BEYOND SINGULAR VALUES: GEOMETRY AND NUMERICAL FIDELITY IN MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

When does spectral flattening improve the loss decrease from a matrix update? We study this question for Muon through rotations that preserve the candidate spectrum, update norm, and entire ideal polar update while changing the orientation of a spectrum-preserving comparator. On one fixed objective, candidates with identical spectra and gradient alignments can favor opposite updates; a local theorem identifies the curvature projection responsible for this variation. At finite step sizes, the gradient at the comparator endpoint determines the first-order response to a small rotation and accounts for curvature along the update path. Tests span 30 frozen state–block combinations from three 124-million-parameter language-model trajectories. In the first layer, scalars fitted on old text recover most of the endpoint predictor's gain. In middle and final layers, the same-objective endpoint expansion remains accurate, while calibration-to-evaluation differences dominate prediction error. We then separate the exact five-round spectral map from its floating-point implementation. In a first-layer same-input test with raw output lengths, single precision (FP32) reduces paired-response squared error by 99.97% compared with brain floating point (BF16). FP32 preserves the ideal response sign in all 36 first-layer and 24 deeper-layer constructed directions; the high-precision finite-round map reproduces the remaining random-direction reversals. These results distinguish orientation and path effects from data-transfer error and numerical distortion at the intervention scale.

1 INTRODUCTION

Muon makes spectral reshaping a practical ingredient of language-model optimization (Jordan et al., 2024; Liu et al., 2025). It transforms momentum matrices with an approximate polar decomposition before updating hidden-layer parameters. The ideal transformation retains the singular directions and sets every nonzero singular value to one, shifting relative update magnitude toward directions with smaller candidate singular values. To explain the loss benefit of this redistribution, we need to know how those directions meet the loss landscape at the step actually taken. This connection matters for optimizer design and for experiments that attribute a loss benefit to spectral flattening. Such a benefit can also depend on update length, gradient alignment, or the numerical approximation, and each explanation suggests a different algorithmic change.

We ask a concrete question: *at a fixed model state and finite step size, what information determines whether flattening a matrix update lowers the loss?* We compare an ideal polar update with a spectrum-preserving comparator of the same Frobenius norm, the square root of the sum of squared matrix entries. Equal length isolates the allocation of update magnitude across directions. The remaining question is whether orientation can change the preference while the candidate spectrum, gradient alignments, and polar update remain fixed. A controlled answer would distinguish information supplied by the spectrum from information supplied by its orientation relative to the loss.

Our starting point is a family of joint rotations of the two singular bases. These rotations preserve the candidate singular values, both singular subspaces, both update norms, and the entire ideal polar update. With the polar endpoint fixed, every change in ideal relative loss comes from the spectrum-preserving comparator. Within this family, two candidates on one strictly convex quadratic loss have identical gradient alignments but opposite update preferences. A local theorem identifies the curvature component that changes the advantage along exactly alignment-preserving rotations. The

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.

construction turns the broad observation that curvature matters into a test with explicit invariants and a specified direction of variation.

The neural-network test must resolve two distinct scales: the optimizer step and the smaller orientation perturbation. A small rotation moves the comparator endpoint only slightly, although the optimizer step leading there can be substantial. The gradient at that endpoint therefore determines the local response and incorporates curvature along the update path. Different text samples introduce a second source of error: the gap between calibration and evaluation gradients can dominate the local expansion remainder. Separately, the matrix routine must resolve the small difference induced by the rotation. Accuracy of an individual output and accuracy of this difference are different requirements: a numerical implementation can move the polar endpoint that the ideal intervention holds fixed. We separate the exact finite-round map from input rounding and matrix-arithmetic error, then project the resulting endpoint motion onto the loss. This distinguishes a geometric effect from its distortion by the computation used to measure it.

We make three contributions, combining theory with one-step loss-response measurements across 30 state–block combinations:

- **An orientation test with a fixed polar endpoint.** We give a same-objective preference reversal and a generic local theorem under exact alignment constraints, showing which curvature information remains after spectral and first-order quantities have been fixed (Section 3).
- **A finite-step account of the response.** We use endpoint gradients to account for the update path and separate data-transfer error from the local expansion remainder (Section 4). First-layer tests show that simple scalar controls recover most of the predictive gain. Middle- and final-layer tests separate an accurate same-objective expansion from a larger calibration-to-evaluation gap (Sections 6.2 and 6.5).
- **A loss-level separation of approximation and arithmetic.** For the tested five-round routine, we derive a paired-response decomposition and a sufficient sign-preservation bound (Section 5). In the first-layer same-input test, single precision (FP32) reduces raw-output mean squared response error by 99.97% relative to brain floating point (BF16). All 36 first-layer and 24 deeper-layer constructed directions retain the ideal response sign. A high-precision finite-round reference reproduces the remaining random-direction reversals (Sections 6.3–6.5).

2 Related work

Adaptive and matrix-aware optimization. AdaGrad adapts the update geometry to accumulated gradients (Duchi et al., 2011), while Adam uses gradient moments and AdamW separates weight decay from the adaptive update (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Shampoo preserves tensor structure through matrix preconditioners (Gupta et al., 2018), and SOAP applies Adam in a preconditioner eigenbasis (Vyas et al., 2025). Norm-based steepest descent offers a common view of several such directions (Bernstein & Newhouse, 2024). Our equal-Frobenius-length comparison tests the allocation of update magnitude across directions at one fixed state.

Muon and loss geometry. Muon implementations and scaling studies establish the practical role of approximate orthogonalization (Jordan et al., 2024; Liu et al., 2025); fractional spectral powers interpolate between flattening and spectrum preservation (Dong & Sawin, 2026). Curvature-based accounts compare Muon with Adam (Wang et al., 2026), while river-valley analysis describes stage-dependent behavior (Shen et al., 2026). Mousse and Newton-Muon incorporate curvature-aware preconditioning (Zhang et al., 2026; Du & Su, 2026), following the broader use of structured curvature approximations (Martens & Grosse, 2015). We instead vary orientation under joint spectral, alignment, and fixed-polar constraints. Full Hessian–vector products (Pearlmutter, 1994) select local intervention directions, and endpoint gradients measure their response at the chosen step size.

Polar computation and constrained perturbations. Classical polar algorithms analyze Newton and weighted Halley iterations (Higham, 1986; Nakatsukasa et al., 2010); Polar Express develops polynomial matrix-sign methods for Muon (Amsel et al., 2026). Gawlik & Leok (2017) compute

derivatives of the exact polar map, and matrix-manifold methods supply the standard tools for constrained tangent motion (Absil et al., 2008). Our joint rotations leave that exact polar map constant but allow its finite-round approximation to vary. The resulting spectral-gap bound links this variation to the comparator chord and then to the paired loss response.

Precision and response fidelity. Numerical analysis distinguishes approximation, conditioning, and rounding error (Higham, 2002), while mixed-precision methods balance accuracy against computational cost (Higham & Mary, 2022; Micikevicius et al., 2018). We examine the accuracy of a small *difference* between outputs of one fixed polynomial routine. The exact finite-round reference separates approximation bias from arithmetic error, so higher precision and a different approximation become distinct interventions with measurable loss effects.

3 Orientation changes utility without changing the spectrum

To see what spectral information leaves unresolved, we begin with an equal-length comparison and then vary orientation while fixing the ideal polar endpoint.

3.1 A length-matched comparison

To separate spectral shape from update magnitude, let L be a twice continuously differentiable loss of one matrix parameter block $W \in \mathbb{R}^{m\times n}$, with all other parameters fixed. Write $G = \nabla L(W)$ and $H = \nabla^2 L(W)$, and use the Frobenius inner product and norm throughout. For a nonzero candidate with compact singular value decomposition $M = U\Sigma V^\top$ and rank r, define the ideal polar update P and its length-matched, spectrum-preserving comparator C by

$$P = UV^\top, \qquad D = \frac{\sqrt{r}}{\|\Sigma\|_F}\Sigma, \qquad C = UDV^\top, \qquad \|C\|_F = \|P\|_F = \sqrt{r}. \tag{1}$$

The compact factors U, V have orthonormal columns, and Σ contains the r positive singular values; P is the polar factor on these active subspaces. The experiments use a momentum candidate M; we also examine the special case $M = G$. At step size $\eta > 0$, positive finite-step advantage means the polar update produces lower loss:

$$\mathcal{A}(\eta) = L(W - \eta C) - L(W - \eta P), \tag{2}$$

$$\mathcal{A}(\eta) = \eta a + \frac{\eta^2}{2}\Delta q + R(\eta), \quad a = \langle G, P - C\rangle_F, \quad \Delta q = \langle C, H[C]\rangle_F - \langle P, H[P]\rangle_F. \tag{3}$$

For a twice continuously differentiable loss, $R(\eta) = o(\eta^2)$; a ρ-Lipschitz Hessian along the two update segments gives $|R(\eta)| \le \rho\eta^3 r^{3/2}/3$. The first term measures gradient alignment, and the second measures the difference in directional curvature costs. When $M = G$, the identity $a = \|G\|_* - \sqrt{r}\,\|G\|_F \le 0$ shows that equal-length flattening incurs a first-order alignment cost unless the nonzero spectrum is flat. Here $\|G\|_*$ is the sum of singular values, and Cauchy–Schwarz gives the inequality. For a quadratic loss with $a\Delta q < 0$, the nonzero crossover is exactly $\eta^* = -2a/\Delta q$.

