WHEN HISTORICAL ENERGY IS INSUFFICIENT FOR LOCAL TRAINING RESPONSE

Anonymous authors
Paper under double-blind review

ABSTRACT

A training trajectory can be nearly one-dimensional in historical energy, yet discarding its weak second direction can lose an order of local prediction accuracy. For smooth fixed-loss gradient descent over fixed finite windows, we characterize the necessary and sufficient gradient and curvature capture rates for preserving the unrestricted linearized model's $O(\eta^2)$ normalized parameter-error guarantee. With nonzero gradient and transverse curvature, even the best line through the anchor, fixed across the window and chosen using the complete future, has sharp $\Theta(\eta)$ error. Both the gradient–curvature plane and the exact noiseless historical top-two space attain $O(\eta^2)$ error. History therefore contains a sufficient response space, but every fixed explained-energy threshold below 100% eventually discards its necessary second direction, whose energy fraction vanishes quadratically. Multi-scale tests in selected LoRA coordinates of eight LLM model–data blocks support the predicted orders and assess the sharp best-line coefficient, with model-dependent finite-scale deviations. Across four families, matched experiments locate over 96% of the one-direction predictor's squared residual in a direction carrying under 2% of future update energy (family medians). A varying-driver motion law separates driver and curvature contributions to historical compression; shuffled-AdamW diagnostics examine direction selection beyond fixed-loss GD.

1 INTRODUCTION

Training trajectories are compressed to predict subsequent updates or to restrict future optimization to a small space (Wei et al., 2026; Chen et al., 2026; Li et al., 2023). A space can retain nearly all historical update energy yet discard a weak direction needed to preserve local prediction accuracy. We ask: *which directions must compression retain, and how accurately, to preserve a local training model's accuracy order?*

We study parameter displacements over a fixed number of gradient-descent steps on one smooth loss. The unrestricted linearized-gradient model has $O(\eta^3)$ absolute error, or $O(\eta^2)$ error relative to the $\Theta(\eta)$ displacement. Compression should preserve this guarantee: otherwise space truncation can dominate the error of the local approximation itself. Historical energy and response importance have different scales. Transverse curvature contributes quadratically small historical energy, but omitting it leaves a first-order normalized response error.

We give necessary and sufficient capture rates for the gradient and its first Hessian action. With nonzero gradient and transverse curvature, even the best line through the anchor, fixed across the window and selected using the complete future, has sharp $\Theta(\eta)$ error. Both the gradient–curvature plane and the exact noiseless historical top-two space preserve the $O(\eta^2)$ target. Yet every fixed energy threshold below 100% eventually selects one dimension (Figure 1). This is a loss of local accuracy order: the best-line error still vanishes, and fixed windows shrink in elapsed optimization time as η decreases.

Classical model reduction already distinguishes reconstruction energy from dynamical adequacy, and Krylov methods match operator response (Moore, 1981; Ilak & Rowley, 2008; Grimme, 1997). Our contribution is the quantitative link for local training response: necessary and sufficient capture rates, a sharp optimum over every admissible fixed line, and sufficiency of exact historical top-two spaces. Together, they identify the minimum dimension and explain why fixed energy thresholds discard sufficient information already present in history.

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

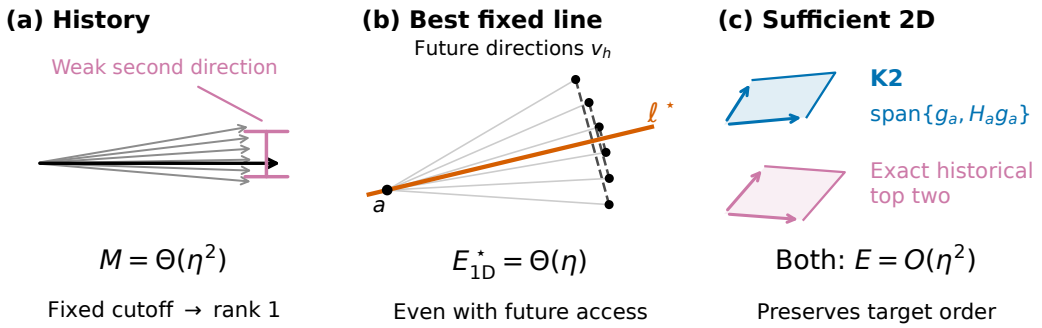


Figure 1: **Energy selects one direction; preserving local accuracy order requires two.** Schematic for a fixed-loss local GD window with nonzero gradient and transverse curvature. History contains a weak second direction, but any fixed explained-energy threshold below 100% eventually selects one. Even the best future-informed line through the common anchor loses one accuracy order; the gradient–curvature plane and exact historical top-two space preserve the $O(\eta^2)$ normalized-error target. Both constructions use the anchor gradient and local operator. The geometry depicts the leading local expansion, not an exactly planar nonlinear path.

Our results establish:

- **A sharp characterization of response accuracy.** Necessary and sufficient gradient and curvature capture rates specify adequate spaces. The best anchored fixed line loses one accuracy order even with future access; a two-dimensional space suffices.
- **Sufficient history, inadequate energy selection.** Exact noiseless historical top-two spaces preserve the target, but fixed energy thresholds discard the necessary second direction. The same curvature quantitatively links historical tail energy to omitted response.
- **Matched validation and scope.** Multi-scale LLM experiments test the orders and sharp constant; four-family controls localize the weak-direction residual. Varying-driver theory and shuffled-AdamW diagnostics delimit the mechanism.

2 Response Target and Compression Protocol

Let $\theta_t \in \mathbb{R}^d$ denote the selected trainable coordinates, updated by gradient descent with step size η. After P observed updates, all predictions start at the true anchor $a = \theta_P$ and forecast the next H steps. TRUE denotes nonlinear continuation on the same loss. Here, the response is the sequence of parameter displacements from the common anchor; compression restricts the space in which the local model evolves.

History and local response. For adjacent updates $u_t = \theta_t - \theta_{t-1}$, define $U = [u_1^\top; \ldots; u_P^\top]$ and

$$O_1 = \frac{\sigma_1(U)^2}{\|U\|_F^2}, \qquad M = 1 - O_1. \tag{1}$$

The tail-energy fraction M measures temporal increment compression; $\sqrt{M}$ is the normalized Frobenius residual amplitude. Neither is per-tensor spatial rank or cumulative-displacement occupancy. At the anchor, let $g_a = \nabla L(a)$ and $H_a = \nabla^2 L(a)$. The unrestricted linearized-gradient model, LocalFull, evolves

$$z_{h+1} = z_h - \eta(g_a + H_a z_h), \quad z_0 = 0, \qquad \widehat{\theta}^{\mathrm{LocalFull}}_{P+h} = a + z_h. \tag{2}$$

For an orthonormal basis Q, fixed throughout the forecast window, its matched projected response is

$$c_{h+1} = c_h - \eta(Q^\top g_a + Q^\top H_a Q c_h), \quad c_0 = 0, \qquad \widehat{\theta}^{Q}_{P+h} = a + Q c_h. \tag{3}$$

K1 uses $Q = [q_a]$, where $q_a = -g_a/\|g_a\|$. K2 adds

$$q_2 = \frac{(I - q_a q_a^\top) H_a q_a}{\|(I - q_a q_a^\top) H_a q_a\|}. \tag{4}$$

HistSVD1/2 use the leading one/two right singular vectors of U instead. All four use the same anchor gradient, local operator, step size and horizon; only their retained space changes. History selects the HistSVD basis, but does not replace the operator in its recurrence. We also write $q_1 = q_a$ when naming the first K2 basis vector.

