code: `https://github.com/skydancerosel/mini_gpt`

Abstract

We analyze the cumulative parameter trajectory of transformer training under AdamW and identify a stable, low-dimensional drift direction that captures the majority of long-horizon displacement from initialization. This "backbone" accounts for 60–80% of cumulative drift across blocks and remains temporally stable throughout training. Strikingly, instantaneous per-batch gradients are nearly isotropic relative to this direction, and it is orthogonal to top Fisher curvature modes. Instead, the backbone emerges from optimizer integration: momentum amplifies weak but temporally coherent gradient bias, while adaptive normalization suppresses incoherent transverse fluctuations.

Replacing AdamW with SGD-family optimizers eliminates this structure. Even at matched validation loss, SGD trajectories remain nearly colinear and fail to develop the multi-dimensional slow–fast geometry observed under AdamW, demonstrating that the backbone is optimizer-induced rather than loss-landscape-determined.

Oscillatory regime switching between competing objectives occurs primarily in directions transverse to the backbone. Reheating experiments show that transverse modes can be transiently re-excited from late-training checkpoints without erasing accumulated backbone drift, consistent with a slow–fast decomposition of training dynamics.

These results shift attention from instantaneous gradient geometry to cumulative trajectory structure and provide a concrete empirical characterization of optimizer-induced implicit bias in transformer training.

1 Introduction

Training dynamics in deep neural networks are typically analyzed through the geometry of the loss landscape: curvature, sharpness, and stochastic gradient noise are taken to determine how optimization proceeds. While this perspective captures important local properties of learning, it emphasizes instantaneous gradient structure rather than the accumulated trajectory of parameters over long training horizons.

In high-dimensional models, these two viewpoints need not coincide. Per-step gradients may be large and highly variable, yet their cumulative displacement can concentrate in a small number of coherent directions. Understanding this cumulative geometry is essential for characterizing optimizer-induced implicit bias and long-horizon training behavior.

In this work, we study the global parameter trajectory of transformer training under AdamW [Loshchilov and Hutter, 2019]. Rather than analyzing local curvature or single-step gradients, we examine cumulative displacement from initialization across checkpoints. We find that training admits a stable low-dimensional drift direction—which we term the *backbone*—that captures the majority of cumulative parameter movement. Across blocks and seeds, the first

[*] `abbyxu@gmail.com`

principal component of uncentered trajectory PCA explains 60–80% of total drift, and this direction remains nearly fixed throughout training.

Crucially, the backbone is not aligned with instantaneous gradient directions and is nearly orthogonal to leading Fisher curvature modes [Martens, 2020]. Per-batch gradients are close to isotropic relative to the backbone. However, the optimizer-integrated update—after momentum accumulation and adaptive per-parameter normalization [Kingma and Ba, 2015]—exhibits strong alignment with it. This demonstrates that the backbone is not a property of the loss landscape alone, but an emergent property of optimizer dynamics.

To test this interpretation, we replace AdamW with SGD-family optimizers while holding model, data, and schedule fixed. Under SGD with or without momentum, trajectories remain nearly colinear and fail to develop the multi-dimensional structure observed under AdamW—even at matched validation loss. This establishes that the backbone is optimizer-induced rather than a generic feature of the objective.

We further examine the dynamical consequences of this structure. Oscillatory regime switching between competing objectives occurs primarily in directions transverse to the backbone. Reheating experiments show that these transverse modes can be transiently re-excited from late-training checkpoints without substantially altering accumulated backbone drift. This behavior is consistent with a slow–fast decomposition: a low-dimensional, optimizer-shaped drift manifold governs long-horizon evolution, while high-dimensional transverse dynamics mediate switching.

Together, these findings shift attention from instantaneous gradient geometry to cumulative trajectory structure. They provide a concrete empirical characterization of optimizer-induced implicit bias in transformer training and suggest that adaptive optimization reshapes not only convergence rates but the geometry of learning itself.

Relation to prior work. The separation of dynamics into slow and fast components has classical roots in dynamical systems theory, particularly in slow manifold and time-scale separation results such as Fenichel-type theorems [Saxe et al., 2014]. In optimization and deep learning, related ideas appear in analyses of momentum methods and adaptive optimizers, where Adam-type algorithms are understood as inducing effective geometry changes through preconditioning and sign-consistent updates [Kingma and Ba, 2015, Loshchilov and Hutter, 2019, Cohen et al., 2021]. Recent work has also emphasized implicit bias and trajectory-level properties of high-dimensional training dynamics [Lewkowycz et al., 2020, Power et al., 2022, Frankle et al., 2020].

Our contribution differs in emphasis and object of study. Rather than analyzing local curvature, stationary points, or instantaneous update rules, we examine the cumulative geometry of training trajectories and identify a stable drift direction that dominates long-horizon parameter displacement. We show that this backbone direction is not aligned with per-batch gradients or with top curvature modes, but instead emerges from optimizer-integrated temporal coherence. This shifts attention from static loss-landscape structure to trajectory-level geometry, providing a concrete empirical characterization of optimizer-induced slow-manifold behavior in transformer training.

2 Experimental Setup

2.1 Model and Data

We train a decoder-only Transformer [Vaswani et al., 2017] in the GPT-2 family [Radford et al., 2019]: 8 layers, $d_{\text{model}} = 512$, 16 attention heads, $d_{\text{ff}} = 2048$, totalling 51M parameters. The training corpus is TinyStories [Eldan and Li, 2023]. With probability $p_{\text{probe}} = 0.10$, a training sequence is replaced by a probe sequence containing a codeword–value pair; the model must predict the value token given the codeword at out-of-distribution gap distances.

2.2 Training Configuration

Table 1: Training hyperparameters.

Parameter	Value
Optimizer	AdamW (β_1=0.9, β_2=0.95)
Learning rate	10^{-3}, cosine decay, 1500-step warmup
Weight decay	0.5
Probe loss weight λ	2.0 (steps 1–3999), 4.0 (steps 4000–10000)
Effective batch size	128 (64 × 2 gradient accumulation)
Total steps	10,000
Checkpoint interval	200 steps (51 checkpoints)
Seeds	42, 271

The composite loss at each training step is

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \mathcal{L}_{\mathrm{LM}}(\boldsymbol{\theta}) \;+\; \lambda\,\mathcal{L}_{\mathrm{probe}}(\boldsymbol{\theta}), \tag{1}$$

where $\mathcal{L}_{\mathrm{LM}}$ is the standard next-token prediction cross-entropy and $\mathcal{L}_{\mathrm{probe}}$ is the cross-entropy on the codeword retrieval task. The weight λ is doubled at step 4000 to intensify probe competition.

2.3 Oscillation Phenomenology (Brief)

Over 10,000 steps, the out-of-distribution probe accuracy p_{ood} oscillates between 0.40 and 0.78 (seed 42) or 0.20 and 0.67 (seed 271), while the LM validation loss decreases monotonically from $\sim$10 to $\sim$1.2. The oscillations damp after step $\sim$7000, and the model settles into an LM-dominant regime ($p_{\mathrm{ood}} < 0.20$). We set aside the full oscillation phenomenology and focus on the geometric structure of the underlying training trajectory.

3 The Backbone

3.1 Trajectory PCA: One Direction Dominates

We analyze the cumulative parameter drift using *uncentered* principal component analysis. Let $\boldsymbol{\theta}(t) \in \mathbb{R}^D$ denote the vectorized parameters of a single transformer block at checkpoint t, with block dimensionality $D \approx 3.1 \times 10^6$.

Definition 1 (Drift matrix)**.** *The drift matrix* $\mathbf{X} \in \mathbb{R}^{T\times D}$ *has rows*

$$\mathbf{x}(t) \;=\; \boldsymbol{\theta}(t) - \boldsymbol{\theta}(0), \qquad t = 1, \dots, T, \tag{2}$$

where $T = 51$ *is the number of checkpoints.*

We deliberately omit mean centering before computing the SVD. Standard (centered) PCA would subtract the mean drift $\bar{\mathbf{x}} = T^{-1}\sum_t \mathbf{x}(t)$, which removes the monotonic component and conflates it with the first principal component. Since all drifts are relative to initialization, there is no reason to assume zero-mean variation.