A separable curvature model makes this orientation dependence explicit: if $H[X] = BXA$ for symmetric positive-semidefinite factors, then

$$\Delta q = \operatorname{tr}\big[(U^\top BU)\{D(V^\top AV)D - V^\top AV\}\big]. \tag{4}$$

The effect depends on how left and right singular directions pair with curvature, not just on each matrix's eigenvalues or singular values.

3.2 A fixed-polar intervention and a generic local result

We isolate orientation by jointly rotating both singular bases with an orthogonal matrix $R \in O(r)$ while fixing the polar endpoint:

$$M_R = UR\Sigma R^\top V^\top, \qquad C_R = URDR^\top V^\top, \qquad P_R = P. \tag{5}$$

This family, which we call a fixed-polar fiber, preserves the singular values, both singular subspaces, the update norms, and $\|P - C_R\|_F$. Because candidates share the ideal polar loss, advantage changes equal comparator-loss changes.

Theorem 1 (Opposite preferences on one fixed objective)**.** *There exist a strictly convex quadratic loss, a state, and two candidates in the same fixed-polar fiber with equal individual gradient alignments but opposite exact advantages at one common positive step. Both comparator updates and their shared polar update decrease this loss.*

For an explicit example, take $W = 0$, $G = \mathrm{diag}(3, 1)$, $D = \mathrm{diag}(7/5, 1/5)$, and $L(X) = \langle G, X\rangle_F + \frac{1}{2}\mathrm{tr}(XAX^\top)$ with $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. Rotations by $\pm\pi/8$ give $P = I_2$ and, at $\eta = 1/2$, exact advantages 0.14544 and -0.19397. Appendix A verifies all invariants and descent inequalities. These opposite preferences isolate orientation information with spectrum, norm, gradient alignment, step size, and the polar endpoint fixed.

To identify the missing local quantity, set $E = \mathrm{sym}(U^\top GV)$ and $F = \mathrm{sym}(U^\top H[C]V)$, where $\mathrm{sym}(X) = (X + X^\top)/2$. The possible update velocities are $UBV^\top$ with B in the symmetric moving space $\mathcal{S}_D$, and first-order alignment preservation restricts them to $\mathcal{T}_D$:

$$\mathcal{S}_D = \{B = B^\top : B_{ij} = 0 \text{ if } d_i = d_j\}, \quad E_D = \Pi_{\mathcal{S}_D} E, \quad \mathcal{T}_D = \mathcal{S}_D \cap E_D^\perp, \quad F_T = \Pi_{\mathcal{T}_D} F. \tag{6}$$

Here Π denotes orthogonal projection under the Frobenius inner product.

Theorem 2 (Generic local non-identifiability)**.** *Fix U, D, V, W, G with $s = \dim \mathcal{S}_D \geq 2$ and $E_D \neq 0$. Within the vector space of Frobenius-self-adjoint Hessian operators, the set $\{H : F_T = 0\}$ is a linear subspace of codimension $s - 1$. For every Hessian outside this subspace, there is a smooth local fiber curve $C(t)$ with $C(0) = C$, $\left\|\dot{C}(0)\right\|_F = 1$, and exactly constant $\langle G, C(t)\rangle_F$. For the quadratic loss with gradient G and Hessian H at W, this curve satisfies*

$$\frac{d}{dt}\mathcal{A}(C(t))\bigg|_{t=0} = \eta^2 \|F_T\|_F > 0. \tag{7}$$

For a twice continuously differentiable loss with the same G, H at W and a ρ-Lipschitz Hessian along $W - s\eta C$, $0 \leq s \leq 1$, the derivative differs from the right-hand side by at most $\rho\eta^3 r/2$.

The proof represents every moving velocity as a commutator, preserves exact alignment with a scalar correction, and establishes surjectivity of the Hessian-to-response map (Appendix A). The exceptional Hessian set has measure zero, also within the positive-definite cone. The regularity condition $E_D \neq 0$ makes the exact alignment level set smooth; $s \geq 2$ leaves a moving direction on that set. For a nonquadratic loss, the same direction increases the advantage whenever $\rho\eta r/2 < \|F_T\|_F$. Appendix A describes the degenerate cases. Equation (7) characterizes local variation of the advantage; a preference reversal occurs when the advantage itself crosses zero.

4 The relevant curvature lies along the update path

A finite optimizer step may traverse changing curvature, so we evaluate orientation changes through the gradient at the baseline comparator endpoint. Write $C' = C + \Delta$ for two points on the fiber and define their ideal advantage response Y_L and endpoint linear prediction p_L by

$$Y_L = L(W - \eta C') - L(W - \eta C), \qquad p_L = -\eta \langle \nabla L(W - \eta C), \Delta\rangle_F . \tag{8}$$

Proposition 3 (Endpoint and path identities)**.** *If L is twice continuously differentiable near the baseline path and endpoint chord, then*

$$p_L = -\eta \langle G_L(W), \Delta\rangle_F + \eta^2 \left\langle \int_0^1 H_L(W - s\eta C)[C]\, ds, \Delta \right\rangle_F , \tag{9}$$

$$Y_L = p_L + R_{C,L}, \qquad |R_{C,L}| \leq \frac{\beta_C \eta^2}{2} \|\Delta\|_F^2 , \tag{10}$$

where the bound assumes that the Hessian operator norm is at most β_C along the endpoint chord.

The remainder is quadratic in the intervention chord $\eta\Delta$. The origin-Hessian approximation replaces the integral by $H_L(W)[C]$ and changes p_L by at most $\rho\eta^3 \|C\|_F^2 \|\Delta\|_F /2$ under a ρ-Lipschitz Hessian. A small orientation perturbation therefore supports an endpoint expansion even when curvature changes substantially over the baseline step.

Because the direction is calibrated on data S and assessed on data X, their endpoint gradients must be separated:

$$Y_X - p_S = (p_X - p_S) + (Y_X - p_X). \tag{11}$$

The first term is the change of objective between data samples, and the second is the finite-chord remainder on the evaluation objective. Exact gradient neutrality on S removes the origin-gradient term on S, but generally leaves that term on X.

5 From the ideal polar map to numerical computation

The ideal analysis fixes the polar endpoint, but finite iteration and floating-point arithmetic can move it.

5.1 Finite iteration and floating-point error are separate

We separate the real-arithmetic finite-round map from input rounding and matrix-arithmetic error. The tested Newton–Schulz-style routine applies the quintic polynomial $q(z) = 3.4445z - 4.775z^3 + 2.0315z^5$ for k rounds after Frobenius normalization with $\varepsilon = 10^{-7}$ (Jordan et al., 2024). The main setting uses five rounds ($k = 5$).

Proposition 4 (Exact finite-round response)**.** *Let $z_i^{(0)} = \sigma_i/(\|\Sigma\|_F + \varepsilon)$, $z_i^{(j+1)} = q(z_i^{(j)})$, and $Z_k = \mathrm{diag}(z^{(k)})$. In real arithmetic, the matrix routine satisfies*

$$T_k^*(M_R) = URZ_kR^\top V^\top, \qquad \delta T_k^* = U(RZ_kR^\top - Z_k)V^\top, \tag{12}$$

$$\|\delta T_k^*\|_F \le \gamma_k \|\Delta\|_F, \qquad \gamma_k = \max_{d_i \neq d_j} \frac{|z_i^{(k)} - z_j^{(k)}|}{|d_i - d_j|}, \tag{13}$$

with an empty maximum defined as zero.

The ideal polar response is zero; any nonuniformity in Z_k allows the finite-round output to change under rotation. For norm-matched outputs, replace Z_k by $\sqrt{r}Z_k/\|Z_k\|_F$ when $\|Z_k\|_F > 0$.

Let Q_{in} denote input rounding and $\widehat{T}$ the complete numerical map, and define input and arithmetic errors by

$$E_{\mathrm{in}}(M) = T_k^*(Q_{\mathrm{in}}M) - T_k^*(M), \qquad E_{\mathrm{arith}}(M) = \widehat{T}(M) - T_k^*(Q_{\mathrm{in}}M),$$
$$\delta\widehat{T} = \delta T_k^* + \delta E_{\mathrm{in}} + \delta E_{\mathrm{arith}}. \tag{14}$$

Here $\delta f = f(M_R) - f(M)$ for every map f, and T_k^* denotes the polynomial matrix routine on any input, including a rounded input with a different rank. The identity also holds with output norm matching, provided all maps use the same fixed target norm $\sqrt{r}$ and every normalized output is nonzero. Since $\delta P = 0$, we measure intervention errors against the nonzero comparator chord or finite-round reference response.