Accuracy target. For TRUE displacement $\Delta_h(\eta) = \theta_{P+h} - a$, define

$$D_0(X, Y) = \frac{1}{H} \sum_{h=1}^{H} \frac{\|X_{P+h} - Y_{P+h}\|_2}{\|\Delta_h(\eta)\|_2}. \tag{5}$$

The common denominators are nonzero for small η under the assumptions below. Write $E(\widehat{\Delta}; \eta) = D_0(\mathrm{TRUE}, a + \widehat{\Delta})$ for displacement predictions. A family is *locally order-adequate* if $E = O(\eta^2)$, preserving LocalFull's guarantee; LocalFull can be higher order or exact for special losses. AUE denotes the horizon-averaged normalized parameter forecast error: $\mathrm{AUE}(K) = D_\gamma(\mathrm{TRUE}, K)$, where D_γ adds $\gamma = 10^{-30}$ to each denominator in Equation 5. The guard is numerically inactive in the retained conditions, but changes the strict limit; asymptotic statements use D_0 (Appendix A.12).

The vector identity

$$\mathrm{TRUE} - K1 = (\mathrm{TRUE} - \mathrm{LocalFull}) + (\mathrm{LocalFull} - K1) \tag{6}$$

distinguishes local-model mismatch from response-space restriction. Distances use the same TRUE denominators; the corresponding scalar errors need not add.

3 What Information Preserves Local Accuracy?

Throughout this section, d is finite, $L \in C^{2,1}$ near θ_0, and exact stateless GD uses this fixed loss. The integers $P, H \geq 2$ are fixed as $\eta \to 0^+$. Define

$$g_0 = \nabla L(\theta_0), \quad q_0 = -g_0/\|g_0\|, \quad H_0 = \nabla^2 L(\theta_0), \quad \beta_0 = \|(I - q_0 q_0^\top) H_0 q_0\|, \tag{7}$$

and assume $g_0 \neq 0$, $\beta_0 > 0$. Anchor quantities are defined analogously. No convexity assumption is required. All remainders are uniform over the fixed window; they are not growing-horizon bounds.

Capture the first two response terms. Writing $G_a = \|g_a\|$, the local expansion is

$$\Delta_h = h\eta G_a q_a - \binom{h}{2} \eta^2 G_a H_a q_a + O(\eta^3). \tag{8}$$

The first term depends on the gradient; the second also needs its Hessian action. Their different scales determine how accurately each must be captured by a compressed space.

Theorem 1 (Necessary and sufficient response capture)**.** *Under the assumptions above, let Π_η be any orthogonal projector, fixed across the forecast window but otherwise arbitrarily dependent on η. Equation 3, with $QQ^\top = \Pi_\eta$, has $O(\eta^2)$ TRUE-normalized error if and only if*

$$\|(I - \Pi_\eta) q_a\| = O(\eta^2), \qquad \|(I - \Pi_\eta) H_a q_a\| = O(\eta). \tag{9}$$

These conditions also characterize the best representation in $\mathrm{range}(\Pi_\eta)$ *with unrestricted coefficients at each horizon.*

The first two horizons give necessity; Equation 8 gives sufficiency (Appendix A.4). This is local moment matching at explicit error orders (Saad, 2003). K1 captures the gradient exactly but omits transverse curvature; K2 captures both. The criterion concerns the forced response from this anchor, not all possible initial perturbations or observables.

Optimize over every common line. Let $\mathcal{A}_1(\eta)$ contain all tuples $(\ell_\eta, \widehat{\Delta}_1, \ldots, \widehat{\Delta}_H)$ with $\dim \ell_\eta = 1$ and $\widehat{\Delta}_h \in \ell_\eta$. The parameter line $a + \ell_\eta$ passes through the anchor and stays fixed across the window; its coefficients may vary freely with h. Its orientation can depend arbitrarily on η and the complete future. This stronger comparison distinguishes a dimensional limitation from a poor choice of K1 direction.

Theorem 2 (Sharp common-line optimum and minimum dimension)**.** *Under the assumptions above, put* $v_h = \Delta_h/\|\Delta_h\|_2$ *and* $J_\eta(\ell) = H^{-1}\sum_h \|(I - \Pi_\ell)v_h\|_2$*. Then*

$$\inf_{\mathcal{A}_1(\eta)} E = \min_{\ell \in \mathrm{Gr}(1,d)} J_\eta(\ell) = c_H^* \beta_0 \eta + O(\eta^2), \qquad c_H^* = \frac{\lfloor H^2/4 \rfloor}{2H}. \tag{10}$$

For K2 on $\mathcal{S}_2 = \mathrm{span}\{g_a, H_a g_a\}$*,*

$$D_0(\mathrm{LocalFull}, K2) = O(\eta^2), \qquad D_0(\mathrm{TRUE}, K2) = O(\eta^2). \tag{11}$$

Thus two is the minimum dimension preserving the local accuracy target among fixed linear response spaces through the common anchor.

With $r_a = (I - q_a q_a^\top) H_a q_a$, normalized future directions satisfy $v_h = q_a - \eta(h-1) r_a/2 + O(\eta^2)$. Every optimal line localizes near q_a; minimizing its leading error reduces to absolute deviation on $\{0, \ldots, H-1\}$. Medians give the sharp constant, including for discontinuous, future-informed choices (Appendix A). K2 contains q_a, r_a and therefore matches the first two terms. Free-offset affine lines, rotating spaces and nonlinear representations are different approximation classes.

Corollary 3 (Exact history contains a sufficient space)**.** *Let* $\Pi_{\mathrm{hist},2}$ *project onto the top-two right singular subspace of the exact noiseless prefix* U*. Under the same assumptions it satisfies Equation 9, and its matched response obeys*

$$D_0(\mathrm{LocalFull}, \mathrm{HistSVD2}) = O(\eta^2), \qquad D_0(\mathrm{TRUE}, \mathrm{HistSVD2}) = O(\eta^2).$$

The normalized prefix has second singular value $\Theta(\eta)$ and tail after rank two of amplitude $O(\eta^2)$; its first two rows imply the required gradient and curvature capture rates (Appendix A.8). Thus K2 is one sufficient construction, not the unique one. Exact-history sufficiency does not establish stability to noise or rounding.

Why a fixed energy threshold discards this information.

Theorem 4 (Historical energy and omitted response)**.** *Under the assumptions above,*

$$M = C_P(\eta\beta_0)^2 + O(\eta^3), \qquad C_P = \frac{P^2 - 1}{12}, \tag{12}$$

$$R := D_0(\mathrm{LocalFull}, K1) = C_H \eta \beta_a + O(\eta^2), \qquad C_H = \frac{H-1}{4}, \tag{13}$$

where $\beta_a = \beta_0 + O(\eta)$*. Consequently,*

$$R = \frac{C_H}{\sqrt{C_P}}\sqrt{M} + O(M). \tag{14}$$

The weak update component is order η relative to dominant motion, so its squared energy fraction is order η^2. Its omitted displacement is order η^2, yielding order-η normalized error. Historical energy therefore calibrates the omitted response within this protocol.