The singular value decomposition

$$\mathbf{X} \;=\; \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{V}^\top, \tag{3}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_{\min(T,D)})$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, yields the principal directions as the columns of $\mathbf{V}$. The fraction of total squared drift captured by the k-th component is

$$\rho_k \;=\; \frac{\sigma_k^2}{\sum_{j=1}^{\min(T,D)} \sigma_j^2}. \tag{4}$$

Table 2: Variance explained by PC1 (ρ_1, %) per transformer block.

Seed	**Blk 0**	**Blk 1**	**Blk 2**	**Blk 3**	**Blk 4**	**Blk 5**	**Blk 6**	**Blk 7**
42	80.5	81.2	80.7	80.2	79.7	79.0	78.8	77.9
271	78.6	80.6	80.4	80.4	80.1	79.1	79.3	78.3

Table 2 shows that PC1 captures **78–81%** of the total squared drift in every block, in both seeds. The training trajectory is overwhelmingly one-dimensional. We call this direction the **backbone**, denoted $\mathbf{v}_\mathrm{b}$.

Remark 1 (Why uncentered PCA?)**.** *Standard PCA centers the data by subtracting the column mean before SVD. For trajectory analysis, centering removes the dominant monotonic drift and distributes it across all components. Uncentered PCA preserves the absolute direction of displacement from initialization. In our setting, this correctly identifies the persistent LM-driven drift as the leading component. The mathematical difference is that centered PCA diagonalizes the covariance matrix* $\frac{1}{T}\mathbf{X}^\top\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top$*, while uncentered PCA diagonalizes* $\frac{1}{T}\mathbf{X}^\top\mathbf{X}$ *directly.*

3.2 Temporal Stability of the Backbone

A rolling window analysis (width $W = 10$ checkpoints, ≈2000 steps) tracks the local PC1 direction $\mathbf{v}_\mathrm{b}^{(w)}$ at each window position w and measures its alignment with the global $\mathbf{v}_\mathrm{b}$. Define

$$c(w) \;=\; |\langle \mathbf{v}_\mathrm{b}^{(w)},\, \mathbf{v}_\mathrm{b}\rangle| \;=\; |\cos\angle(\mathbf{v}_\mathrm{b}^{(w)},\, \mathbf{v}_\mathrm{b})|. \tag{5}$$

The mean of $c(w)$ across all windows is 0.997–0.998 for both seeds, indicating that the backbone direction is essentially fixed from early training onward. It is not an artifact of averaging over distinct dynamical phases.

3.3 Backbone–Residual Decomposition

Definition 2 (Backbone decomposition)**.** *The parameter vector at step t is decomposed as*

$$\boldsymbol{\theta}(t) \;=\; \boldsymbol{\theta}(0) \;+\; a(t)\,\mathbf{v}_\mathrm{b} \;+\; \mathbf{r}(t), \tag{6}$$

where the backbone coordinate *is the signed projection*

$$a(t) \;=\; \langle \boldsymbol{\theta}(t) - \boldsymbol{\theta}(0),\, \mathbf{v}_\mathrm{b}\rangle, \tag{7}$$

and the residual $\mathbf{r}(t) \perp \mathbf{v}_\mathrm{b}$ *captures all non-backbone displacement:*

$$\mathbf{r}(t) \;=\; \big[\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\big] \;-\; a(t)\,\mathbf{v}_\mathrm{b}. \tag{8}$$

The backbone coordinate $a(t)$ grows monotonically—it tracks the steady LM-driven drift. The residual $\mathbf{r}(t)$ contains the oscillatory dynamics: its norm $\|\mathbf{r}(t)\|$ fluctuates in phase with p_ood oscillations. At the final checkpoint, the backbone fraction

$$f_\mathrm{b}(t) \;=\; \frac{a(t)^2}{\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\|^2} \;=\; \frac{a(t)^2}{a(t)^2 + \|\mathbf{r}(t)\|^2} \tag{9}$$

is 68–72%. This establishes a clean separation: the backbone carries the monotonic LM drift, while the residual carries the switching dynamics.

4 Mechanism: How the Backbone Emerges

This section presents the central result. We show that the backbone is an emergent property of optimizer integration, not of instantaneous gradient alignment.

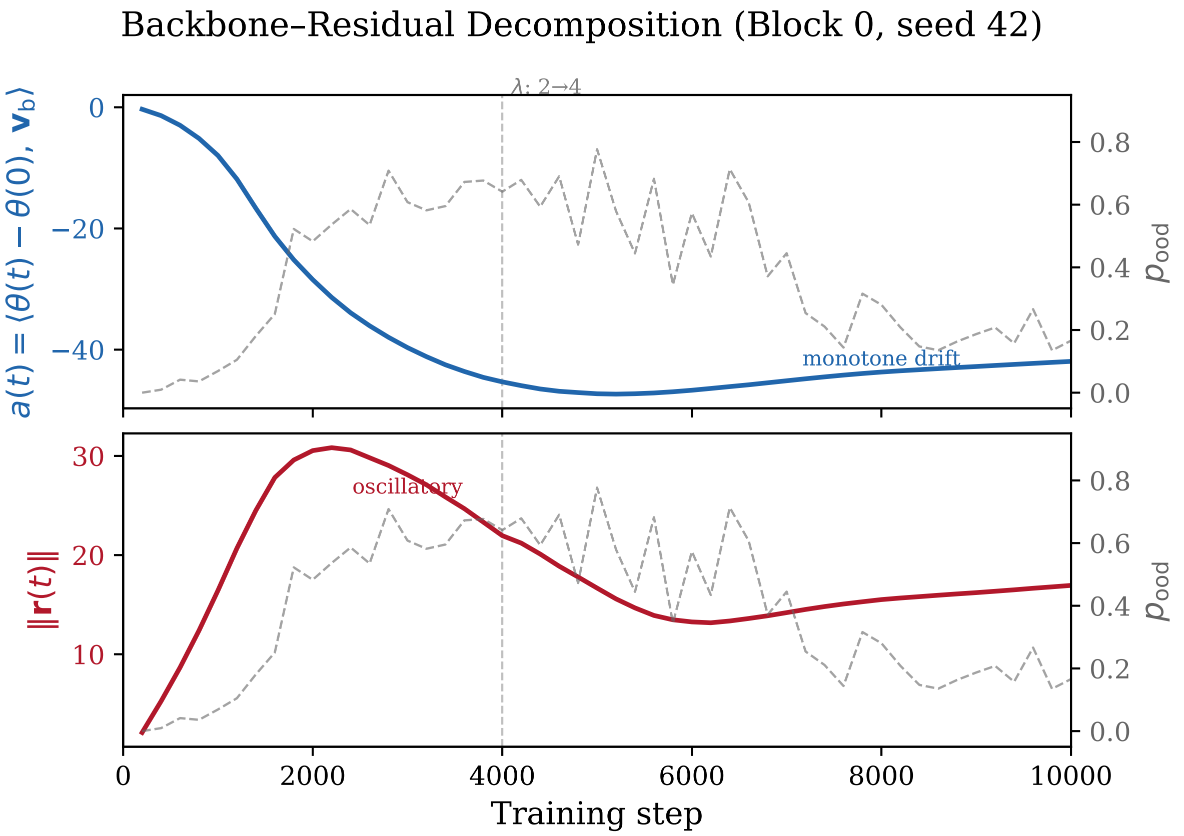

Figure 1: **Backbone–residual decomposition (seed 42, Block 0).** *Left:* the backbone coordinate $a(t)$ grows monotonically while the residual norm $\|\mathbf{r}(t)\|$ oscillates and then decays. *Right:* the out-of-distribution probe accuracy p_ood (grey) fluctuates in phase with the residual—but the backbone is impervious. The vertical dashed line marks the λ-transition at step 4000.