5.2 How a moving numerical endpoint changes the loss response

We next translate this matrix motion into a loss response. For $F_L(A) = L(W - \eta A)$, let $\widehat{T}_0 = \widehat{T}(M)$, $\widehat{T}_1 = \widehat{T}(M_R)$, and define the actual paired advantage response and its realization bias by

$$\widehat{Y}_L = Y_L - [F_L(\widehat{T}_1) - F_L(\widehat{T}_0)], \qquad B_L = \widehat{Y}_L - Y_L. \tag{15}$$

Proposition 5 (Loss-level realization bound)**.** *Suppose L is twice continuously differentiable and its Hessian operator norm is at most β_T on a convex neighborhood containing $W - \eta P$, $W - \eta\widehat{T}_0$, and $W - \eta\widehat{T}_1$. Set $G_P = \nabla L(W - \eta P)$ and $E_i = \widehat{T}_i - P$; then*

$$B_L = \eta\left\langle G_P, \delta\widehat{T}\right\rangle_F + R_{T,L}, \qquad |R_{T,L}| \le \frac{\beta_T\eta^2}{2}(\|E_0\|_F + \|E_1\|_F)\left\|\delta\widehat{T}\right\|_F. \tag{16}$$

The positive sign follows from subtracting the approximate-polar loss difference in Equation (15). The gradient is at the polar endpoint, rather than the spectrum-preserving endpoint in Equation (8). A shared forward error can cancel from a paired response, while a small forward error alone gives no relative guarantee for a sufficiently small intervention. Appendix B gives a counterexample and a combined sufficient condition for preserving a predicted sign.

6 EXPERIMENTS

We test the orientation, finite-step prediction, and numerical-fidelity claims at fixed model states.

6.1 DESIGN, COMPARISON UNITS, AND METRICS

We study spectral insufficiency on controlled quadratic families, and finite-step prediction and numerical realization on frozen states of a 124-million-parameter autoregressive Transformer. We train the language model with Muon to predict the next token on FineWeb-Edu, an educational-text subset of FineWeb (Penedo et al., 2024). We measure probe losses on the separate Pile test corpus (Gao et al., 2020). The first-layer panel contains three training seeds (903–905), three snapshot steps (50, 300, 1000), and two layer-zero matrices: the feed-forward expansion and attention value projection. The three independent training trajectories yield 18 state–block combinations; a depth replication adds 12 combinations at layers 6 and 11, indexed from zero. Calibration constructs positive and negative projected-curvature directions, a random feasible direction, and a nonzero direction with zero tangent curvature. We match their comparator-chord norms and correct their calibration alignment. The first three directions form the frozen 54-point primary panel; we report the 18 zero-tangent controls separately. The primary setting uses 512 calibration sequences, relative chord $\|\Delta\|_F/\sqrt{r} = 0.0003$, and half the implementation-matched reference step at each snapshot.

Each response is a paired difference of mean cross-entropy losses, measured in nats per token, with lower loss indicating better next-token prediction. We average paired sequence responses within each state and direction, then compute errors across these means. For targets y_i and compared values z_i, we define mean absolute error (MAE), mean squared error (MSE), and normalized squared error (NSE) as

$$\mathrm{MAE} = \frac{1}{n}\sum_i |z_i - y_i|, \qquad \mathrm{MSE} = \frac{1}{n}\sum_i (z_i - y_i)^2, \qquad \mathrm{NSE} = \frac{\sum_i (z_i - y_i)^2}{\sum_i y_i^2}. \tag{17}$$

All three are lower-is-better; for a nonzero target vector, NSE below one beats a prediction of zero at every point. Pearson correlation measures centered co-variation, sign agreement measures the fraction of matching signs, and the through-origin slope checks amplitude calibration; all have ideal value one. For prediction we report slope $\sum_i p_i Y_i / \sum_i p_i^2$; for realization we use $\sum_i Y_i \widehat{Y}_i / \sum_i Y_i^2$, making the independent variable explicit. We report pooled results for the frozen panel and results within each independent trajectory. Direction counts are descriptive; the three training trajectories are the replication units.

Orientation controls. As a broader control on Theorem 1, quadratic sweeps exhibit opposite update preferences in 1,251 of 4,096 eligible cases (30.54%); flat-spectrum and isotropic-curvature controls follow the predicted null behavior. These sweeps co-rotate the candidate and gradient while holding curvature fixed, complementing the fixed-gradient construction in Theorem 1. Appendix C.3 reports the sweep design and earlier language-model controls.

6.2 ENDPOINT PREDICTION ON FRESH CONFIRMATION TEXT

We next compare the frozen origin-curvature tangent prediction p_0, given by $\eta^2 t \left\langle H_S(W)[C], \dot{C} \right\rangle_F$, with the endpoint prediction p_S from Equation (8) on the actual finite chord and fresh confirmation text. We committed all endpoint predictions before evaluation on 256 fresh sequences per snapshot, with every sequence excluded from prior selection, calibration, and evaluation roles. On this first-layer panel, the endpoint predictor reduces pooled MSE from 6.6125×10^{-15} to 3.5959×10^{-16}, a 94.56% reduction relative to p_0, with lower MSE in each trajectory. To test a simpler amplitude correction, we subsequently fit one global scalar or three stage-specific scalars on old evaluation responses, excluding every confirmation sequence index across all seeds. Leave-one-seed-out (LOSO) controls fit the coefficients on the other two trajectories; all controls use the same 54 confirmation points, without fitting to their outcomes (Appendix C.5).

Scalar controls recover most of the gain. A scalar rescaling already recovers most of the endpoint gain: stage-wise scaling reduces NSE to 0.035001, compared with 0.031111 for the endpoint

Table 1: Prediction on the same 54 confirmation points. Scalar controls are a supplementary analysis fitted on old responses; lower errors and slope closer to one are better. MAE is in 10^{-8} nats/token.

Prediction	NSE	MAE	Slope	Signs
Origin tangent p_0	0.572106	3.5446	0.573490	43/54
Origin, actual chord	0.574694	3.5629	0.572976	43/54
Global scalar αp_0	0.044034	1.4871	1.041444	43/54
Stage scalars $\alpha_t p_0$	0.035001	1.2402	1.039065	43/54
LOSO global scalar	0.081484	1.9290	1.142133	43/54
LOSO stage scalars	0.074408	1.6713	1.139491	43/54
Endpoint p_S	0.031111	1.2744	0.951500	44/54

predictor. The endpoint has 11.11% lower pooled MSE but 2.76% higher MAE, and lower MSE in one of three trajectories and 8/18 state–block combinations. Relative to the original predictor, it improves absolute error in 31/54 points and state-wise MSE in 10/18 combinations. These controls identify amplitude calibration as the dominant predictive effect; the endpoint identity explains its connection to the update path. On the old 1,536-sequence set, the same-objective endpoint diagnostic reduces residual MAE from 2.45×10^{-8} for the origin-gradient-plus-curvature chord expansion to 7.43×10^{-12}, a 99.97% reduction. The calibration-to-evaluation endpoint gap remains 1.37×10^{-8} in mean absolute magnitude. This calculation uses the evaluation gradient, whereas the predictions in Table 1 do not.

6.3 A PRECISION INTERVENTION RESTORES THE MAIN RESPONSE

We next test whether the implemented polar routine preserves the ideal response by fixing the states, candidates, directions, step, five-round polynomial, and first 256 old evaluation sequences while changing only the numerical path. BF16, FP32, and double precision (FP64) receive the same inputs rounded first to FP32, while a full-FP64 control removes that initial rounding. A spectral reference applies the same five scalar polynomial rounds to the stored singular values and reconstructs the rotated output in FP64. All model, gradient, and loss computations use FP64. We evaluate both raw, unrescaled outputs and outputs matched to norm $\sqrt{r}$.

Table 2: Realization error relative to the ideal response Y on the existing 256-sequence precision subset, pooled over 54 primary points. All model losses are measured in FP64; the rows vary the matrix computation.

Length	Computation	NSE	MAE	Corr.	Signs
Raw	BF16	27.97636	4.183×10^{-7}	-0.03286	22/54
	FP32	0.008739	3.848×10^{-9}	0.99574	52/54
	FP64, same input	0.008733	3.849×10^{-9}	0.99574	52/54
	FP64, full	0.008733	3.849×10^{-9}	0.99574	52/54
	Spectral five-round	0.008733	3.849×10^{-9}	0.99574	52/54
Matched	BF16	36.17684	4.683×10^{-7}	-0.07389	23/54
	FP32	0.013474	4.542×10^{-9}	0.99341	52/54
	FP64, same input	0.013464	4.544×10^{-9}	0.99341	51/54
	FP64, full	0.013465	4.544×10^{-9}	0.99341	51/54
	Spectral five-round	0.013465	4.544×10^{-9}	0.99341	51/54

For raw outputs, FP32 reduces MSE from 3.4064×10^{-13} to 1.0641×10^{-16} (99.9688%) and MAE by 99.0800% relative to BF16. Every primary point has lower absolute error and every state has lower MSE, for both output-length conventions. Across seeds 903, 904, and 905, raw-output NSE falls from 6.3948, 178.7400, and 742.6795 to 0.0080, 0.0097, and 0.0527, respectively. Unlike the endpoint-predictor improvement, the numerical improvement occurs across all three measured training stages.