Corollary 5 (K1 loses one accuracy order)**.** *With* $E_F = D_0(\mathrm{TRUE}, \mathrm{LocalFull})$ *and* $E_{K1} = D_0(\mathrm{TRUE}, K1)$*,*

$$|E_{K1} - R| \le E_F, \qquad E_{K1} = C_H \beta_0 \eta + O(\eta^2) = \Theta(\eta). \tag{15}$$

The reverse triangle inequality uses common TRUE denominators and remains valid when LocalFull is higher order or exact. For $H = 24$, the best line's coefficient is 3, versus K1's 5.75: orientation improves the constant, not the order.

Corollary 6 (Fixed energy thresholds select an inadequate dimension)**.** *For any fixed* $\tau \in (0, 1)$*, define*

$$r_\tau(\eta) = \min\left\{ r \ge 1 : \frac{\sum_{j=1}^r \sigma_j(U_\eta)^2}{\|U_\eta\|_F^2} \ge \tau \right\}.$$

Eventually $r_\tau(\eta) = 1$*, while every predictor in* $\mathcal{A}_1(\eta)$ *fails to preserve the* $O(\eta^2)$ *target.*

A scale-dependent threshold can retain the sufficient space. In the exact noiseless regime, choose $\tau(\eta) = 1 - \kappa M(\eta)$, $0 < \kappa < 1$. The relative tail after rank two is $T_2 = O(\eta^4)$, so $T_2 \le \kappa M < M$ eventually selects exactly two directions (Appendix A.9). Merely setting $1-\tau = O(\eta^2)$ is insufficient unless its coefficient is small enough. This is an asymptotic spectral consequence, not a noise-robust or finite-tolerance certificate.

4 Matched Evidence for the Weak-Direction Mechanism

The primary panel uses Qwen3-1.7B, SmolLM2-1.7B, Llama-3.2-1B, and OLMo-2-0425-1B, sharing 32 mathematical problem–solution examples from `siyanzhao/Openthoughts_math_30k_opsd` in eight fixed four-example batches (source identities and exclusions: Appendices B and H). Each condition takes 32 stateless GD steps on assistant-only cross-entropy, with only final-block query/value LoRA-B coordinates trainable. Model computation is FP32, dropout is disabled, and the step size is fixed by $\eta\|g_0\| = 0.03$. We forecast $H = 24$ updates after a prefix of $P = 8$; Appendix B gives the full protocol. Fixed drivers and stateless GD isolate local geometry; the selected coordinates keep repeated matrix-free Hessian actions tractable and match derivatives to TRUE continuation in the same space. This is a controlled probe of local response, not full-parameter post-training.

The four-model fixed-driver panel tests whether the weak transverse direction improves response accuracy when the anchor, local operator, step size, horizon, and metric are held fixed. Across all 32 conditions, historical rank-one occupancy is 0.99664–0.99999, yet the direct LocalFull–K1 restriction exceeds the TRUE–LocalFull mismatch in every condition (Figure 2(a)). The space restriction therefore exceeds the error already present in the unrestricted local approximation; both distances use the same TRUE displacement denominator.

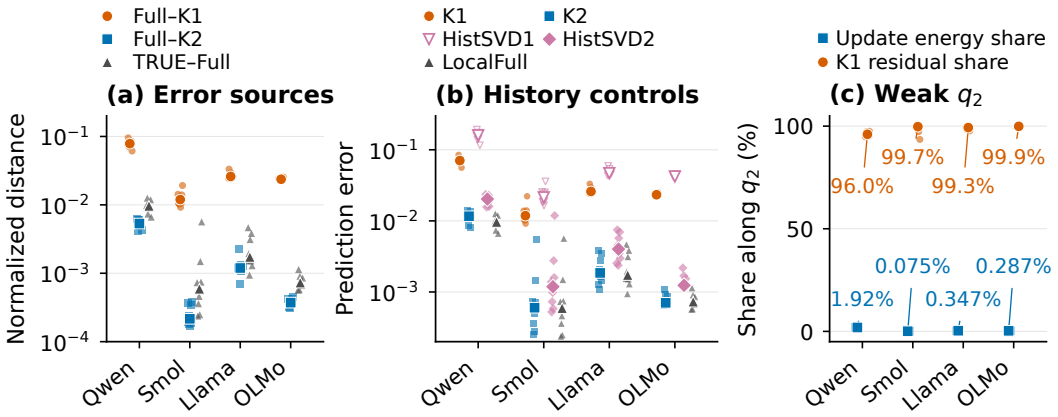


Figure 2: Matched fixed-driver responses: eight conditions per model, with $P = 8$ historical and $H = 24$ future updates. Points are conditions and larger markers are family medians; Full denotes LocalFull. (a) Direct response-space restriction compared with local-model mismatch, using TRUE-normalized distances. (b) TRUE-normalized forecast errors for the same anchor and operator. K2 and HistSVD2 improve their one-dimensional counterparts in all conditions. (c) The weak curvature direction q_2 carries little future-update energy but most K1 squared residual; the two percentages use different denominators. Labels give family medians. This single-scale panel tests the mechanism, not asymptotic rates.

Adding the first transverse curvature direction gives K2 and improves K1 in all 32 conditions. The leading historical singular-vector response HistSVD1 is also less accurate than LocalFull throughout. HistSVD2 improves HistSVD1 in every condition, while K2 remains more accurate at this finite step scale (Figure 2(b); Appendix D.3). Thus useful response information is already present in the history, although the historical and curvature-informed planes need not perform identically at finite scale. K2 also has lower total AUE than a two-dimensional history control containing the same anchor gradient in all 32 conditions (Appendix D.1); its finite-scale advantage is therefore not explained solely by gradient capture.

Localization identifies the weak direction's role. Family-median shares along q_2 are 0.075%–1.92% of future-update energy but 96.0%–99.9% of K1's squared residual over the same horizons (Figure 2(c); Appendix D.2). These fractions have different denominators: total realized update energy and total squared prediction error, respectively.

Historical tail energy and direct K1 restriction have pooled Spearman correlation 0.999 within this panel. This ordering is consistent with the theory, but rank correlation does not test the square-root form or its coefficient; Appendix C examines those quantitatively. The criterion for retaining the local model's *accuracy order* nevertheless requires capturing the weak direction, not merely retaining most historical energy.

5 Quantitative Validation in Nonlinear LLM Adaptation

The multi-scale study tests the predicted motion order, the sharp one-dimensional obstruction, and the accuracy guarantee for suitable two-dimensional spaces. Qwen3-1.7B and Llama-3.2-1B each contribute four independent model–data blocks, evaluated at five nested scales $\delta = \eta\|g_0\|_2$ from 0.03 to 0.001875. Each scale resets to the same initialization, with fixed loss, coordinates, $P = 8$, $H = 24$, and exact automatic-differentiation Hessian-vector products (HVPs). The replication unit is the model–block ($n = 8$), not each of the 40 scale conditions. Slopes are fitted within blocks; Appendix C.1 specifies the numerical resolution rule and all included scales.