4.1 The Puzzle: Per-Batch Gradients Do Not Align

An intuitive expectation is that the backbone arises because gradients consistently point along it. This is false.

At each checkpoint, we compute the per-batch combined-loss gradient $\mathbf{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t))$ and measure its cosine similarity with $\mathbf{v}_\mathrm{b}$. Across 16 mini-batches per checkpoint:

$$|\cos \angle(\mathbf{g}_t, \mathbf{v}_\mathrm{b})| \;\approx\; 0.008\text{–}0.012. \tag{10}$$

To understand why, recall that for two independent random unit vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^D$, the expected absolute cosine similarity is

$$\mathbb{E}\big[|\langle \mathbf{a}, \mathbf{b} \rangle|\big] \;=\; \sqrt{\frac{2}{\pi D}}. \tag{11}$$

For $D = 3.1 \times 10^6$, this yields $\approx 4.5 \times 10^{-4}$. The observed $|\cos| \approx 0.01$ is modestly above this (due to gradient structure), but still at the noise floor.

Proposition 1 (Noise-floor scaling)**.** *For a gradient vector $\mathbf{g} \in \mathbb{R}^D$ with effective dimensionality $d_\mathrm{eff} \ll D$ (i.e., the gradient energy concentrates on d_eff directions), the expected projection onto a fixed unit vector $\mathbf{v}_\mathrm{b}$ scales as*

$$\mathbb{E}\big[|\langle \hat{\mathbf{g}}, \mathbf{v}_\mathrm{b} \rangle|\big] \;\sim\; \frac{1}{\sqrt{d_\mathrm{eff}}}, \tag{12}$$

where $\hat{\mathbf{g}} = \mathbf{g}/\|\mathbf{g}\|$. Since gradients in deep networks typically have $d_\mathrm{eff} \gg 1$, the alignment with any fixed direction is small.

No individual gradient step "points along" the backbone.

4.2 Optimizer Integration Resolves It

The backbone is not produced by strong instantaneous gradient alignment. Per-batch gradients are nearly isotropic with respect to the backbone direction, and their cosine alignment is only marginally above random. Instead, the backbone emerges from the interaction between optimizer dynamics and temporal coherence.

Under AdamW [Loshchilov and Hutter, 2019], the applied parameter update is not the raw gradient but a momentum-accumulated, variance-normalized step with weight decay. At step t, the first and second moment estimates are

$$\mathbf{m}_t \;=\; \beta_1 \, \mathbf{m}_{t-1} \;+\; (1-\beta_1)\, \mathbf{g}_t, \tag{13}$$

$$\mathbf{v}_t \;=\; \beta_2 \, \mathbf{v}_{t-1} \;+\; (1-\beta_2)\, \mathbf{g}_t^2, \tag{14}$$

where $\mathbf{g}_t^2$ denotes element-wise squaring, $\beta_1 = 0.9$, and $\beta_2 = 0.95$. With bias correction ($\hat{\mathbf{m}}_t$, $\hat{\mathbf{v}}_t$), the effective parameter update (including decoupled weight decay with coefficient μ) is

$$\boxed{\boldsymbol{\theta}_{t+1} \;=\; \boldsymbol{\theta}_t \;-\; \eta_t \left(\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \;+\; \mu\, \boldsymbol{\theta}_t \right),} \tag{15}$$

where η_t is the learning rate at step t and $\epsilon = 10^{-8}$. The net update direction is therefore

$$\mathbf{u}_t \;=\; -\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \;-\; \mu\, \boldsymbol{\theta}_t. \tag{16}$$

Three properties of this update conspire to create the backbone:

(i) Momentum integration of signed bias. The exponential moving average in eq. (13) acts as a low-pass filter with time constant $\tau_m = -1/\ln \beta_1 \approx 9.5$ steps. Even when the instantaneous projection $\langle \mathbf{g}_t, \mathbf{v}_\mathrm{b} \rangle$ is small, if its *sign* exhibits a weak but persistent bias across steps, the momentum term $\mathbf{m}_t$ integrates this bias. Concretely, suppose the i-th coordinate of the gradient has a mean μ_i and variance σ_i^2 across time. Then

$$\mathbb{E}[(\mathbf{m}_t)_i] \;\approx\; \mu_i, \qquad \mathrm{Var}[(\mathbf{m}_t)_i] \;\approx\; \frac{1-\beta_1}{1+\beta_1}\, \sigma_i^2, \tag{17}$$

so the signal-to-noise ratio of the momentum is

$$\mathrm{SNR}_i \;=\; \frac{|\mu_i|}{\sqrt{\mathrm{Var}[(\mathbf{m}_t)_i]}} \;\approx\; \frac{|\mu_i|}{\sigma_i} \cdot \sqrt{\frac{1+\beta_1}{1-\beta_1}} \;=\; \frac{|\mu_i|}{\sigma_i} \cdot \sqrt{19}, \tag{18}$$

amplifying the per-coordinate SNR by a factor of $\sqrt{19} \approx 4.4$. Coordinates aligned with the backbone (where μ_i is nonzero due to the persistent LM gradient bias) benefit from this amplification.

(ii) Adaptive normalization suppresses transverse variance. The denominator $\sqrt{\hat{\mathbf{v}}_t} + \epsilon$ in eq. (15) normalizes each coordinate by its root-mean-square gradient magnitude. Coordinates with large but incoherent gradients (high σ_i, low μ_i—typical of transverse directions) are divided by a large value, suppressing their contribution to the update. Coordinates with smaller but coherent gradients (moderate σ_i, nonzero μ_i—typical of backbone-aligned directions) are normalized by a smaller value, preserving their contribution. Formally, the effective update along coordinate i scales as

$$\frac{(\hat{\mathbf{m}}_t)_i}{\sqrt{(\hat{\mathbf{v}}_t)_i} + \epsilon} \;\approx\; \frac{\mu_i}{\sqrt{\mu_i^2 + \sigma_i^2}} \;=\; \frac{\mathrm{SNR}_i^0}{\sqrt{1 + (\mathrm{SNR}_i^0)^2}}, \tag{19}$$

where $\mathrm{SNR}_i^0 = \mu_i/\sigma_i$. This is a *sign-preserving squashing function*: it amplifies directions with high SNR (backbone) relative to those with low SNR (transverse).

Table 3: Update-direction alignment $|C(t)|$ by training phase (seed 42). The early-training alignment is 20–30× above the per-batch gradient noise floor of ∼0.01.

Phase	**Block 0**	**Blocks 1–7 (median)**	**Interpretation**
Early ($t < 2000$)	**0.27** (peak 0.34)	0.20–0.21	Strong alignment
Mid ($2000 \leq t \leq 6000$)	0.054	0.054–0.076	Weakened at λ-transition
Late ($t > 6000$)	0.110	0.061–0.143	Moderate, sign reversed

(iii) Cumulative coherence vs. instantaneous isotropy. In $D = 3.1 \times 10^6$ dimensions, each gradient step has enormous freedom. Large transverse components can dominate the norm of each update while still cancelling over time if their directions fluctuate. A much smaller but temporally coherent component can therefore dominate *cumulative displacement.* If the per-step transverse displacement has variance $\sigma_\perp^2$ per dimension and the backbone displacement has magnitude $\delta_\parallel$ per step, then after T steps the expected squared displacements are

$$\|\Delta_\parallel\|^2 \;=\; (T\,\delta_\parallel)^2 \;=\; T^2\,\delta_\parallel^2, \qquad \mathbb{E}[\|\Delta_\perp\|^2] \;=\; T\,(D-1)\,\sigma_\perp^2, \tag{20}$$

where the backbone displacement grows *linearly* ($\propto T$) while the transverse displacement grows as a random walk ($\propto \sqrt{T}$). The backbone fraction therefore approaches 1 as $T \to \infty$:

$$f_\mathrm{b} \;\approx\; \frac{T^2\,\delta_\parallel^2}{T^2\,\delta_\parallel^2 + T\,(D-1)\,\sigma_\perp^2} \;\xrightarrow{T\to\infty}\; 1. \tag{21}$$

This explains why $\rho_1 \approx 0.80$ with only $T = 51$ checkpoints: even a weak backbone bias, amplified by momentum and adaptive normalization, is sufficient to dominate cumulative drift.