At the matrix level, the median residual $\left\|\delta\widehat{T} - \delta T_5^*\right\|_F / \|\Delta\|_F$ falls from 58.6803 to 0.005172, and the median cosine with the finite-round reference rises from 0.003710 to 0.999663. Both matrix

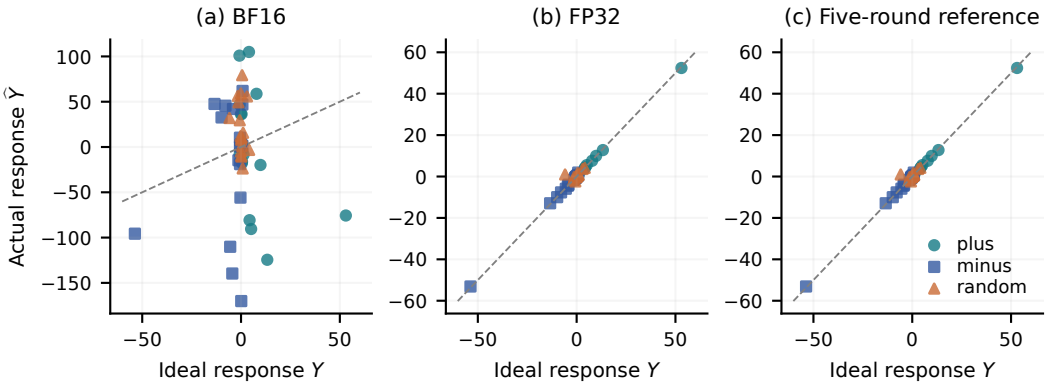

Figure 1: The same 54 first-layer interventions under three numerical paths, using raw output length. Both axes are in 10^{-8} nats/token. The dashed line is exact agreement with the ideal response; the BF16 panel needs a much wider vertical range. FP32 closely matches the five-round spectral reference, including its remaining random-direction discrepancies.

and loss measurements show recovery of the paired intervention. Meanwhile the median natural-output distance $\left\|\widehat{T}_0 - P\right\|_F / \|P\|_F$ changes only from 28.7404% in BF16 to 28.6448% in full FP64, illustrating the distinction between pointwise closeness and paired-response fidelity.

6.4 The residual bias belongs to the finite-round map

To separate arithmetic error from finite-round bias, we compare raw FP32 with the spectral reference: their response MAE is 3.66×10^{-11}. This difference is 0.95% of the 3.85×10^{-9} MAE between the FP32 and ideal responses. Full FP64 and the spectral reference agree to MAE 3.95×10^{-17}, directly reproducing the remaining finite-round discrepancy. The polar-endpoint projection in Equation (16) leaves FP32 remainder MAE 1.38×10^{-11}, about 0.36% of the total realization-bias MAE. Both percentages compare mean absolute magnitudes on the same evaluation data.

Table 3: Direction-specific realization: raw FP32 output, with absolute signal and bias reported alongside normalized error.

Direction (count)	mean$\lvert Y\rvert$	mean$\lvert B\rvert$	NSE	Slope	Signs
Constructed (36)	5.777×10^{-8}	2.313×10^{-9}	0.000710	0.98868	36/36
Random (18)	1.150×10^{-8}	6.917×10^{-9}	0.782230	0.44084	16/18
Zero-tangent (18)	6.709×10^{-9}	1.246×10^{-8}	9.517659	0.75946	12/18

Constructed directions retain both amplitude and sign, whereas the random and zero-tangent controls expose a substantial finite-round contribution relative to their weaker ideal signals. For seed 903, step 50, and the value-projection random direction, the ideal response is -5.8479×10^{-8}, but both full FP64 and the spectral reference give approximately $+9.9355 \times 10^{-9}$. The finite-round bias of $+6.8414 \times 10^{-8}$ exceeds and opposes the ideal signal, so higher arithmetic precision alone cannot repair this reversal. For $Y \neq 0$, the sign is preserved exactly when $-B/Y < 1$, and $|B| < |Y|$ is a stronger sufficient condition. In the spectral reference, this stronger condition holds in 30/36 constructed directions, 13/18 random directions, and 10/18 zero-tangent controls, while the corresponding sign counts are 36/36, 16/18, and 12/18. Here the condition diagnoses the observed signal and bias on the precision subset.

6.5 Depth replication separates geometry from data transfer

We then repeat the intervention at attention value projections in layers 6 and 11 to test whether the same separation persists deeper in the network. We retain the 512 calibration sequences, reuse the

256 previously evaluated confirmation sequences at each snapshot, and freeze all new directions and predictions before measuring losses. This design tests new blocks on existing text. At the fixed relative chord of 0.0003, all 24 constructed directions are available; the solver reaches the target for six of twelve random directions, all at step 1000. We report the six unavailable early random directions without replacement (Appendix C.6).

Table 4: Depth replication on 30 available directions. NSE compares each value with the ideal response on the same evaluation set; lower is better. p_X uses evaluation gradients and is a diagnostic, while p_0 and p_S use calibration data. Signs report raw FP32 agreement with the ideal response on all available directions.

Layer (count)	Origin p_0	Endpoint p_S	Diagnostic p_X	FP32	FP32 signs
6 (15)	1.218514	1.211861	5.78×10^{-10}	0.006430	14/15
11 (15)	1.180475	1.178111	8.82×10^{-9}	0.001751	15/15
Both (30)	1.207111	1.201744	3.05×10^{-9}	0.005027	29/30

Data transfer dominates prediction error. Unlike in the first layer, the endpoint predictor lowers pooled MSE by only 0.44% relative to the origin predictor, and both have NSE above one. By contrast, the same-objective endpoint diagnostic has residual MAE 2.09×10^{-13}, versus a mean absolute calibration-to-evaluation gap of 4.62×10^{-9}. The local expansion remains accurate at depth, but calibration endpoint gradients transfer poorly to evaluation responses. This separation explains why the first-layer prediction gain does not recur in these blocks.

Paired numerical fidelity persists at depth. Numerical fidelity is more stable: raw FP32 outputs preserve the ideal sign in 24/24 constructed and 5/6 available random directions. The five-round spectral reference gives the same signs at every point; its response differs from FP32 by MAE 1.73×10^{-12}, or 0.56% of the FP32-to-ideal MAE of 3.08×10^{-10}. The constructed directions have NSE 0.004389, while the weaker random responses have NSE 2.963594 despite five matching signs. These measurements reproduce the distinction between accurate arithmetic and finite-round distortion of a small ideal signal. The mean absolute ideal response falls from 5.55×10^{-9} in constructed directions to 1.78×10^{-10} in random directions, whereas the corresponding bias falls only from 3.32×10^{-10} to 2.11×10^{-10}. A weaker ideal signal can therefore suffer greater relative distortion despite a smaller absolute bias. Appendix C.6 reports seed-wise errors, absolute signal sizes, and the sign reversal; Appendix C.3 reports matrix-only size and iteration sweeps and compute costs.

What the controls identify. Together, the scalar controls test whether response amplitude accounts for the first-layer prediction gain. The evaluation-gradient diagnostic then tests the endpoint expansion on its own objective, separating local approximation error from transfer across text samples. Finally, the spectral reconstruction holds the five-round approximation fixed while removing matrix-arithmetic error. These comparisons distinguish three possible interventions: recalibrating a predictor, changing its calibration data, and changing the numerical map used for the update. They also explain why a more accurate endpoint expansion and more accurate arithmetic need not remove the same discrepancy.

7 Conclusion

We asked what determines whether spectral flattening improves loss at a fixed model state and finite step size. Fixing the ideal polar endpoint assigns the response to the spectrum-preserving comparator and shows that spectrum and first-order alignment are insufficient. For small orientation changes, the endpoint gradient captures the traversed path. Scalar controls explain most first-layer prediction gain, whereas calibration-to-evaluation differences dominate at depth. Higher precision removes arithmetic error, but the finite-round map explains the remaining random-direction reversals.

Implementations should therefore be validated at the scale of the interpreted change, in both matrix space and loss. Once high precision matches the finite-round reference, further reducing ideal-response bias requires changing the approximation. Orientation, endpoint geometry, and paired numerical fidelity thus capture distinct aspects of finite-step Muon behavior.

AI USE STATEMENT

We used generative AI tools to polish the writing, support literature search and reference checks, check and improve the theoretical proofs, and assist with the design and implementation of experimental code and the analysis of synthetic experiments. We verified the theoretical and empirical claims through explicit derivations, numerical tests, saved-record checks, and primary-source verification. The authors take responsibility for the final claims and artifacts.

ETHICS STATEMENT

This work studies optimizer behavior using existing text corpora and frozen model states. Web-derived text can contain personal information and social biases; these properties remain relevant to the underlying language models. The reported measurements concern aggregate next-token losses and matrix responses, and the anonymous supplement contains experimental and analysis code with embedded configuration templates.

REPRODUCIBILITY STATEMENT

Appendices A and B provide complete proofs; Appendix C specifies the experimental protocol and result checks. Fresh-text and depth-replication predictions were frozen before their respective loss evaluations; the depth replication reuses previously evaluated text and reports unavailable directions. The precision comparison uses identical inputs and paired losses. The manuscript source includes the bibliography and vector figures. The separate anonymous code supplement contains training and intervention routines, configuration templates, analysis scripts, and numerical tests; it excludes datasets, model weights, and result files.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations*, 2026.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *16th International Workshop on Optimization for Machine Learning (OPT 2024), NeurIPS Workshop*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936. Association for Computational Linguistics, 2019.