We bound the best anchored, window-fixed line using the sum of disjoint-pair projective distances as a lower bound and the line oriented by the TRUE $h = 12$ displacement as a feasible upper bound (Appendix C.2). Both use the complete future only for geometric evaluation. They bound the same line class as the theorem and require no numerical optimization over line directions.

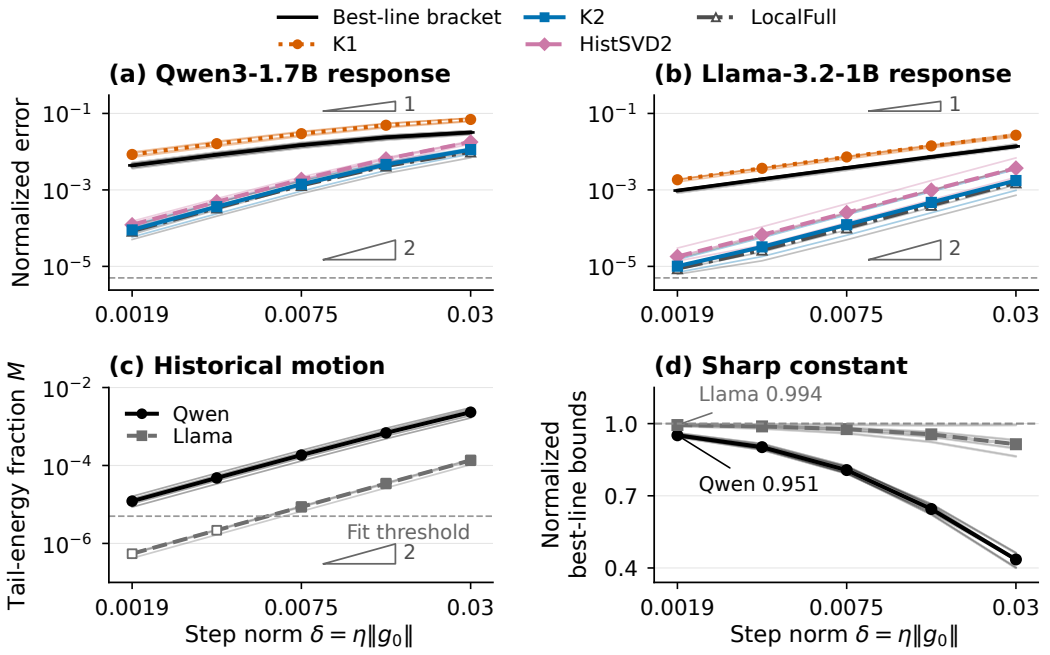


Figure 3: Five-scale exact-HVP validation. Thin curves: four blocks per model; thick curves: medians at common δ. (a,b) TRUE-normalized errors and tight best-line lower/upper bounds. Triangular guides labeled 1 and 2 show theoretical slopes, not fits. (c) Historical tail energy M; hollow points retain measured values below the dashed 5×10^{-6} threshold and are excluded from slope fits. (d) Bounds $[L_{\mathrm{disj}}, U_{\mathrm{med}}]/(3\eta\beta_0)$, reference one; labels give smallest-scale upper-bound medians. In (c,d), circles/solid lines denote Qwen, squares/dashes Llama. Bounds are geometric, not confidence intervals. The 99% energy rule selects rank one in every block and scale.

Historical tail energy scales approximately quadratically over the resolved scales (Figure 3(c)). Finite-scale agreement differs by model. K1 has median slope 0.77 for Qwen and 0.98 for Llama, whereas K2, HistSVD2, and LocalFull have median slopes of about 1.73–1.92. Llama is closer to the leading asymptotic predictions; Appendix C reports the complete fits and geometric bounds. The results support the difference in accuracy orders; they do not recover limiting exponents from five scales.

The sharp coefficient provides a separate quantitative check. For $H = 24$, the best-line leading error is $3\eta\beta_0$, compared with $5.75\eta\beta_0$ for K1. At the smallest scale, the model-median normalized lower and upper bounds nearly coincide, at about 0.951 for Qwen and 0.994 for Llama. Thus line orientation improves the constant, not the order; Qwen still shows about 4.9% finite-scale deviation from the predicted constant. All eight blocks select $r_{0.99} = 1$ at all five scales, even though this space loses the target accuracy order. The historical sufficiency theorem concerns exact noiseless histories; these tests use floating-point gradients and expose the corresponding numerical resolution boundary.

6 When Does Motion Compress?

The response results assume a fixed loss. A separate local law explains when curvature governs historical compression in the presence of varying drivers. For fixed finite P and losses with uniformly bounded, locally Lipschitz Hessians near θ_0, consider

$$L_t^\lambda = (1-\lambda)L_{\mathrm{ref}} + \lambda L_t, \qquad \theta_t = \theta_{t-1} - \eta \nabla L_t^\lambda(\theta_{t-1}), \tag{16}$$

where $\lambda \in [0,1]$, $\eta > 0$, and $g_0 = \nabla L_{\mathrm{ref}}(\theta_0) \neq 0$. Let $D_\perp$ and $K_\perp$ be the normalized driver-gradient variation and first curvature correction, respectively, after centering across steps and projecting orthogonally to g_0; Appendix A.11 gives their explicit definitions.

Proposition 1 (Varying-driver motion law). As $(\lambda, \eta) \to (0,0)$ jointly, with $\rho = \lambda + \eta$,

$$M(\lambda,\eta) = \frac{1}{P}\|\lambda D_\perp + \eta K_\perp\|_F^2 + O(\rho^3) = A\lambda^2 + 2B\lambda\eta + C\eta^2 + O(\rho^3), \tag{17}$$

where

$$(A,B,C) = \frac{1}{P}\big(\|D_\perp\|_F^2, \langle D_\perp, K_\perp\rangle_F, \|K_\perp\|_F^2\big). \tag{18}$$

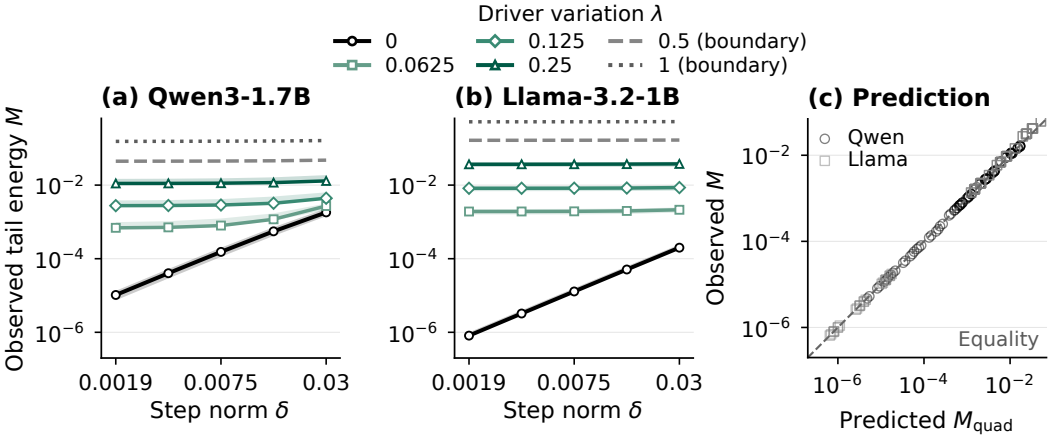


Figure 4: Driver variation and curvature in historical motion. (a,b) Model medians of tail energy M across eight blocks at each (λ, δ); bands show block interquartile ranges in the primary driver range. Gray dashed/dotted $\lambda = 0.5, 1$ curves are boundary diagnostics. (c) Each point compares measured M with the coefficient-based quadratic prediction in one primary condition ($\lambda \leq 0.25$, three smallest δ): 96 per model, nested within eight blocks. The diagonal is equality, not a fitted line; median absolute $\log_{10}(M/M_{\mathrm{quad}})$ is 0.013 for Qwen and 0.042 for Llama. The plateaus are finite-scale observations.