4.3 Evidence: Update-Direction Alignment

We compute the 200-step update direction

$$\mathbf{u}(t) \;=\; \boldsymbol{\theta}(t) - \boldsymbol{\theta}(t-200), \tag{22}$$

which reflects the net effect of all optimizer operations (momentum, adaptive learning rates, weight decay, gradient clipping) accumulated over 200 steps. We measure the cosine similarity with the backbone:

$$C(t) \;=\; \cos\angle\big(\mathbf{u}(t),\,\mathbf{v}_\mathrm{b}\big) \;=\; \frac{\langle \mathbf{u}(t),\,\mathbf{v}_\mathrm{b}\rangle}{\|\mathbf{u}(t)\|}. \tag{23}$$

The early-training $|C(t)|$ of 0.20–0.34 (Table 3) is **20–30× above** the per-batch gradient noise floor of ∼0.01. This confirms that optimizer integration transforms isotropic per-batch gradients into structured, backbone-aligned updates.

Sign structure. The signed cosine $C(t)$ is persistently negative throughout early training (the optimizer drifts in the $-\mathbf{v}_\mathrm{b}$ direction), then flips to positive around step 5000–5200. Defining the sign-flip step t^* as the first checkpoint where $C(t)$ crosses zero:

Table 4: Sign flip in update–backbone alignment.

Seed	t^*	**Early** $C(t)$ **(Blk 0)**	**Late** $C(t)$ **(Blk 0)**
42	∼5200	−0.27	+0.11
271	∼5000	−0.29	+0.07

The sign reversal coincides with the λ-transition (step 4000, $\lambda : 2 \to 4$) and the onset of oscillation damping. The reversal is present in both seeds (Table 4).

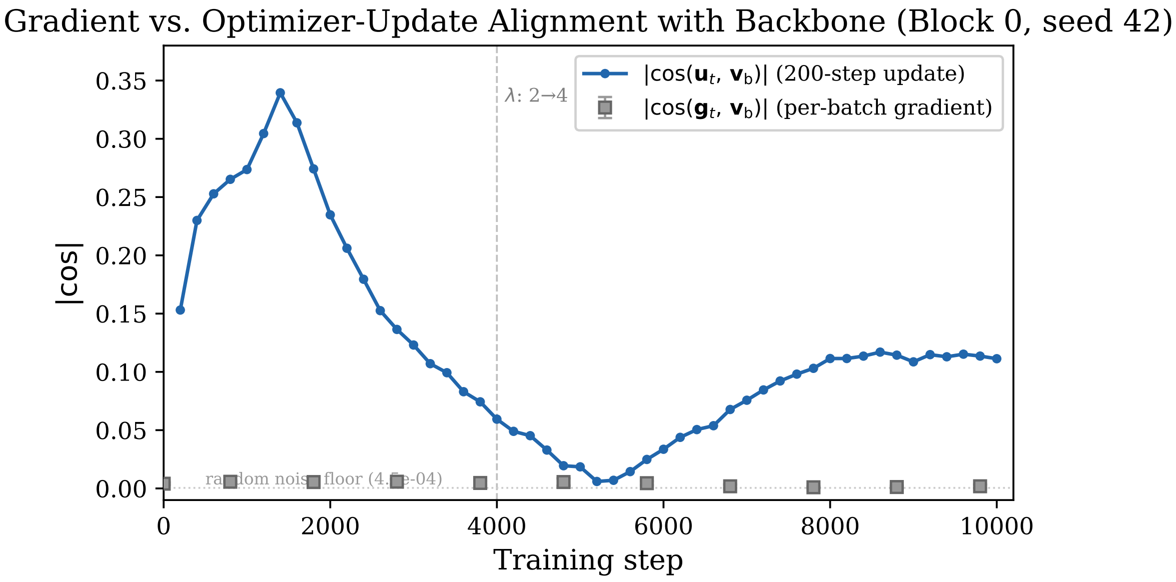

Figure 2: **The linchpin: update vs. gradient alignment with the backbone (seed 42, Block 0).** Per-batch gradient alignment $|\cos(\mathbf{g}_t, \mathbf{v}_\mathrm{b})|$ (blue circles with error bars) hovers at the noise floor (∼0.005), while the 200-step update alignment $|\cos(\mathbf{u}_t, \mathbf{v}_\mathrm{b})|$ (solid red) reaches 0.20–0.34 in early training. The ∼40× gap demonstrates that the backbone emerges from optimizer integration, not from instantaneous gradient structure. The signed update alignment (dashed red) reveals a persistent negative bias that flips sign near step 5000.

Table 5: Mean signed gradient projection $b_\ell(t)$ per block (seed 42, $\lambda = 2.0$ phase). Block 0 dominates by 3–10×.

Checkpoint	**Blk 0**	**Blk 1**	**Blks 2–5**	**Blk 6**	**Blk 7**
Step 200 (init)	**0.014**	0.003	0.001	0.000	0.002
Step 1800 (peak)	**0.019**	0.001	0.001	0.003	0.004
Step 2000 (trough)	**0.029**	0.001	0.001	0.005	0.005

4.4 Block Localization: The First Transformer Block Drives the Backbone

Which blocks contribute the gradient bias that creates the backbone? At each checkpoint, we compute the signed projection of the combined-loss gradient onto the backbone, per block ℓ:

$$b_\ell(t) \;=\; \langle \mathbf{g}_\ell(t),\, \mathbf{v}_\mathrm{b}^{(\ell)} \rangle, \tag{24}$$

where $\mathbf{g}_\ell(t)$ is the gradient restricted to block ℓ and $\mathbf{v}_\mathrm{b}^{(\ell)}$ is the block-ℓ backbone.

Block 0 has 3–10× larger signed projection than any other block (Table 5). Blocks 6–7 contribute a secondary positive bias. Blocks 1–5 are uniformly near zero. By late training (step 9600+), all projections collapse to < 0.001.

The cumulative signed bias, obtained by summing $b_\ell(t)$ across checkpoints,

$$B_\ell \;=\; \sum_{t \in \mathcal{T}} b_\ell(t), \tag{25}$$

confirms this: Block 0 accumulates monotonically to $B_0 \approx 0.12$, while blocks 1–7 remain near 0.01. The same pattern holds in seed 271.

Interpretation. Block 0 (the first transformer block) sits directly above the token embeddings and is the first to process the LM-relevant token representations. Its attention and MLP weight

Table 6: Rayleigh quotients and anisotropy (seed 42). $M = 32$ mini-batches.

Step	$q(\mathbf{v}_{\mathrm{b}})$	$q(\mathbf{v}_{\mathrm{sw}})$	$q(\mathbf{v}_{\mathrm{PC2}})$	**Anisotropy** α
200 (init)	2.4×10^{-6}	2.9×10^{-6}	3.2×10^{-6}	1.3×
1800 (peak)	2.5×10^{-6}	2.2×10^{-6}	5.2×10^{-6}	1.9×
4800 (transition)	$\mathbf{1.6 \times 10^{-4}}$	8.2×10^{-5}	4.6×10^{-5}	**12.4**×
9600 (late)	$\mathbf{8.1 \times 10^{-3}}$	1.8×10^{-3}	7.0×10^{-4}	4.8×

matrices are the most directly constrained by the language modeling objective, giving the LM gradient at this level the highest temporal coherence [Saxe et al., 2014]—it consistently pulls these parameters in a stable direction determined by the language statistics.

4.5 Fisher Curvature: The Backbone Stiffens

Does the backbone become progressively harder to move along? We answer this using the Fisher information matrix, which characterizes the local curvature of the loss landscape.