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

Zhehang Du and Weijie Su. The Newton-Muon optimizer. arXiv:2604.01472, 2026.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027, 2020.

Evan S. Gawlik and Melvin Leok. Iterative computation of the Fréchet derivative of the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1354–1379, 2017.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018.

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

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, second edition, 2002.

Nicholas J. Higham and Theo Mary. Mixed precision algorithms in numerical linear algebra. *Acta Numerica*, 31:347–414, 2022.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Technical blog, 2024.

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

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

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

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

Yuji Nakatsukasa, Zhaojun Bai, and François Gygi. Optimizing Halley's iteration for computing the matrix polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2700–2720, 2010.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. Datasets and Benchmarks Track.

Tianqi Shen, Jinji Yang, Runze Shi, Jianhao Ma, Jiaye Teng, and Ziye Ma. Towards understanding the power and limits of the Muon optimizer: A river-valley perspective. arXiv:2606.21514, 2026.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective. arXiv:2606.04662, 2026.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106. Association for Computational Linguistics, 2017.

Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the geometry of Muon with curvature-aware preconditioning. arXiv:2603.09697, 2026.

A PROOFS AND GEOMETRIC CONDITIONS

This appendix supplies the proofs and boundary cases behind the geometric and finite-step results in the main text. We follow the order of the argument there and use the notation summarized below.

Table 5: Quantities in the paired comparison. A positive advantage favors the polar update; a positive response means that advantage increases between two candidates.

Symbol	Meaning
P, C	Ideal polar update and equal-length spectrum-preserving comparator.
C', Δ	Rotated comparator and its finite chord $\Delta = C' - C$.
$\mathcal{A}(C)$	Advantage $L(W - \eta C) - L(W - \eta P)$ for one candidate.
Y_L	Ideal response $\mathcal{A}(C') - \mathcal{A}(C)$; the polar endpoint is shared.
p_L	Linear prediction from the comparator endpoint, $-\eta \langle \nabla L(W - \eta C), \Delta \rangle_F$.
$\widehat{Y}_L$	Response when the numerical map supplies both polar-side updates.
B_L	Realization bias $\widehat{Y}_L - Y_L$, distinct from comparator response Y_L.
$\delta\widehat{T}$	Numerical polar-side motion $\widehat{T}(M_R) - \widehat{T}(M)$.
F_T	Curvature action projected onto alignment-neutral moving velocities.

A.1 UTILITY EXPANSION AND THE FIXED-OBJECTIVE REVERSAL

We first justify the length-matched loss expansion and then verify the explicit reversal on a single fixed objective. Taylor expansion along each of $-\eta C$ and $-\eta P$ gives Equation (3); under a ρ-Lipschitz Hessian, the individual third-order bounds are $\rho\eta^3 \|C\|_F^3 /6$ and $\rho\eta^3 \|P\|_F^3 /6$. The sum of these bounds gives the stated remainder because both update norms equal $\sqrt{r}$. For $H[X] = BXA$, cyclicity of trace gives $\langle USV^\top, H[USV^\top] \rangle_F = \operatorname{tr}[(U^\top BU)S(V^\top AV)S^\top]$, and subtraction at $S = D, I$ proves Equation (4).

Proof of Theorem 1. At $W = 0$, set

$$G = \operatorname{diag}(3, 1), \quad A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad D = \operatorname{diag}(7/5, 1/5), \quad L(X) = \langle G, X \rangle_F + \tfrac{1}{2} \operatorname{tr}(XAX^\top). \tag{18}$$

The Hessian $H[X] = XA$ has eigenvalues $1, 1, 3, 3$, so the objective is strictly convex. Let $R_\pm$ be plane rotations by $\pm\pi/8$ and take $M_\pm = C_\pm = R_\pm D R_\pm^\top$, with $P = I_2$. Both matrices have singular values $7/5, 1/5$, norm $\sqrt{2}$, and polar distortion $\|P - C_\pm\|_F^2 = 4/5$, with

$$\langle G, P \rangle_F = 4, \qquad \langle G, C_\pm \rangle_F = (16 + 3\sqrt{2})/5, \tag{19}$$

$$a_\pm = (4 - 3\sqrt{2})/5, \qquad \Delta q_\pm = \pm 24\sqrt{2}/25. \tag{20}$$

At the common step $\eta = 1/2$, the exact advantages and loss values are

$$\mathcal{A}_+ = \frac{2}{5} - \frac{9\sqrt{2}}{50} > 0, \qquad \mathcal{A}_- = \frac{2}{5} - \frac{21\sqrt{2}}{50} < 0, \tag{21}$$

$$L(-P/2) = -\frac{3}{2}, \qquad L(-C_\pm/2) = -\frac{11}{10} - \frac{(15 \mp 6)\sqrt{2}}{50} < 0. \tag{22}$$

Thus the loss, gradient, Hessian, spectrum, alignments, and polar endpoint are fixed across the pair, all updates descend, and the preferences differ. □

A.2 MOVING VELOCITIES AND EXACT ALIGNMENT CORRECTION

The explicit example proves that a reversal is possible, but it does not yet describe the local degrees of freedom available on a fixed-polar fiber. We now extend its rotation principle to smooth curves that preserve gradient alignment exactly. Equation (5) is a compact singular value decomposition with bases UR and VR, so its polar factor is $(UR)(VR)^\top = P$ and all stated norm invariants follow

by orthogonal invariance. For $R(t) = \exp(tK)$ with $K^\top = -K$, the derivative is $\dot{C} = U[K, D]V^\top$, where $[K, D] = KD - DK$. Every $B \in \mathcal{S}_D$ is realized by the skew-symmetric generator

$$K_{ij} = \begin{cases} B_{ij}/(d_j - d_i), & d_i \neq d_j, \\ 0, & d_i = d_j. \end{cases} \tag{23}$$

Consequently, constant gradient alignment has tangent constraint $\langle E_D, B\rangle_F = 0$, while the derivative of $\langle C, H[C]\rangle_F$ is $2\langle F, B\rangle_F$. On unit Frobenius velocities, the largest alignment-neutral curvature derivative is $2\|F_T\|_F$, attained by $B = F_T/\|F_T\|_F$ when $F_T \neq 0$. We next construct an exact alignment-preserving curve for any $B \in \mathcal{T}_D$ when $E_D \neq 0$. Choose K from Equation (23), and choose a skew-symmetric N such that $[N, D] = E_D$ by the same formula. Define $R(t, u) = \exp(tK)\exp(uN)$ and $h(t, u) = \langle G, UR(t, u)DR(t, u)^\top V^\top - C\rangle_F$. At $(t, u) = (0, 0)$, we have $h = 0$, $\partial_t h = \langle E_D, B\rangle_F = 0$, and $\partial_u h = \|E_D\|_F^2 > 0$. The implicit function theorem gives a smooth $u(t)$ with $u(0) = u'(0) = 0$ and $h(t, u(t)) = 0$ near zero. Thus $C(t) = UR(t, u(t))DR(t, u(t))^\top V^\top$ preserves alignment exactly and has $\dot{C}(0) = UBV^\top$. This construction makes the local level-set argument explicit in standard matrix-manifold coordinates (Absil et al., 2008). For multiplicities n_a of the distinct diagonal values of D, the moving dimension is $s = \sum_{a<b} n_a n_b$, and the regular alignment constraint leaves dimension $s - 1$. A nonzero tangent orthogonal to both alignment and a nonzero projected curvature response therefore requires $s \geq 3$.

Proof of Theorem 2. Consider the linear map $\Phi(H) = \Pi_{\mathcal{T}_D}\,\mathrm{sym}(U^\top H[C]V)$. For any $B \in \mathcal{T}_D$, write $Y = UBV^\top$, $c^2 = \|C\|_F^2$, and define

$$H_Y[X] = \frac{\langle C, X\rangle_F\, Y + \langle Y, X\rangle_F\, C}{c^2} - \frac{\langle C, Y\rangle_F\, \langle C, X\rangle_F\, C}{c^4}. \tag{24}$$

This operator is self-adjoint and satisfies $H_Y[C] = Y$, hence $\Phi(H_Y) = B$ and Φ is surjective. Rank–nullity gives codimension $s-1$ for its kernel, which is closed, nowhere dense, and of Lebesgue measure zero. The complement remains relatively open, dense, and of full measure within the positive-definite cone because that cone is open in the ambient operator space. Choose the unit velocity $B = F_T/\|F_T\|_F$ and integrate it on the exact alignment level set as above. On a quadratic objective the linear term is constant along the curve, so differentiation gives $\dot{\mathcal{A}} = \eta^2\langle F_T, B\rangle_F = \eta^2\|F_T\|_F$. For a general loss, expand the endpoint gradient as $\nabla L(W - \eta C) = G - \eta H[C] + e$, where $\|e\|_F \leq \rho\eta^2\|C\|_F^2/2$. The inner product with $-\eta\dot{C}(0)$ proves the discrepancy bound $\rho\eta^3 r/2$, since $\left\langle G, \dot{C}(0)\right\rangle_F = 0$ and $\left\|\dot{C}(0)\right\|_F = 1$. □

The regularity condition is essential: for a distinct-entry diagonal D and $E = D$, the alignment $\mathrm{tr}(DRDR^\top)$ has a strict maximum at the orbit point D, although every first derivative vanishes there. Indeed, $\mathrm{tr}(D^2) - \mathrm{tr}(DRDR^\top) = \frac{1}{2}\left\|RDR^\top - D\right\|_F^2$, so equality forces the comparator to remain unchanged. The equality level therefore contains no nearby nonconstant moving curve, despite the vanishing linearized constraint. At $s = 1$, a regular alignment constraint leaves a zero-dimensional tangent space. For a pure-gradient candidate $M = G$, E is diagonal and $E_D = 0$; the regular theorem applies to candidates satisfying $E_D \neq 0$, as can occur with momentum.