Thus $A, C \geq 0$ and $B^2 \leq AC$, but B has either sign: driver variation and curvature can reinforce or partially cancel. At $\lambda = 0$, this recovers $M = C_P(\eta\beta_0)^2 + O(\eta^3)$. When $D_\perp \neq 0$, the leading driver contribution is $A\lambda^2$; this joint-local statement neither proves a fixed-λ positive floor nor asserts monotonicity. The expansion applies to stateless GD near $(\lambda, \eta) = (0, 0)$.

The nonlinear validation contains 480 grid conditions within 16 model–data blocks, eight per model. Its primary region is $\lambda \leq 0.25$ and the three smallest step scales: 96 nested conditions per model. Observed-versus-predicted log–log slopes are 1.009 (Qwen) and 1.023 (Llama), with median absolute base-10 log-ratios 0.013 and 0.042. The tested nonzero-driver slices plateau as step size decreases, consistent with the leading driver term. Both $\lambda = 0.5$ and 1 are outside the primary fit (Figure 4; Appendix E.1). A separate 2×2 study finds more compressed motion under repeated than shuffled drivers, with prevalence also depending on B-only versus A+B parameterization (Appendix E.3). These studies locate the compressed-motion regime; they do not extend the fixed-loss response theorem to changing drivers.

7 What Should Training Compression Preserve?

Match the compression criterion to the target. The capture conditions in Equation 9 characterize preservation of the local model's $O(\eta^2)$ normalized-error guarantee. The best common-line error also vanishes, so a single direction can eventually meet a fixed nonzero tolerance while losing that order. Fixed windows shrink in elapsed optimization time as $\eta \to 0$; the result does not give a fixed-time or long-horizon forecasting guarantee.

Direction selection under shuffled AdamW. A separate two-model study uses shuffled mini-batches and final-block q/v LoRA A+B. At the same energy-selected dimension, a response-informed control has lower LocalFull restriction error than historical SVD in 32/32 windows and all eight trajectory medians. Including the current update direction in the history space while keeping its total dimension fixed leaves the control better in 29/32 windows and all trajectory medians (Appendix E.5). Trajectories are replication units; windows are nested observations. The control uses extra schedule-conditioned gradients, HVPs, and local predictions, without nonlinear future TRUE states. This supports direction-selection mismatch at equal dimension, with unequal information and computation; it establishes neither a matched-budget advantage nor two-dimensional sufficiency. Driver and optimizer-state changes introduce additional forecasting errors (Appendix F).

Functional and continuation diagnostics. An oracle representation diagnostic projects TRUE future displacements onto one or two historical directions. The median held-out NLL gain from the second direction is 4.25×10^{-6} nats/token; same-objective continuation gains are also below the predeclared materiality threshold (Appendix G.1). The coefficients use future information, so this is not an online predictor evaluation. These effects do not establish material learning gains or practical acceleration.

8 Related Work

Historical energy and dynamical accuracy. POD and snapshot SVD optimize captured energy (Sirovich, 1987; Eckart & Young, 1936), but weak-energy modes can control transient response (Ilak & Rowley, 2008). Solution-only POD snapshots can give suboptimal pointwise error bounds; difference-quotient augmentation restores optimal bounds under the respective assumptions (Iliescu & Wang, 2014; Koc et al., 2021). Local SVD also recovers Frenet frames and curvature ratios (Álvarez-Vizoso et al., 2019), and fixed-rank matrix geometry provides our two-sided normal projection (Vandereycken, 2013). These results motivate distinguishing historical reconstruction from response accuracy. We quantify that distinction for adjacent GD updates: exact history contains an order-adequate space, while a fixed energy threshold selects a dimension whose best possible common-line response loses one order.

Krylov response and optimization. Balanced reduction uses controllability and observability (Moore, 1981); Krylov methods match operator response (Grimme, 1997; Saad, 2003), and Krylov subspace descent explicitly uses $\mathrm{span}\{g, Hg, \ldots\}$ in deep learning (Vinyals & Povey, 2011). K2 and

its first two matched response terms are classical ingredients. For quadratic recovery, Piontkovskaia & Nikolenko (2026) give Krylov containment and exact invariant-space conditions. Our target is approximate response over the whole finite window: Theorem 1 gives the necessary and sufficient capture rates, and Theorem 2 optimizes the leading error over every anchored fixed line, including future-informed choices. Together with historical sufficiency, these results identify the minimum dimension and its energy-selection consequence.

Training subspaces and extrapolation. Trajectory PCA motivates projected retraining (Li et al., 2023), while Song et al. (2025) show that dominant-Hessian-subspace alignment need not make projected optimization effective. DMD and training-Jacobian methods model evolution from data or derivatives (Schmid, 2010; Tu et al., 2014; Belrose & Scherlis, 2024); intrinsic dimension, LoRA and gradient projection address parameterization or storage (Li et al., 2018; Aghajanyan et al., 2021; Hu et al., 2022; Zhao et al., 2024).

Recent RLVR extrapolation studies use cumulative checkpoint deltas or per-tensor spatial rank (Cai et al., 2026; Wei et al., 2026; Chen et al., 2026; Wang et al., 2026). RELEX combines trajectory SVD, coefficient linearity and task evaluations; its rank-one extrapolator outperforms ranks five and ten in its ablation. NExt learns nonlinear evolution of low-rank checkpoint-delta representations (Chen et al., 2026); its per-tensor spatial rank does not specify a fixed one-dimensional parameter response space. ExPO amplifies the change from an initial SFT checkpoint to a partially aligned model, motivated by a first-order approximation (Zheng et al., 2025). Our common-line bound permits arbitrary, including nonlinear, coefficients across horizons: coefficient flexibility alone cannot recover an omitted transverse direction. Evolving spaces and nonlinear representations fall outside that class, and whether NExt captures the identified curvature requires a separate test. Task gains also need not preserve a particular parameter trajectory's local accuracy order. These differences in targets, representations and information budgets allow successful extrapolation without contradicting our result.

9 Conclusion

Trajectory compression preserves the stated local response accuracy when it captures the gradient and its first curvature action to the required orders. Under fixed-loss GD with nonzero transverse curvature, even the best anchored fixed line loses one order, whereas K2 and exact historical top-two retain it. The historical top-two space contains sufficient information that a fixed explained-energy threshold below 100% eventually discards. Energy still calibrates the omitted response; it is the retention criterion that must match the accuracy target. Nonlinear LLM experiments support this mechanism within the tested scales. Next steps include extending the guarantees to noisy histories and stateful optimizers, estimating response information with few directional queries, and testing retention criteria against output accuracy at matched computational budgets.