Definition 3 (Empirical Fisher and Rayleigh quotient)**.** *Given M mini-batch gradients $\{\mathbf{g}_1, \dots, \mathbf{g}_M\}$ stacked into a matrix $\mathbf{G} \in \mathbb{R}^{M \times D}$, the empirical Fisher is*

$$\hat{\mathbf{F}} = \frac{1}{M} \mathbf{G}^\top \mathbf{G} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{g}_i \, \mathbf{g}_i^\top. \tag{26}$$

The Rayleigh quotient of a unit vector $\mathbf{v}$ with respect to $\hat{\mathbf{F}}$ is

$$q(\mathbf{v}) = \mathbf{v}^\top \hat{\mathbf{F}} \, \mathbf{v} = \frac{1}{M} \|\mathbf{G}\mathbf{v}\|^2 = \frac{1}{M} \sum_{i=1}^{M} \langle \mathbf{g}_i, \mathbf{v} \rangle^2. \tag{27}$$

Remark 2 (Computational trick)**.** *Computing eq.* (27) *requires only a single matrix-vector product $\mathbf{G}\mathbf{v} \in \mathbb{R}^M$, avoiding construction of the $D \times D$ Fisher matrix. Since $D \approx 25 \times 10^6$ (full trunk) and $M = 32$, this reduces memory from $O(D^2)$ to $O(MD)$.*

We define the anisotropy ratio as the Rayleigh quotient along the backbone relative to the average over random orthogonal directions:

Definition 4 (Anisotropy ratio)**.**

$$\alpha = \frac{q(\mathbf{v}_{\mathrm{b}})}{\frac{1}{K} \sum_{k=1}^{K} q(\mathbf{w}_k)}, \tag{28}$$

where $\{\mathbf{w}_1, \dots, \mathbf{w}_K\}$ are $K = 10$ random unit vectors orthogonal to $\mathbf{v}_{\mathrm{b}}$, generated by Gram–Schmidt orthogonalization of random Gaussian vectors. $\alpha > 1$ indicates that the loss landscape is stiffer along the backbone than along typical directions.

Three observations emerge from Tables 6 and 7:

1. **The backbone stiffens progressively.** $q(\mathbf{v}_{\mathrm{b}})$ increases by three orders of magnitude from initialization to late training (seed 42: $2.4 \times 10^{-6} \to 8.1 \times 10^{-3}$), outpacing all other directions.

2. **Anisotropy spikes at the λ-transition.** The moment when the probe loss weight doubles ($\lambda : 2 \to 4$ at step 4000) creates a sudden curvature increase along the backbone: $\alpha = 12.4\times$ in seed 42, $40.8\times$ in seed 271. This is the sharpest curvature event in training.

Table 7: Rayleigh quotients and anisotropy (seed 271).

Step	$q(\mathbf{v}_{\mathrm{b}})$	$q(\mathbf{v}_{\mathrm{sw}})$	$q(\mathbf{v}_{\mathrm{PC2}})$	**Anisotropy** α
200 (init)	2.0×10^{-6}	1.1×10^{-6}	1.4×10^{-6}	1.6×
2200 (peak)	1.4×10^{-5}	1.1×10^{-5}	2.5×10^{-5}	5.9×
4400 (transition)	9.5×10^{-5}	5.1×10^{-5}	5.2×10^{-5}	**40.8**×
9200 (late)	1.0×10^{-4}	4.2×10^{-5}	7.6×10^{-5}	8.3×

Table 8: Optimizer configurations for the control experiment. All runs share the same model, data, warmup (1500 steps), cosine schedule (10% floor), gradient clipping (1.0), and seed.

Run	**Optimizer**	**LR**	**Momentum**	**WD**	**WD type**
A	AdamW	10^{-3}	(β_1=0.9)	0.5	decoupled
B	SGD (no momentum)	10^{-3}	0.0	0.5	L2
C	SGD + momentum	10^{-2}	0.9	0.05	L2
C′	SGD + Nesterov (SGDW)	10^{-2}	0.9	0.5	decoupled

3. **The backbone is *not* the top Fisher eigenvector.** Extracting the leading Fisher eigenvector $\mathbf{u}_1$ via the $M \times M$ Gram matrix trick

$$(\mathbf{G}\mathbf{G}^{\top})\,\mathbf{a} \;=\; \sigma^2\,\mathbf{a} \;\implies\; \mathbf{u}_1 \;=\; \mathbf{G}^{\top}\mathbf{a}_1 \,/\, \|\mathbf{G}^{\top}\mathbf{a}_1\|, \tag{29}$$

the overlap $|\langle \mathbf{u}_1,\, \mathbf{v}_{\mathrm{b}} \rangle| \approx 0.001$—essentially zero. The backbone captures *cumulative drift*, not the direction of steepest instantaneous curvature. It is a slow, persistent mode of the optimizer dynamics rather than a mode of the loss landscape Hessian.

5 Optimizer Ablation: SGD-Family Controls

The preceding sections established that the backbone emerges from optimizer integration rather than from instantaneous gradient structure. A direct test of this claim is to replace AdamW with SGD-family optimizers while holding everything else constant.

5.1 SGD-Family Control Experiment

We trained the same model under four optimizer configurations, using identical data, initialization (seed 42), schedule, and checkpointing protocol (Table 8). The SGD variants differ from AdamW only in the optimizer; in particular, they lack per-parameter adaptive scaling.

Runs A–C trained for 4000 steps; Run C′ trained for 2000 steps with an early-stop decision rule (stop if val > 5.1 and $p_{\mathrm{ood}} < 0.02$ at step 2000). Run C′ implements SGDW—decoupled weight decay identical to AdamW's scheme—by setting `weight_decay=0` in the optimizer and manually applying $\boldsymbol{\theta} \leftarrow (1 - \eta_t \cdot \mathrm{wd})\,\boldsymbol{\theta}$ after each gradient step, isolating the effect of weight-decay coupling from adaptive scaling.

SGD without momentum failed to train (val loss remained > 8 and probe accuracy stayed at chance). Both momentum-SGD variants (C and C′) trained slowly but reached val ≈ 5.1, with weak probe OOD signal ($p_{\mathrm{ood}} \leq 0.015$) and no pronounced oscillatory switching. Nesterov momentum and decoupled weight decay produced negligible improvements over standard momentum-SGD at matched step: at step 2000, val loss differed by < 0.03 and probe OOD accuracy by ∼0.001 (Table 9).

The trajectory geometry tells a starker story. Over the analysis window $[600, 2000]$, AdamW develops a non-degenerate trajectory with $\rho_1 \approx 0.62$ and $k_{95} = 9$, while all SGD-family trajectories