A.3 Proof of the endpoint and path identities

Having characterized the admissible perturbations, we next connect one such perturbation to its finite-step loss response.

Proof of Proposition 3. The chain rule applied to $g(s) = \nabla L(W - s\eta C)$ gives $g'(s) = -\eta H_L(W - s\eta C)[C]$. The integral over $[0, 1]$, paired with $-\eta\Delta$, proves Equation (9). The integral Taylor formula at $W - \eta C$ along the chord $-\eta\Delta$ gives

$$R_{C,L} = \eta^2 \int_0^1 (1 - t)\,\langle \Delta, H_L(W - \eta C - t\eta\Delta)[\Delta]\rangle_F\, dt. \tag{25}$$

The Hessian bound and $\int_0^1 (1 - t)dt = 1/2$ prove Equation (10). Finally, the integral of $\|H_L(W - s\eta C) - H_L(W)\|_{\mathrm{op}} \leq \rho s\eta\|C\|_F$ gives the origin-to-path approximation bound. □

For a nominal tangent displacement $t\dot{C}$, the endpoint correction to the origin tangent predictor separates exactly into

$$\begin{aligned} p_L - \eta^2 \left\langle H_L(W)[C], t\dot{C} \right\rangle_F = & -\eta \left\langle G_L(W), \Delta \right\rangle_F \\ & + \eta^2 \left\langle (\overline{H}_{L,C} - H_L(W))[C], \Delta \right\rangle_F \\ & + \eta^2 \left\langle H_L(W)[C], \Delta - t\dot{C} \right\rangle_F , \end{aligned} \tag{26}$$

where $\overline{H}_{L,C} = \int_0^1 H_L(W - s\eta C) ds$, and the three terms are origin-gradient residual, path-curvature change, and tangent–chord discrepancy.

B Numerical response proofs and a preservation condition

We now complete the numerical part of the argument by connecting finite-round matrix motion to loss bias and, ultimately, to sign preservation. The first two proofs establish the matrix- and loss-level identities; the corollary then combines them with the endpoint error decomposition.

Proof of Proposition 4. Write the real-arithmetic recurrence as $X_0 = M/(\|M\|_F + \varepsilon)$ and

$$X_{j+1} = 3.4445 X_j - 4.775 (X_j X_j^\top) X_j + 2.0315 (X_j X_j^\top)^2 X_j. \tag{27}$$

If $X_j = U R Z_j R^\top V^\top$, orthonormality of U, V reduces each multiplication to the corresponding scalar polynomial on Z_j. The initial Frobenius norm is invariant under rotation, so induction proves Equation (12). The transpose convention for tall inputs yields the same recurrence because $(XX^\top)^\ell X = X(X^\top X)^\ell$. For any real diagonal $Z = \mathrm{diag}(z_i)$ and orthogonal R,

$$\left\| RZR^\top - Z \right\|_F^2 = 2\,\mathrm{tr}(Z^2) - 2\,\mathrm{tr}(ZRZR^\top) = \sum_{i,j} R_{ij}^2 (z_i - z_j)^2. \tag{28}$$

Apply this identity to Z_k and D, observing that $d_i = d_j$ implies $z_i^{(k)} = z_j^{(k)}$, and bound each squared gap by $\gamma_k^2 (d_i - d_j)^2$. The isometries U, V preserve the Frobenius norm and complete the proof. □

Proof of Proposition 5. Let $E(t) = (1-t)E_0 + tE_1$ and integrate the gradient along the segment joining the two actual update endpoints:

$$B_L = \eta \int_0^1 \left\langle \nabla L(W - \eta(P + E(t))), \delta\widehat{T} \right\rangle_F dt. \tag{29}$$

The Hessian bound on the convex neighborhood gives $\|\nabla L(W - \eta(P + E(t))) - G_P\|_F \le \beta_T \eta \|E(t)\|_F$. Subtract the constant-gradient term, use Cauchy–Schwarz, and integrate $\|E(t)\|_F \le (1-t)\|E_0\|_F + t\|E_1\|_F$ to obtain Equation (16). □

An alternative expansion at the actual baseline endpoint has remainder bound $\beta_T \eta^2 \left\|\delta\widehat{T}\right\|_F^2 / 2$, but uses $\nabla L(W - \eta \widehat{T}_0)$ instead of G_P.

Corollary 6 (A sufficient sign-preservation condition)**.** *Under the evaluation-loss assumptions of Propositions 3 and 5, use the finite-round map and γ_k from Proposition 4 with the chosen output-length convention. Assume the calibration loss is differentiable at $W - \eta C$, and all numerical maps in Equation* (14) *are defined. Set $q_C = \|\Delta\|_F$, $b_{\mathrm{num}} = \|\delta E_{\mathrm{in}}\|_F + \|\delta E_{\mathrm{arith}}\|_F$, and*

$$A_X = \eta \|\nabla L_X(W - \eta P)\|_F + \frac{\beta_T \eta^2}{2} (\|E_0\|_F + \|E_1\|_F). \tag{30}$$

Then the total prediction error satisfies

$$|\widehat{Y}_X - p_S| \le |p_X - p_S| + \frac{\beta_C \eta^2}{2} q_C^2 + A_X(\gamma_k q_C + b_{\mathrm{num}}). \tag{31}$$

If $p_S \neq 0$ and the right side is less than $|p_S|$, the actual response has the predicted sign.

Proof. Split $\widehat{Y}_X - p_S = (Y_X - p_X) + (p_X - p_S) + B_X$, apply the two loss bounds, and use $\left\|\delta\widehat{T}\right\|_F \le \gamma_k q_C + b_{\mathrm{num}}$ from Equation (14). An error interval with radius strictly smaller than the nonzero prediction magnitude cannot cross zero. □

B.1 Why forward accuracy is a different requirement

The preservation bound depends on paired output motion, not only on the distance from each numerical output to the ideal polar factor. A smooth counterexample shows why uniform forward closeness alone cannot control relative fidelity for a shrinking intervention. Take $D = \sqrt{2/5}\,\mathrm{diag}(2,1)$, $C(t) = R(t)DR(t)^\top$, $P = I_2$, and $B = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$. For $0 < \epsilon < 1/2$ and $|t| < \pi/4$, the curve is injective, so $\widetilde{T}_\epsilon(C(t)) = P + \epsilon\sin(t/\epsilon^2)B$ defines a smooth map on this local fiber. Its forward error is uniformly at most ϵ, but its derivative at zero has norm $1/\epsilon$. Set $L(X) = \langle B, X\rangle_F$, $W = 0$, $\eta = 1$, and $h = \pi\epsilon^2/2$. The approximate-polar endpoint moves by ϵB and induces realization bias ϵ; the ideal response is $-\sqrt{2}(d_1 - d_2)\sin h\cos h = O(\epsilon^2)$. The relative error therefore diverges despite a zero Hessian and vanishing forward error. This smooth construction establishes the mathematical separation between forward accuracy and relative response fidelity; the measured BF16 behavior is evaluated separately in Section 6.

For comparison, the classical cubic recurrence $z_{j+1} = 1.5z_j - 0.5z_j^3$ converges monotonically to one from $0 < z_0 < 1$, since $1 - z_{j+1} = \frac{1}{2}(1 - z_j)^2(2 + z_j)$. The scalar-only check on the 18 saved spectra reaches maximum singular-value error 4.20×10^{-13} within 29–41 rounds at a declared 10^{-12} tolerance. This check evaluates scalar spectral convergence; model-loss evaluation in this study uses the five-round quintic.

C Experimental details and evidence provenance

This appendix provides the implementation details and evidence trail behind the experimental comparisons in Section 6. We first describe the shared setup, then turn to supplementary controls, audit procedures, and the depth replication.

C.1 Training and intervention construction

Every intervention begins from a saved training state, which keeps the model and optimizer history fixed across the paired comparison. We first describe these states and then explain how the alignment-preserving rotations are constructed from them. Training and probe examples are drawn

Table 6: Frozen language-model configuration.

Item	Value
Layers / width / heads / feed-forward width	12 / 768 / 12 / 3072
Context length / total steps	256 / 1000
Tokens per step	8192
Momentum / Nesterov / polynomial rounds	0.95 / enabled / 5
Peak Muon / auxiliary AdamW learning rates	0.02 / 0.0003

from fixed caches, with 256 prediction targets per sequence. We save each snapshot immediately before the labeled optimizer update and retain its actual Nesterov input candidate. For a block of shape $m \times n$, the reference coefficient is $\eta_{\mathrm{ref}} = \mathrm{LR}\sqrt{\max(1, m/n)}$, and the primary probe uses $\eta = \eta_{\mathrm{ref}}/2$. These probes change one stored candidate block with other parameters fixed and omit a separate weight-decay displacement from the paired comparison.