Reproducibility Statement

Appendix A gives complete proofs of the local response theorems, the K1 and historical-space corollaries, and the fixed and scale-dependent energy-threshold consequences. The appendix guide separates empirical cohorts, repeated analyses, and replication units. The subsequent appendices report full protocols, synthetic checks, numerical validation, and optimizer, driver, and parameterization boundaries. Appendix H records model, data, and adapter identities and distinguishes aggregate recomputation, vector-based analysis, and trajectory re-execution. The accompanying code-and-evidence supplement indexes condition-level results, protocol metadata, and model-free verification tools.

AI Use Statement

Generative AI assisted literature search, theoretical derivation and proof development, code development, experiment design and orchestration, analysis, and manuscript preparation. The authors independently verified the mathematical arguments, citations, experimental results, and released code and remain responsible for the scientific content.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.acl-long.568/.

J. Álvarez-Vizoso, Robert Arn, Michael Kirby, Chris Peterson, and Bruce Draper. Geometry of Curves in $\mathbb{R}^n$ from the Local Singular Value Decomposition. *Linear Algebra and its Applications*, 571:180–202, 2019. doi: 10.1016/j.laa.2019.02.006.

Nora Belrose and Adam Scherlis. Understanding Gradient Descent through the Training Jacobian. *arXiv preprint arXiv:2412.07003*, 2024. URL https://arxiv.org/abs/2412.07003.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guiquan Liu, and Junfeng Fang. On Predictability of Reinforcement Learning Dynamics for Large Language Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4c454d34f3a4c8d6b4ca85a918e5d7ba-Paper-Conference.pdf.

Zhipeng Chen, Tao Qian, Wayne Xin Zhao, and Ji-Rong Wen. Low-rank Optimization Trajectories Modeling for LLM RLVR Acceleration. *arXiv preprint arXiv:2604.11446*, 2026. URL https://arxiv.org/abs/2604.11446.

Carl Eckart and Gale Young. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Eric James Grimme. *Krylov Projection Methods for Model Reduction*. PhD thesis, University of Illinois at Urbana-Champaign, 1997. URL https://hdl.handle.net/2142/81180.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Milos Ilak and Clarence W. Rowley. Modeling of Transitional Channel Flow Using Balanced Proper Orthogonal Decomposition. *Physics of Fluids*, 20:034103, 2008. URL https://arxiv.org/abs/0707.4112.

Traian Iliescu and Zhu Wang. Are the Snapshot Difference Quotients Needed in the Proper Orthogonal Decomposition? *SIAM Journal on Scientific Computing*, 36(3):A1221–A1250, 2014. doi: 10.1137/130925141.

Birgul Koc, Samuele Rubino, Michael Schneier, John R. Singler, and Traian Iliescu. On Optimal Pointwise in Time Error Bounds and Difference Quotients for the Proper Orthogonal Decomposition. *SIAM Journal on Numerical Analysis*, 59(4):2163–2196, 2021. doi: 10.1137/20M1371798.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the Intrinsic Dimension of Objective Landscapes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08838.

Tao Li, Lei Tan, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low Dimensional Trajectory Hypothesis is True: DNNs can be Trained in Tiny Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3411–3420, 2023. doi: 10.1109/TPAMI.2022.3178101. URL https://arxiv.org/abs/2103.11154.

Bruce C. Moore. Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981. doi: 10.1109/TAC.1981.1102568.

Irina Piontkovskaia and Sergey Nikolenko. Recoverable but Not Stationary: Local Linear Structures in Weights and Activations. *arXiv preprint arXiv:2606.10929*, 2026. URL https://arxiv.org/abs/2606.10929.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, second edition, 2003.

Peter J. Schmid. Dynamic Mode Decomposition of Numerical and Experimental Data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

Lawrence Sirovich. Turbulence and the Dynamics of Coherent Structures. I. Coherent Structures. *Quarterly of Applied Mathematics*, 45(3):561–571, 1987. doi: 10.1090/qam/910462.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD Really Happen in Tiny Subspaces? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.16002.

Jonathan H. Tu, Clarence W. Rowley, Dirk M. Luchtenburg, Steven L. Brunton, and J. Nathan Kutz. On Dynamic Mode Decomposition: Theory and Applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014. URL https://arxiv.org/abs/1312.0041.

Bart Vandereycken. Low-Rank Matrix Completion by Riemannian Optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. URL https://arxiv.org/abs/1209.3834.

Oriol Vinyals and Daniel Povey. Krylov Subspace Descent for Deep Learning. *arXiv preprint arXiv:1111.4259*, 2011. URL https://arxiv.org/abs/1111.4259.

Tianle Wang, Jiayu Liu, Zhongyuan Wu, Shenghao Jin, Wei Chen, Hao Xu, and Ning Miao. Linear Dynamics in the RLVR Training of Large Language Models. *arXiv preprint arXiv:2601.04537*, 2026. URL https://arxiv.org/abs/2601.04537.

Zhepei Wei, Xinyu Zhu, Wei-Lin Chen, Chengsong Huang, Jiaxin Huang, and Yu Meng. You Only Need Minimal RLVR Training: Extrapolating LLMs via Rank-1 Trajectories. *arXiv preprint arXiv:2605.21468*, 2026. URL https://arxiv.org/abs/2605.21468.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.03507.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Model Extrapolation Expedites Alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1025–1041, 2025. doi: 10.18653/v1/2025.acl-long.51. URL https://aclanthology.org/2025.acl-long.51/.

APPENDIX GUIDE: RESULTS, COHORTS, AND ARTIFACTS

Appendix A proves the local-response results. The table below separates the main empirical tests from supporting analyses; counts of windows, scales, and methods are not additional independent replications. Throughout, HistSVD1/2 denotes the matched recurrence in the first one/two historical singular directions. Archived field names and filenames may retain the alias `LocalSVD1/2`.

Evidence group	Observations and grouping	Question supported / location
Primary fixed-driver panel	4 models $\times$ 8 shared data batches; 32 conditions, one step scale	Matched response errors and weak-direction localization. History, geometry, mapping, and KL analyses reuse these conditions (Appendices D.1, D.2, D.3, D.4, G).
Primary exact-HVP multi-scale study	2 models $\times$ 4 data blocks $\times$ 5 nested scales; 8 model–blocks, 40 conditions	Finite-scale orders, best-line bounds, and fixed-threshold ranks. Sharp-constant calibration reuses the saved vectors (Appendices C.1, C.2).
Varying-driver grid	2 models $\times$ 8 blocks $\times$ 6 driver strengths $\times$ 5 scales; 16 blocks, 480 cells	Local occurrence law; primary fit uses 192 cells in total, nested within the same 16 blocks (Appendix E.1).
Shuffled A+B AdamW history study	2 models $\times$ 4 trajectories $\times$ 4 windows; 8 trajectories, 32 windows	Energy-selected versus response-informed spaces outside the theorem. Re-executes the S-AB occurrence arm; practical-budget diagnostics reuse these windows (Appendices E.5, F).
Supporting cross-model scaling	2 models $\times$ 12 data blocks $\times$ 5 scales; 24 blocks, 120 conditions	Motion/restriction coefficients with finite-difference actions and their numerical limits (Appendices C.3, C.7).
Full-FP32 exact-HVP Qwen3-8B confirmation	8 data blocks $\times$ 5 scales; 40 conditions	Motion/K1-restriction scaling (Appendix C.5).
Weak-direction functional study	Reuses the 8 multi-scale blocks; 15 scale–horizon cells per block, then one continuation setting	Oracle historical projections test NLL and continuation effects, which are small. These are representation diagnostics, not HistSVD recurrence forecasts (Appendix G.1).
Other scope and control studies	Cohort sizes and dependencies stated locally	AdamW, secant, numerical, initialization, and random-direction diagnostics answer distinct questions. CORR2 has 128 windows nested in 16 trajectories; its controls share those windows (Appendix E.6).