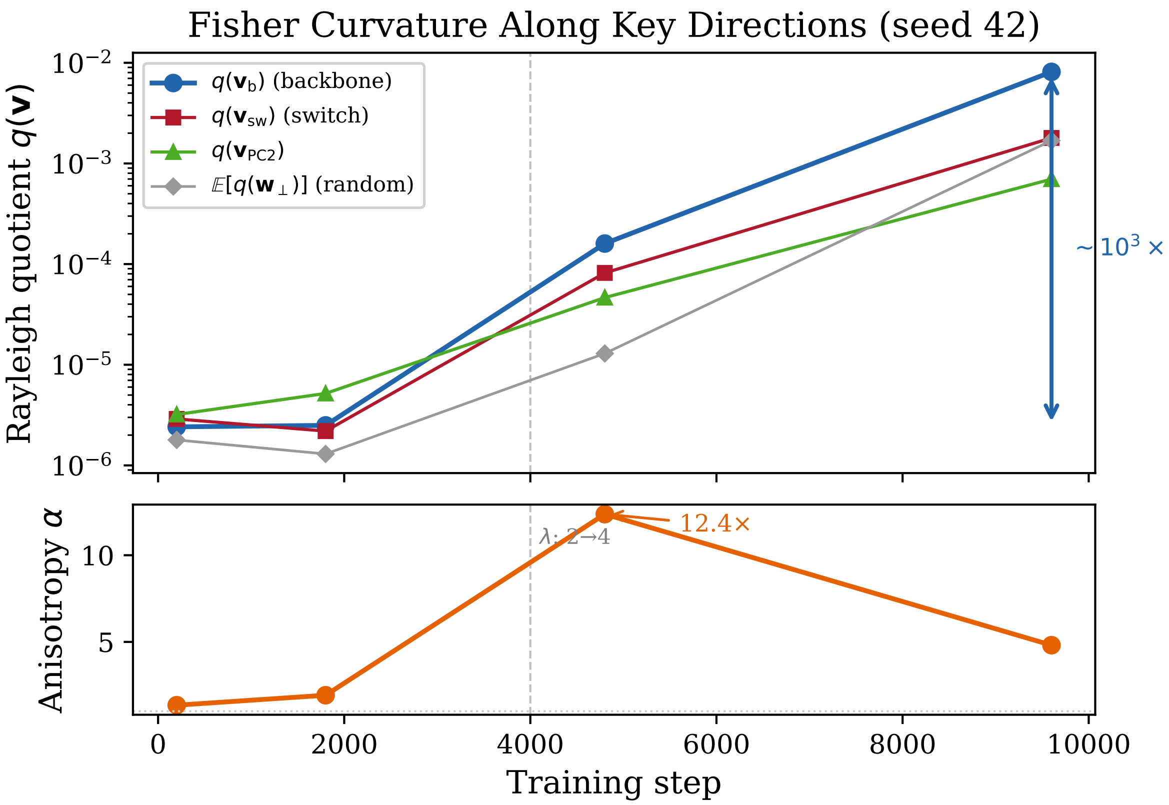

Figure 3: **Fisher curvature along the backbone stiffens over training (seed 42).** *Left:* Rayleigh quotients $q(\mathbf{v})$ for the backbone (red), switch direction (blue), and random orthogonal directions (grey). The backbone curvature increases by three orders of magnitude. *Right:* Anisotropy ratio $\alpha = q(\mathbf{v}_\mathrm{b})/\,\mathbb{E}[q(\mathbf{w}_\perp)]$ spikes at the λ-transition (step 4000) before partially relaxing.

remain nearly colinear ($\rho_1 \approx 1.0$, $k_{95} = 1$). In particular, the difference in geometry is visible before any probe oscillations occur in the AdamW run (the first probe peak at step 1800 falls within the analysis window), suggesting it reflects baseline optimizer geometry rather than oscillation-specific effects.

These results indicate that momentum alone does not produce the multi-dimensional slow–fast structure observed under AdamW, and that the key ingredient is AdamW's adaptive per-parameter scaling (Section 4.2).

5.2 Matched-Loss Geometry

A potential confound in the above comparison is unequal training progress: AdamW reaches val$\approx$ 1.6 by step 4000, far ahead of SGD+momentum's val$\approx$ 5.1. To control for this, we compare trajectory geometry at validation losses around the best regime achieved by momentum-SGD.

Challenge: AdamW has no checkpoint at val$\approx$5.2. AdamW passes through val$\approx$5.2 between steps 1 and 200 (val drops from 10.8 to 4.3), with no intermediate checkpoints. We therefore build backbone estimates on the earliest available windows (Table 10).

Even when AdamW passes through the same loss range early in training, its drift is already non-colinear: PC1 explains only 69–82% of row-normalized displacement energy with $k_{95} = 2$–3 over windows spanning this regime. In contrast, momentum-SGD remains nearly colinear at matched loss ($\rho_1 \approx 0.98$–1.00, $k_{95} = 1$) and exhibits extremely small drift ($\|\Delta\boldsymbol{\theta}\| < 1$) over the

Table 9: Training outcomes for all optimizer variants (seed 42). PC1 and k_{95} are from uncentered drift-matrix PCA on the analysis window $[600, 2000]$ with anchor at step 600.

Run	**Final val**	**Best p_{ood}**	**PC1 (%)**	k_{95}	**Drift**
A (AdamW)	1.59	0.433	61.5	9	113.7
B (SGD no-mom)	8.12	0.000	100.0	1	40.2
C (SGD+mom)	5.10	0.015	100.0	1	54.2
C′ (SGDW+Nesterov)[†]	5.25	0.013	—	—	—

[†] Run C′ stopped at step 2000. At that point, C′ differed from C by <0.03 in val loss and ∼0.001 in p_{ood}; geometry analysis was not pursued.

Table 10: Backbone geometry at matched operating regime. AdamW is analyzed over early windows (val ≈ 4.3–3.0, already below SGD+mom's best); SGD+mom is analyzed over its plateau (val ≈ 5.1–5.2). Drift is $\|\boldsymbol{\theta}(t_{\text{end}}) - \boldsymbol{\theta}(t_{\text{start}})\|$.

Optimizer	**Window**	**PC1 (%)**	k_{95}	k_{99}	**Drift**
At/near val ≈ 5.2 (matched regime)					
AdamW	$[1, 200, 400]$[a]	77.4	2	2	—
AdamW	$[1, 200, 400, 600]$[a]	69.2	3	3	24.9
AdamW	$[200, 400, 600]$[b]	81.9	2	2	—
SGD+mom	$[2000, 2200, 2400]$	98.1	1	2	0.26
SGD+mom	$[1800, \dots, 2600]$	97.7	1	2	0.60
Standard analysis window (reference)					
AdamW	$[200, \dots, 1000]$	77.8	4	8	50.7
AdamW	$[600, \dots, 2000]$	61.5	9	19	113.7
SGD+mom	$[600, \dots, 2000]$	100.0	1	1	54.2

[a] Window spans val from 10.8 to 3.0–3.4; passes through val ≈ 5.2 between steps 1 and 200.

[b] Window starts at val = 4.3, already below SGD+mom's best; this makes the comparison conservative.

corresponding plateau windows. Decoupled weight decay (SGDW) and Nesterov momentum produce negligible changes relative to standard momentum-SGD (Table 9).

The drift magnitudes are revealing: AdamW traverses 24.9 units of parameter drift between steps 1 and 600, while SGD+momentum moves only 0.26 units across its entire plateau window $[2000, 2400]$. The SGD trajectory is not merely low-rank; it is nearly *stationary* during its low-loss plateau.

These results indicate that AdamW's adaptive per-parameter scaling induces qualitatively richer trajectory geometry than SGD-family variants, even at comparable validation loss, supporting an optimizer-specific mechanism for the emergence of non-degenerate slow–fast structure.

6 Reheating: Re-Entering the Probe Basin

6.1 Protocol

From the endpoint of training (step 10,000; $p_{\text{ood}} \approx 0.16$, deep in the LM-dominant regime), we resume training with doubled probe loss weight $\lambda = 4.0$ and a *fresh* AdamW optimizer (zeroed momentum and second-moment buffers). Three learning rates are tested: $\eta \in \{10^{-3}, 6 \times 10^{-4}, 3 \times 10^{-4}\}$. Each reheating run lasts 2,000 steps with a cosine learning rate schedule.

The rationale is straightforward: if the probe attractor still exists in the loss landscape, a sufficiently strong gradient signal should be able to push the model back into it. The fresh

Table 11: Reheating results (seed 42). The model starts from $p_{\text{ood}} = 0.16$.