The fiber construction uses eight predetermined singular coordinates, projects curvature into the alignment-neutral moving space, and chooses positive, negative, random, and zero-curvature unit-velocity directions there. A finite alignment correction and scalar chord solve convert each tangent into a rotation with the prescribed common chord. The coordinates are the unique rounded values of eight equally spaced indices from zero to $r - 1$ in the saved singular-value ordering. The solver uses

$R(t) = \exp(tK)\exp(s(t)N)$, where K is the selected tangent generator and the scalar $s(t)$ corrects alignment along the fixed normal generator N. The t in p_0 is the positive curve parameter returned by the final chord-matching solve. We select the largest common feasible relative chord from the frozen list $\{0.0003, 0.0001, 0.00003\}$ across all states, calibration sizes, and four directions. This selection uses calibration information only. The construction uses fixed angle, alignment, and chord tolerances and records failed directions rather than replacing them after loss evaluation. All first-layer state–direction points are available at the primary size, while some points are unavailable in the larger-size scan.

C.2 Data roles and computational controls

Once the states and directions are fixed, the remaining design choices concern which text is used at each stage and how the numerical paths are controlled. We keep these roles separate so that the paired comparisons can be reproduced and their possible sources of leakage can be identified. The fresh confirmation allocator excludes all previous role assignments and allocates the three snapshots jointly within each seed, producing disjoint 256-sequence snapshot sets within that seed. Texts can overlap across seeds, and we apply no additional deduplication between the separate training and probe corpora. We committed predictions before the first confirmation evaluation. The primary confirmation criterion is a frozen descriptive comparison: lower endpoint-predictor MSE than origin-predictor MSE in each training trajectory. Sign scoring retains the common set where the larger absolute value of the two frozen predictions exceeds 10^{-12}, which includes all 54 primary points.

The realization data provide 256 paired sequence responses per state and direction. All numerical paths store their outputs in 64-bit containers; this storage conversion preserves the lower-precision results without adding significant digits. TensorFloat-32 is disabled, FP32 matrix multiplication uses its highest-precision setting, and BF16 reduced-precision reduction remains enabled to reproduce the original path. Stored losses reproduce, parameter-restoration checks pass, and repeat discrepancies remain below the declared numerical tolerances.

C.3 Additional results and historical controls

With the common setup in place, we now examine how the main comparisons depend on intervention size, iteration count, and several earlier experimental settings. We begin with broader controls for

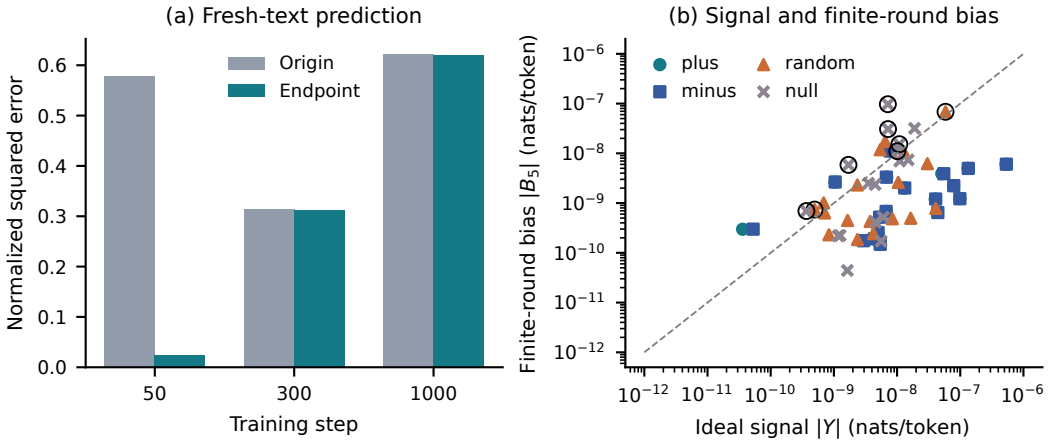

Figure 2: First-layer path prediction and finite-round bias. (a) Fresh-text prediction errors by training step, comparing the unscaled origin predictor with the endpoint predictor. Table 8 adds scalar controls. (b) On the existing 256-sequence precision subset, absolute ideal signal versus five-round spectral-reference bias; black rings mark sign reversals. Points below the diagonal satisfy $|B_5| < |Y|$, a sufficient sign-preservation condition. "Plus/minus", "random", and "null" denote the constructed, random, and zero-tangent directions.

the orientation claim. The quadratic study spans dimensions 12 and 16, three update-spectrum

types, four curvature geometries, and left, right, or joint rotations, yielding 8,064 seven-angle sweeps over eight seeds. After algebraic null exclusions, exact advantages take both signs in 1,251 of 4,096 eligible sweeps (30.54%; family–seed bootstrap interval 25.46–35.35%). The reversal rate rises from 0/1,024 at a quarter of the base step to 625/1,024 at four times that step, while flat-spectrum and isotropic-curvature controls behave as predicted. These sweeps co-rotate candidate and gradient while holding curvature fixed; Theorem 1 supplies the same-gradient, same-objective construction. Earlier held-out left-basis interventions on the same model family found continuous utility effects in 6/12 early target states and passed the frozen origin-based second-order bridge in 3/12. A natural-direction study confirmed finite-loss crossovers in 11/11 eligible critical states from two new trajectories. A crossover for one update pair and a much smaller response between orientations have different error margins.

The next controls ask whether numerical fidelity changes with the size of the intervention or the number of polynomial rounds. The shared-size panel differs from the complete 54-point original-

Table 7: Raw-output numerical residual medians, normalized by comparator-chord norm; lower is better.

Relative size	BF16	FP32	Shared points
0.25	168.5595	0.014086	47
0.5	94.6649	0.007056	47
1	58.6934	0.003525	47
2	34.8829	0.001762	47
4	20.9071	0.000880	47

size panel, which explains the different original-size medians in this table and the main text. A matrix-only sweep varies the relative chord from 0.0075% to 0.12%, using the same 47 primary directions available at all five sizes. Across that range, the BF16 numerical-residual ratio decreases from 168.56 to 20.91, but its median cosine remains below 0.007; FP32 cosines range from 0.99338 to 0.99997. At the original size, increasing from five to ten rounds raises the BF16 residual ratio from 58.68 to 165.26 and the FP32 ratio from 0.00517 to 0.01907. Since $q(1) = 0.701 \neq 1$, unit singular values are not a fixed point of this quintic: additional rounds change the approximation rather than monotonically refine it. These size and iteration sweeps measure matrix responses; model-loss measurements use the primary size and five rounds. The full 54-point original-size numerical-residual medians for same-input FP64 and full FP64 are 0.000687 and 2.33×10^{-11}, respectively. For matched outputs, the spectral reference agrees in sign on 51/54 points, compared with 52/54 for FP32. The extra FP32 match involves a random response close to zero: the ideal value is -7.14×10^{-10} and the full-FP64 value is $+5.59 \times 10^{-12}$. The extra FP32 sign match occurs at this near-zero response, while full FP64 remains closer to the finite-round spectral reference.

A final group of historical checks concerns the accuracy of the loss expansion and the criteria used in earlier intervention studies. The earlier nonlinear accounting control used small two-layer Transformers on BoolQ text (Clark et al., 2019), repurposed for byte-level next-token prediction. At relative parameter radii at most 0.01, the directly measured full-Hessian quadratic expansion achieved $R^2 > 0.999999$ over 1,440 utility measurements, where $R^2 = 1 - \sum_i (Y_i - p_i)^2 / \sum_i (Y_i - \overline{Y})^2$. This coefficient compares fixed Taylor coefficients on the same loss, whereas held-out structured-curvature regressions failed the frozen improvement thresholds. A follow-up using SciQ text (Welbl et al., 2017) separated Kronecker, generalized Gauss–Newton, and full-Hessian contributions, but left the prespecified attribution classification unresolved. These historical controls establish the distinction between a valid loss expansion and a sufficiently accurate data-transfer or curvature proxy. Those historical studies used fixed magnitude, sign-consistency, overlap-validation, and crossover criteria established before the reported counts. These were detection rules rather than familywise significance tests.

C.4 Reproducibility and audit trail

Because the reported responses are small paired differences, their provenance matters alongside their aggregate values. We therefore trace each result to the saved files and code that produced it. We checked every manifest-listed file hash and runtime-source binding against the frozen experiment

releases. The collections preserve configurations, per-sequence results, logs, environment information, and summaries. We recomputed all reported summaries and the signal–bias analysis from these raw records. This post-run audit checks the saved evidence; rerunning model forward passes additionally requires the model weights and text caches retained with the experiments.