Finding and checking evidence. Shared definitions and fixed-driver settings are in Appendix B; matched-history, localization and subspace diagnostics are grouped with the fixed-driver results. Driver/optimizer studies keep their separate cohorts and information budgets. The supplement's evidence index identifies complete subspace diagnostics (S1), historical 8B results (S2), driver schedules (S3), and analytic checks (S4). Scalar aggregation, vector-based geometry, and trajectory re-execution are different reproduction tasks (Appendix H.2).

A PROOFS OF THE LOCAL RESPONSE RESULTS

All fixed-loss proofs use exact stateless GD, exact Hessian actions and strict TRUE normalization D_0. Uniform expansions first establish the motion and K1 rates. The response-capture criterion then supplies K2 and exact-history sufficiency; the sharp common-line optimum and energy-tail orders yield the dimension and threshold conclusions. The varying-driver law has its own joint-local assumptions below. Appendix A.12 separates the implemented denominator guard from the strict limit.

A.1 ASSUMPTIONS AND UNIFORM LOCAL EXPANSIONS

The coordinates lie in $\mathbb{R}^d$ for fixed finite d. The loss is fixed as $\eta \to 0^+$ and belongs to $C^{2,1}$ on an open neighborhood of θ_0. In particular, on a sufficiently small closed ball, the Hessian is bounded and Lipschitz, and

$$\nabla L(x+v) = \nabla L(x) + \nabla^2 L(x)v + O(\|v\|_2^2) \tag{19}$$

holds uniformly when $x, x+v$ remain in that ball. Fix $P, H \geq 2$, $g_0 \neq 0$ and $\beta_0 > 0$. Only finitely many $(P+H)$ exact GD steps are considered. For sufficiently small η, bounded gradients and induction keep all iterates in the ball and give $\theta_k - \theta_0 = O(\eta)$ uniformly over these indices. Writing $G_0 = \|g_0\|_2$, one further application of Equation 19 gives

$$\theta_k - \theta_0 = k\eta G_0 q_0 + O(\eta^2).$$

Thus $a = \theta_P \to \theta_0$. Continuity implies $G_a = \|g_a\|_2$ and β_a are bounded away from zero for sufficiently small η. Local derivative bounds are uniform in the moving anchor. Every remainder below may depend on the fixed windows, dimension and these local bounds, but not on η.

The same fixed-index induction, now expanded about the moving anchor, gives for $1 \leq h \leq H$

$$\Delta_h := \theta_{P+h} - a = h\eta G_a q_a + O(\eta^2). \tag{20}$$

Since $G_a \to G_0 > 0$ and H is fixed, one remainder constant works for all $h \leq H$, and $\|\Delta_h\|_2 \geq h\eta G_a - C\eta^2 > 0$ simultaneously for sufficiently small η. This proves nonvanishing of the strict D_0 denominators without introducing a fixed positive guard.

A.2 PREFIX MOTION NONOCCUPANCY

Applying the uniform gradient expansion at θ_0,

$$u_t = -\eta \nabla L(\theta_{t-1}) \tag{21}$$

$$= \eta G_0 q_0 - (t-1)\eta^2 G_0 H_0 q_0 + O(\eta^3), \quad t = 1, \ldots, P. \tag{22}$$

Let $r_0 = (I - q_0 q_0^\top) H_0 q_0$, so $\|r_0\|_2 = \beta_0$ and $H_0 q_0 = \alpha q_0 + r_0$, with $\alpha = q_0^\top H_0 q_0$. For $\mathbf{j} = (0, 1, \ldots, P-1)^\top$ and $\mathbf{1} = (1, \ldots, 1)^\top$, the normalized row matrix is

$$A = \frac{U}{\eta G_0} = \mathbf{1} q_0^\top - \eta \mathbf{j} (\alpha q_0 + r_0)^\top + O(\eta^2). \tag{23}$$

The remainder is in Frobenius norm since P, d are fixed.

For completeness, the first-order best-rank-one residual must optimize over the singular direction rather than hold it fixed at q_0. At $A_0 = \mathbf{1} q_0^\top$, the leading singular value is $\sqrt{P}$ and the others are zero. The nonzero gap makes the leading singular projectors, and hence the best rank-one approximation, smooth in a neighborhood of A_0. The tangent space of the nonzero rank-one matrices there consists of $x q_0^\top + \mathbf{1} y^\top$. Orthogonal projection of a perturbation E onto its normal space is

$$\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{P} \right) E (I - q_0 q_0^\top).$$

For the first-order perturbation in Equation 23, the parallel term is entirely tangent. The normal component of the coefficient of η is $-(\mathbf{j} - \bar{j}\mathbf{1}) r_0^\top$, with $\bar{j} = (P-1)/2$. Smoothness of the nearest

rank-one projection therefore gives

$$\sum_{i\geq 2}\sigma_i(A)^2 = \eta^2\beta_0^2\sum_{k=0}^{P-1}(k-\bar{j})^2 + O(\eta^3) \tag{24}$$

$$= \eta^2\beta_0^2\frac{P(P^2-1)}{12} + O(\eta^3). \tag{25}$$

Division by $\|A\|_F^2 = P + O(\eta)$ proves Equation 12. If $\beta_0 = 0$, the displayed leading coefficient vanishes; the nondegenerate rate equivalence and ratio conclusion are not asserted.

A.3 Analytically normalized response restriction

At the actual nonlinear prefix endpoint define $\Pi_a = q_a q_a^\top$ and $r_a = (I - \Pi_a)H_a q_a$. The LocalFull recurrence yields, by induction for each fixed h,

$$z_h = h\eta G_a q_a - \binom{h}{2}\eta^2 G_a H_a q_a + O(\eta^3). \tag{26}$$

Write the ambient K1 displacement as $y_h = q_a c_h$. Since the forcing lies in this subspace, Equation 3 becomes

$$y_{h+1} = y_h - \eta(\Pi_a g_a + \Pi_a H_a y_h), \qquad y_0 = 0.$$

The same induction gives

$$y_h = h\eta G_a q_a - \binom{h}{2}\eta^2 G_a \Pi_a H_a q_a + O(\eta^3).$$

Consequently

$$z_h - y_h = -\binom{h}{2}\eta^2 G_a r_a + O(\eta^3). \tag{27}$$

At $h = 1$, both displacements are exactly $-\eta g_a$. At $h = 2$, the vector identity is exactly $z_2 - y_2 = -\eta^2 G_a r_a$ for the exact frozen operator.