Learning Rate	**Peak p_{ood}**	**At Step**	**First ≥ 0.60**	**Final (step 2000)**
10^{-3}	0.705	900	step 600	0.221
$\mathbf{6 \times 10^{-4}}$	**0.782**	**1000**	step 700	0.279
3×10^{-4}	0.578	1500	—	0.421

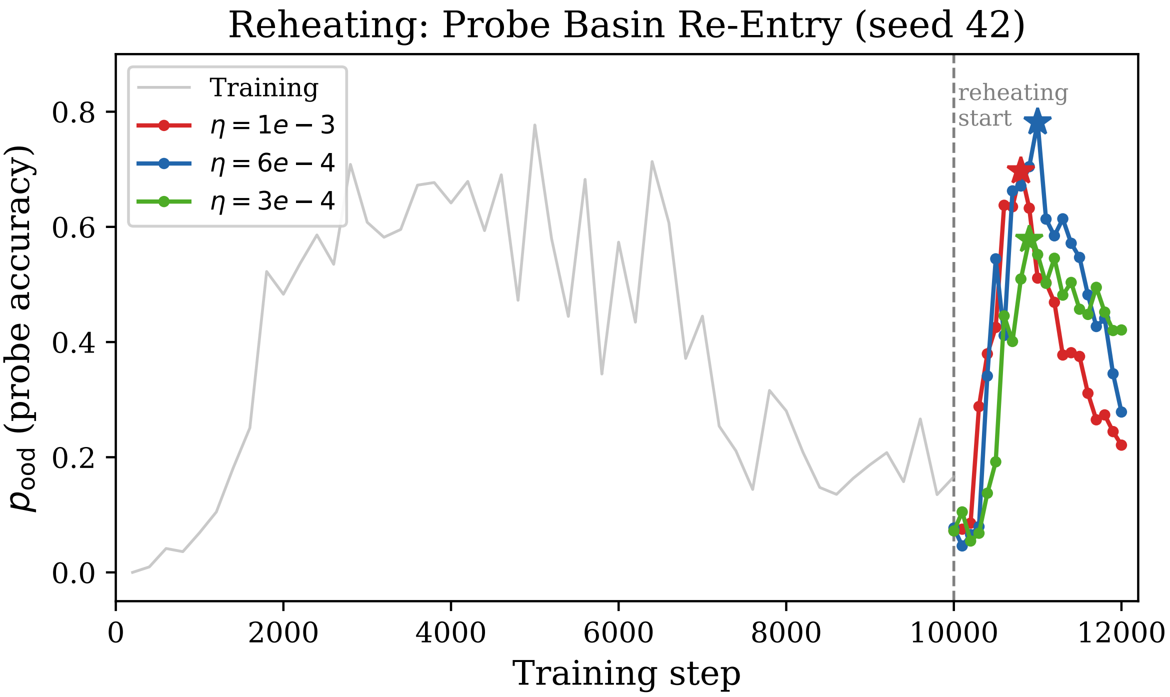

Figure 4: **Reheating trajectories (seed 42).** Three learning rates are tested from the step-10,000 endpoint (grey background: original training). All three achieve probe re-entry, but the effect is transient: p_{ood} peaks and then decays as the cosine schedule reduces η_t. The optimal LR (6×10^{-4}, orange) exceeds the original training peak.

optimizer ensures that momentum from the original training run does not confound the result.

Importantly, reheating primarily perturbs the transverse residual $\|\mathbf{r}(t)\|$ while leaving the accumulated backbone coordinate $a(t)$ largely unchanged. This indicates that the slow manifold persists even when the model temporarily re-enters the probe regime.

6.2 Results

At the optimal learning rate (6×10^{-4}), the model reaches $p_{\text{ood}} = 0.782$—exceeding the training-time peak of 0.777—within 1000 reheating steps (Table 11). The probe basin remains geometrically present in the late-training landscape. Reheating reveals that it is not erased but becomes dynamically suppressed by backbone stiffening.

But re-entry is unstable. After peaking, p_{ood} decays to 0.28 by step 2000. The probe basin has become a transient saddle: reachable but not sustainable under these dynamics.

$\eta = 3 \times 10^{-4}$ is below the threshold for full re-entry—it provides insufficient gradient drive to overcome the curvature barrier separating the LM and probe basins.

6.3 Connection to Backbone Stiffening

The transient nature of reheating is consistent with the backbone stiffening documented in Section 4.5. During training, the backbone direction accumulates Fisher curvature. By step 10,000, $q(\mathbf{v}_\mathrm{b})$ has increased by three orders of magnitude relative to initialization.

Reheating exposes the competition between transverse probe drive and backbone curvature. The backbone coordinate behaves as a stiffened slow variable, while transverse modes respond rapidly to increased probe weighting. Let the component of the probe gradient along the backbone be $g_{\|}(t)$ and the curvature along the backbone be $\kappa(t) \approx q(\mathbf{v}_\mathrm{b})$. The backbone coordinate evolves approximately as

$$\frac{da}{dt} \;\approx\; -\eta_t\, g_{\|}(t) \;-\; \eta_t\, \kappa(t)\, a(t), \tag{30}$$

where the first term is the probe-driven drift and the second is the curvature-mediated restoring force. When κ is large (stiffened backbone), only a large η_t can overcome it; as the cosine schedule reduces η_t, the restoring force dominates and the model returns to the LM basin.

The sign flip in update–backbone alignment (Section 4.3) provides a complementary perspective. In early training, the optimizer drifts in the $-\mathbf{v}_\mathrm{b}$ direction (toward the probe basin). After the sign flip ($t^* \approx 5000$), the drift reverses—the optimizer moves along $+\mathbf{v}_\mathrm{b}$, away from the probe basin. Reheating temporarily reverses this via the strong probe gradient, but once η_t decays, the backbone's natural $+\mathbf{v}_\mathrm{b}$ drift reasserts itself.

6.4 Two-Seed Comparison

Seed 271 reheating shows the same qualitative pattern—transient probe re-entry followed by decay—though the peak p_ood is lower (0.36–0.42 vs. 0.78). Both seeds use identical hyperparameters (Table 1); the quantitative difference likely reflects seed-dependent differences in the late-training loss landscape geometry—in particular, the depth and width of the probe basin at the reheating start point—rather than any difference in the reheating protocol itself.

Thus reheating does not contradict backbone dominance; it demonstrates that switching dynamics are transverse excursions around a persistent slow manifold shaped by optimizer integration.

7 Switching Lives in the Transverse Subspace

With the backbone established as the dominant geometric feature, what can we say about the oscillatory dynamics? They occur primarily in the transverse subspace.

The switching direction between a peak and adjacent trough of p_ood is

$$\mathbf{v}_\mathrm{sw} \;=\; \frac{\boldsymbol{\theta}_\mathrm{peak} - \boldsymbol{\theta}_\mathrm{trough}}{\|\boldsymbol{\theta}_\mathrm{peak} - \boldsymbol{\theta}_\mathrm{trough}\|}. \tag{31}$$

Its alignment with the backbone is:

Seed	$\lvert\langle \mathbf{v}_\mathrm{sw}, \mathbf{v}_\mathrm{b}\rangle\rvert$ **(per-block range)**
42	0.20–0.25
271	0.28–0.31

We can decompose the switch direction into backbone and transverse components:

$$\mathbf{v}_\mathrm{sw} \;=\; \underbrace{\langle \mathbf{v}_\mathrm{sw}, \mathbf{v}_\mathrm{b}\rangle\, \mathbf{v}_\mathrm{b}}_{\text{backbone component}} \;+\; \underbrace{\mathbf{v}_\mathrm{sw} - \langle \mathbf{v}_\mathrm{sw}, \mathbf{v}_\mathrm{b}\rangle\, \mathbf{v}_\mathrm{b}}_{\text{transverse component}}. \tag{32}$$

The backbone component accounts for $\langle \mathbf{v}_{\mathrm{sw}},\, \mathbf{v}_{\mathrm{b}} \rangle^2 \approx 0.04$–$0.10$ of the switching direction's variance. Switching is approximately **80% transverse** to the backbone.

Furthermore, different switching events use near-orthogonal directions (pairwise $|\cos| < 0.08$), and the switching manifold spans at least 10 independent dimensions in the 25M-dimensional trunk space.