C.5 SCALAR RECALIBRATION AND ACTUAL-CHORD CONTROLS

The main text shows that the endpoint predictor improves on the original tangent predictor in the first layer. This supplementary analysis asks how much of that improvement can be reproduced by a scalar amplitude correction or by replacing the nominal tangent with the actual chord. We specified the controls after inspecting the original confirmation results, fitted their coefficients using old responses, and evaluated them on the unchanged 54-point confirmation panel. The analysis adds neither training nor model forward passes.

The fit target for each state and direction is its mean response on the old 1,536-sequence evaluation bank. Before forming those means, we remove all confirmation sequence indices across the three seeds and nine snapshots, preventing them from contributing to scalar-fit responses even across seeds. The original direction-selection data and frozen predictors remain unchanged; the procedure does not establish complete text independence across trajectories or detect duplicate content under different indices. For fitting points i with origin predictions p_{0i} and old-data mean responses y_i^{old}, we use unweighted through-origin least squares,

$$\widehat{\alpha} = \frac{\sum_i p_{0i} y_i^{\mathrm{old}}}{\sum_i p_{0i}^2}. \tag{32}$$

Each state–direction mean receives equal weight, matching the main prediction metric. We apply no clipping, intercept, outcome-based direction filtering, or confirmation-set parameter selection.

The global fit uses one coefficient for all 54 points; stage-wise fits use one coefficient for each of steps 50, 300, and 1000. For each held-out training seed, LOSO fits use only old responses from the other two seeds, either jointly or by stage. We store all coefficients before the analysis script reads confirmation losses. This execution order prevents coefficient leakage, while the supplementary status reflects that the original confirmation results were already known.

We also replace the nominal tangent in p_0 with the actual finite comparator chord:

$$p_{\mathrm{chord}} = -\eta \langle G_S(W), \Delta \rangle_F + \eta^2 \langle H_S(W)[C], \Delta \rangle_F . \tag{33}$$

Both derivatives come from the original calibration records, so this control uses neither endpoint gradients nor confirmation derivatives. Its NSE is 0.574694, compared with 0.572106 for the tangent predictor. Removing its first-order term changes NSE by only 1.50×10^{-9} because calibration alignment is already preserved. The finite-displacement correction therefore does not account for the observed amplitude shift at this intervention size.

Table 8: Supplementary prediction NSE by trajectory and training stage. Lower is better. The endpoint has the lowest pooled NSE, but its advantage over stage-wise scaling is not uniform.

Subset	Origin	Stage scalar	LOSO stage	Endpoint
Seed 903	0.542800	0.031762	0.084380	0.017963
Seed 904	0.459619	0.096460	0.094892	0.178310
Seed 905	0.691400	0.027612	0.038768	0.029593
Step 50	0.578400	0.031584	0.072645	0.023960
Step 300	0.313973	0.162981	0.132602	0.311647
Step 1000	0.621254	0.623028	0.773806	0.619466

Pooled improvement and mixed subgroup results. The pooled metrics favor the endpoint predictor, but the subgroup comparisons show that this advantage is not uniform. Compared with stage-wise scaling, the endpoint reduces pooled MSE from 4.0455×10^{-16} to 3.5959×10^{-16}, while MAE increases from 1.2402×10^{-8} to 1.2744×10^{-8}. It improves absolute error in 24/54 points and MSE in 8/18 states, with a lower trajectory-level MSE only for seed 903. Against LOSO

stage-wise scaling, it lowers pooled MSE by 58.19% and trajectory-level MSE in two of three seeds. The distinction matters: same-trajectory old data calibrate amplitudes effectively, while coefficients fitted on other trajectories transfer less reliably. These are descriptive comparisons over three training trajectories, without a statistical-significance claim.

As a fixed sensitivity analysis, we fit the same scalars on all 1,536 old sequences without the cross-seed exclusion. Global and stage-wise NSE become 0.046050 and 0.037211, respectively; LOSO global and stage-wise NSE become 0.101422 and 0.095156. The conclusion that scaling recovers most of the original gain is unchanged. The scalar methods require old measured responses and one or three fitted parameters; the endpoint method uses a calibration endpoint gradient without fitted response coefficients. Both comparisons concern frozen one-step response prediction, not end-to-end training efficiency.

C.6 Middle- and final-layer replication

The first-layer results motivate a direct check of whether the same geometric and numerical distinctions persist in later blocks. The depth replication therefore applies the same intervention rule to middle- and final-layer parameters while keeping the main design choices fixed.

Frozen design and geometry coverage. We first specify what is held fixed and which directions the constrained solver can realize at the target chord. The depth replication uses the attention value-projection matrices in layers 6 and 11, with zero-based indexing, at steps 50 and 1000 for seeds 903–905. These twelve new state–block combinations use the existing checkpoints, 512-sequence calibration sets, and 256-sequence confirmation sets of their respective snapshots. Calibration and evaluation indices are disjoint within each snapshot, but the evaluation text has already served the first-layer confirmation. We therefore treat this panel as a supplementary replication on new blocks, not a second independent text confirmation. The run uses the frozen relative chord 0.0003, half the reference step, eight predetermined singular coordinates, raw output lengths, and five polynomial rounds. Its two matrix paths are FP32 and the FP64 spectral reconstruction of the same finite-round map; all model parameters, gradients, and loss measurements use FP64. The replication measures one intervention size and does not include a BF16 arm.

The solver attempts positive, negative, and random directions in every state. It constructs all 24 positive/negative directions and six random directions, each from step 1000. For all six step-50 random directions, it reports that the target chord was not found in the connected alignment-preserving chart under the frozen angle budget. This outcome is a construction failure at the requested size, not a certificate of global infeasibility. The run neither substitutes directions nor lowers their target size. All 30 available directions receive the complete planned set of evaluation batches. Every loss record binds to the same all-state prediction commit.

Prediction and realization by trajectory. With the available directions fixed, we compare calibration-based prediction, the same-objective diagnostic, and numerical realization within each trajectory. Table 9 separates calibration predictions from evaluation-gradient diagnostics and numerical realization. The endpoint predictor has lower MSE than the origin predictor in two of three trajectories, with a pooled reduction of 0.44%. Replacing the origin tangent by the full actual-chord expansion gives NSE 1.207111, so this finite-displacement correction also leaves the prediction error essentially unchanged. In Equation (11), the mean absolute data-transfer term is 4.621840×10^{-9} and the mean absolute chord remainder is 2.090604×10^{-13}. These terms explain the error on the measured evaluation objective without estimating a population response.

Absolute signal and the remaining reversal. Aggregate normalized errors can obscure whether a discrepancy reflects a large bias or a weak ideal signal, so we examine their absolute magnitudes next. The mean absolute ideal response is 5.552×10^{-9} for constructed directions and 1.782×10^{-10} for the available random directions. The corresponding FP32 realization-bias magnitudes are 3.317×10^{-10} and 2.110×10^{-10}. The smaller random signals therefore have larger relative distortion even though their absolute bias is smaller. For seed 904, step 1000, and layer 6, the random ideal response is -1.6251×10^{-11}, while FP32 gives $+9.2400 \times 10^{-10}$ and the five-round spectral reference gives $+9.2375 \times 10^{-10}$. The finite-round map reproduces this reversal, and FP32 agrees with its sign on all 30 available directions. These signs describe the measured finite-text means;

Table 9: Depth-replication NSE by trajectory and stage. Every row compares values against its own ideal evaluation response; lower is better. p_X uses evaluation gradients and is a same-objective diagnostic.

Subset	Count	Origin p_0	Endpoint p_S	Diagnostic p_X	FP32	FP32 signs
Seed 903	10	0.550305	0.556490	1.52×10^{-8}	0.010205	10/10
Seed 904	10	1.380967	1.372608	1.48×10^{-9}	0.002180	9/10
Seed 905	10	1.078187	1.074909	9.94×10^{-10}	0.009991	10/10
Step 50	12	1.222205	1.216325	3.33×10^{-9}	0.004762	12/12
Step 1000	18	1.049804	1.049780	9.13×10^{-11}	0.007787	17/18

the ideal mean in this example is close to zero. Using the original frozen prediction-support rule, $\max(|p_0|, |p_S|) > 10^{-12}$, retains 28 of the 30 directions. On this support, the origin and endpoint predictors both match 10/28 signs, the diagnostic matches 28/28, and FP32 and the spectral reference both match 27/28. All squared-error and absolute-error metrics retain all 30 available directions, including the two below the sign-support threshold.

At the matrix level, the median FP32 residual relative to the comparator chord is 0.003328, and the median cosine with the finite-round reference is 0.999693. The polar-endpoint projection in Equation (16) leaves remainder MAE 1.45×10^{-13}, or 0.047% of the mean absolute FP32 realization bias. Thus both endpoint expansions remain accurate on their evaluation objective: one accounts for ideal comparator motion, and the other accounts for the numerical polar motion.

Integrity and precision checks. Finally, we verify the saved evidence and numerical controls of the replication. An independent read-only analysis verifies all file hashes, frozen runtime bindings, record seals, planned batch coverage, and the disjoint calibration/evaluation indices within each snapshot. It recomputes ideal and realized responses from per-sequence losses and reproduces every saved summary metric. The released summarizer also reproduces the stored summary exactly. Alignment, repeat, and parameter-write errors remain below the recorded tolerances. The collection retains raw responses, configurations, logs, environment records, and source code without duplicating model weights or text caches.