To quantify how much of the transverse switching direction is captured by the leading residual PCs, we project out the backbone component to obtain $\mathbf{v}_{\mathrm{sw}}^{\perp} = \mathbf{v}_{\mathrm{sw}} - \langle \mathbf{v}_{\mathrm{sw}},\, \mathbf{v}_{\mathrm{b}} \rangle\, \mathbf{v}_{\mathrm{b}}$ (renormalized to unit length) and compute the energy fraction captured by PCs 2–6:

$$E_{2:6} \;=\; \sum_{k=2}^{6} \langle \hat{\mathbf{v}}_{\mathrm{sw}}^{\perp},\, \mathbf{v}_k \rangle^2, \tag{33}$$

where $\mathbf{v}_k$ are the k-th right singular vectors from the uncentered trajectory PCA. Per block, $E_{2:6}$ ranges from 15–22% in seed 42 and 60–67% in seed 271. In seed 42, the transverse switching direction is largely orthogonal to the low-rank PC subspace, meaning it lives in the high-dimensional tail of the trajectory variance. In seed 271, PCs 2–6 capture the majority of the transverse switch. In both cases the switching dynamics are distributed across multiple residual dimensions rather than concentrated on a single transverse mode.

The picture is: **the optimizer rides a one-dimensional backbone rail while oscillating in a high-dimensional transverse cloud**.

8 Discussion

8.1 Slow–Fast Decomposition of Optimizer Dynamics

The backbone–transverse decomposition reveals a structural separation in transformer training under AdamW. Long-horizon parameter evolution concentrates in a low-dimensional drift direction, while oscillatory objective competition unfolds in a high-dimensional transverse subspace.

This separation is not imposed by architecture or by explicit constraints in the loss. It emerges from optimizer dynamics. Momentum integrates weak but temporally coherent gradient bias, and adaptive normalization suppresses incoherent variance. The result is an effective timescale separation: a slow cumulative drift manifold and faster transverse fluctuations.

Formally, the dynamics resemble a singularly perturbed system,

$$\frac{d\boldsymbol{\theta}}{dt} \;=\; f_{\mathrm{slow}}(\boldsymbol{\theta}) \;+\; \epsilon^{-1}\, f_{\mathrm{fast}}(\boldsymbol{\theta}), \tag{34}$$

where the slow component corresponds to backbone drift and the fast component to transverse switching. The scale separation parameter reflects the ratio between temporal gradient coherence and transverse fluctuation variance. This framing shifts the focus of optimization analysis from static curvature to trajectory-level structure.

8.2 Instantaneous Freedom vs. Cumulative Coherence

The backbone does not imply that gradients are low-dimensional. At each step, gradients remain high-dimensional and largely isotropic relative to the backbone. The distinction lies between instantaneous variability and cumulative displacement.

In high dimensions, large transverse components can dominate the norm of individual updates yet cancel over time if their directions fluctuate. A much smaller but temporally coherent component can therefore dominate long-horizon displacement. The backbone reflects this integrated coherence rather than a restriction of gradient expressivity.

This distinction clarifies why the backbone is nearly orthogonal to top Fisher eigenvectors. It is not the steepest local direction of the loss landscape, but the direction in which optimizer integration accumulates persistent bias.

8.3 Optimizer-Induced Geometry

The SGD-family controls demonstrate that this structure is not a generic property of the objective. Under SGD with or without momentum, trajectories remain nearly colinear and fail to develop multi-dimensional drift structure, even at matched validation loss.

Adaptive per-parameter scaling appears to be the key ingredient. By normalizing coordinates according to gradient variance, AdamW selectively amplifies temporally coherent components and suppresses incoherent ones. This produces qualitatively richer cumulative geometry than momentum alone.

These results suggest that optimizer choice influences not only convergence speed or stability, but the geometric structure of the training trajectory itself.

8.4 Dynamical Consequences: Reheating and Basin Accessibility

Reheating experiments provide a dynamical probe of this structure. Transverse probe dynamics can be transiently re-excited from late-training checkpoints, but accumulated backbone drift remains largely intact. As curvature along the backbone increases, larger learning rates are required to overcome the associated restoring force. When the learning rate decays, the system relaxes back toward the backbone-aligned regime.

Thus reheating does not contradict backbone dominance; it reveals that switching dynamics correspond to transverse excursions around a persistent slow manifold shaped by optimizer integration.

8.5 Limitations and Scope

This study uses a 51M-parameter transformer and a synthetic probe objective. Whether larger models or natural multi-task settings produce higher-dimensional or more distributed backbone structure remains an open question.

The Fisher analysis relies on a low-rank approximation (32 mini-batches), and anisotropy ratios may be underestimates. While results replicate across two seeds, broader variation may reveal additional structure.

Despite these limitations, the optimizer ablation and matched-loss comparisons indicate that the backbone phenomenon is robust and mechanistically grounded.

9 Methods

9.1 Trunk Parameters

All geometric analyses use trunk-only parameters: weight matrices in attention (query, key, value, output projection) and MLP (up-projection, down-projection) across all 8 blocks. This excludes tied embeddings, causal masks, positional embeddings, and layer normalization parameters. Total trunk dimensionality: ∼25M parameters (∼3.1M per block).

9.2 Uncentered PCA

For each transformer block $\ell \in \{0, \ldots, 7\}$, the drift matrix $\mathbf{X}^{(\ell)} \in \mathbb{R}^{T \times D_\ell}$ has rows

$$\mathbf{x}^{(\ell)}(t) \;=\; \mathrm{flatten}_\ell\big(\boldsymbol{\theta}(t)\big) \;-\; \mathrm{flatten}_\ell\big(\boldsymbol{\theta}(0)\big). \tag{35}$$

SVD of $\mathbf{X}^{(\ell)}$ (no mean centering) yields the block backbone $\mathbf{v}_{\mathrm{b}}^{(\ell)} \in \mathbb{R}^{D_\ell}$ as the first right singular vector.

9.3 Update-Direction Alignment

The 200-step update $\mathbf{u}(t) = \boldsymbol{\theta}(t) - \boldsymbol{\theta}(t - 200)$ is computed from consecutive checkpoints. This captures the net effect of AdamW (preconditioner, momentum, weight decay, gradient clipping). Alignment is reported as $C(t) = \langle \mathbf{u}(t), \mathbf{v}_{\mathrm{b}} \rangle / \|\mathbf{u}(t)\|$.

9.4 Rayleigh Quotient Computation

Given a direction $\mathbf{v} \in \mathbb{R}^D$ and a gradient matrix $\mathbf{G} \in \mathbb{R}^{M \times D}$:

$$q(\mathbf{v}) \;=\; \frac{1}{M}\,\|\mathbf{G}\mathbf{v}\|^2 \;=\; \frac{1}{M}\sum_{i=1}^{M} \langle \mathbf{g}_i, \mathbf{v} \rangle^2. \tag{36}$$

This requires one matrix–vector product ($O(MD)$ operations, $O(M)$ storage for the result), avoiding construction of the $D \times D$ Fisher. Anisotropy uses $K = 10$ random orthogonal directions generated by Gram–Schmidt orthogonalization of Gaussian random vectors projected orthogonal to $\mathbf{v}_{\mathrm{b}}$.

9.5 Reheating Protocol

Resume from step 10,000 checkpoint. Fresh AdamW optimizer (zeroed $\mathbf{m}_0$, $\mathbf{v}_0$). Composite loss weight $\lambda = 4.0$. Cosine learning rate schedule over 2,000 steps. Evaluate every 100 steps. Three learning rate values: $\{10^{-3}, 6 \times 10^{-4}, 3 \times 10^{-4}\}$.

Acknowledgments

[To be added.]

References

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

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023. URL `https://arxiv.org/abs/2305.07759`.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, pages 3259–3269, 2020. URL `https://arxiv.org/abs/1912.05671`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL `https://arxiv.org/abs/1412.6980`.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of neural network training. *arXiv preprint arXiv:2002.10434*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1711.05101`.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL `https://arxiv.org/abs/1412.1193`.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR Workshop on MATH-AI*, 2022. URL `https://arxiv.org/abs/2201.02177`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL `https://arxiv.org/abs/1312.6120`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL `https://arxiv.org/abs/1706.03762`.