The denominator is the nonlinear TRUE displacement, not z_h. Using the same fixed loss and step in the Taylor expansion about a gives

$$\begin{aligned}\Delta_h &= h\eta G_a q_a - \binom{h}{2}\eta^2 G_a H_a q_a + O(\eta^3),\\ \|\Delta_h\|_2 &= h\eta G_a + O(\eta^2).\end{aligned} \tag{28}$$

Comparing with Equation 26 gives $(\theta_{P+h} - a) - z_h = O(\eta^3)$ for each fixed h, and division by the displayed denominator gives $D_0(\mathrm{TRUE}, \mathrm{LocalFull}) = O(\eta^2)$. Taking the norm of Equation 27 and dividing yields

$$\frac{\|z_h - y_h\|_2}{\|\theta_{P+h} - a\|_2} = \frac{h-1}{2}\eta\beta_a + O(\eta^2). \tag{29}$$

The reverse triangle inequality controls the norm remainder; $h = 1$ has zero leading term and zero exact numerator. Averaging over $h = 1, \ldots, H$ gives $H^{-1}\sum_h (h-1)/2 = (H-1)/4$, proving Equation 13. The anchor and its gradient, Hessian and transverse curvature depend on η. Since $a - \theta_0 = O(\eta)$, Lipschitzness of the Hessian and the nondegenerate normalization give $\beta_a = \beta_0 + O(\eta)$ uniformly.

Total K1 error. For each horizon, divide the reverse triangle inequality

$$|\|\Delta_h - y_h\|_2 - \|z_h - y_h\|_2| \leq \|\Delta_h - z_h\|_2$$

by the common positive TRUE denominator $\|\Delta_h\|_2$. Averaging and applying the triangle inequality to the average gives

$$|D_0(\mathrm{TRUE}, K1) - D_0(\mathrm{LocalFull}, K1)| \leq D_0(\mathrm{TRUE}, \mathrm{LocalFull}). \tag{30}$$

The right-hand side is $O(\eta^2)$, whereas Equation 13 and $\beta_a = \beta_0 + O(\eta)$ give $D_0(\mathrm{LocalFull}, K1) = C_H\beta_0\eta + O(\eta^2)$. This proves Corollary 5. No division by the LocalFull error is used, so the argument also covers a higher-order or exact LocalFull model.

Energy–response calibration. Since $\beta_0 > 0$, Equation 12 implies $\sqrt{M} = \eta\beta_0\sqrt{C_P} + O(\eta^2)$. Together with Equation 13 and $\beta_a = \beta_0 + O(\eta)$, this gives

$$R = \frac{C_H}{\sqrt{C_P}}\sqrt{M} + O(M). \tag{31}$$

For $P = 8, H = 24$, $C_P = 21/4$, $C_H = 23/4$, and $C_H/\sqrt{C_P} = 23/(2\sqrt{21}) = 2.5095057377\ldots$. The sharp common-line coefficient is $c^*_{24} = \lfloor 24^2/4 \rfloor/(2 \cdot 24) = 3$, so

$$\inf_{\mathcal{A}_1(\eta)} E = \frac{3}{\sqrt{21/4}}\sqrt{M} + O(M) = 1.309307341\ldots\sqrt{M} + O(M).$$

A.4 Response-order preservation for a projected space

We prove Theorem 1. Let $\Pi = \Pi_\eta$ be an orthogonal projector; its range remains fixed within the forecast window, but its rank and orientation may depend arbitrarily on η, including through future information. Set $q = q_a$, $G = G_a$, and $A = H_a$ in this subsection. Define the best representation error by

$$E_{\mathrm{rep}}(\Pi;\eta) = \frac{1}{H}\sum_{h=1}^{H}\frac{\|(I-\Pi)\Delta_h\|_2}{\|\Delta_h\|_2}. \tag{32}$$

Orthogonal projection achieves the minimum over unrestricted horizon coefficients in $\mathrm{range}(\Pi)$. Let $E_{\mathrm{rec}}(\Pi;\eta)$ denote the same TRUE-normalized error for $y_0 = 0$, $y_{h+1} = y_h - \eta\Pi(g_a + Ay_h)$, the full-coordinate form of Equation 3.

Equation 28 and fixed-H induction give

$$\Delta_h = h\eta Gq - \binom{h}{2}\eta^2 GAq + O(\eta^3), \tag{33}$$

$$y_h = h\eta G\Pi q - \binom{h}{2}\eta^2 G\Pi A\Pi q + O(\eta^3). \tag{34}$$

Remainders are uniform over $h \le H$ and orthogonal Π: G, A are bounded at the moving anchor and $\|\Pi\| \le 1$. Neither convergence of Π_η nor exact invariance under A is required.

First, $E_{\mathrm{rep}} \le E_{\mathrm{rec}}$ because every y_h lies in the projected space. Next, suppose $E_{\mathrm{rep}} = O(\eta^2)$. Since H is fixed and the horizon terms are nonnegative, every term is $O(\eta^2)$. At $h = 1$, $\Delta_1 = \eta Gq$ exactly, giving $\|(I-\Pi)q\| = O(\eta^2)$. At $h = 2$, $\|\Delta_2\| = \Theta(\eta)$ and hence

$$(I-\Pi)\Delta_2 = 2\eta G(I-\Pi)q - \eta^2 G(I-\Pi)Aq + O(\eta^3) = O(\eta^3).$$

The first term is already $O(\eta^3)$. Since G is bounded away from zero, division by $\eta^2 G$ gives $\|(I-\Pi)Aq\| = O(\eta)$.

Conversely, suppose these two leakage rates hold. The identity

$$Aq - \Pi A\Pi q = (I-\Pi)Aq + \Pi A(I-\Pi)q$$

and Equation 34 imply $\Delta_h - y_h = O(\eta^3)$ uniformly for $h \le H$. The strict denominators are $\Theta(\eta)$, so $E_{\mathrm{rec}} = O(\eta^2)$. Together these implications prove the equivalence of representation adequacy, the two leakage conditions, and recurrence adequacy.

This local projected moment-matching criterion (Saad, 2003) does not itself require $\beta_0 > 0$; that assumption is needed for the one-dimensional lower bound and minimum dimension of two. The proof gives the asymptotic estimate

$$E_{\mathrm{rec}} \le C\big(\|(I-\Pi)q\| + \eta\|(I-\Pi)Aq\| + \eta^2\big)$$

for sufficiently small η, with C independent of the projector. It does not provide computable finite-tolerance constants without local derivative and denominator bounds.

A.5 K2 SECOND-ORDER SUFFICIENCY

Since $\beta_a = \beta_0 + O(\eta) > 0$ for small η,

$$\mathcal{S}_2(\eta) = \operatorname{span}\{g_a, H_a g_a\} = \operatorname{span}\{q_a, r_a\}$$

has dimension exactly two. Its orthogonal projector Π_2 satisfies

$$\Pi_2 q_a = q_a, \qquad \Pi_2 H_a q_a = H_a q_a = (q_a^\top H_a q_a) q_a + r_a. \tag{35}$$

For $Q_2 = [q_a, r_a/\beta_a]$, the ambient form of Equation 3 is $y_{h+1}^{(2)} = y_h^{(2)} - \eta \Pi_2 (g_a + H_a y_h^{(2)})$, with $y_0^{(2)} = 0$. In Equation 34, both its leading term and its second-order term therefore equal those of LocalFull in Equation 26. Hence $z_h - y_h^{(2)} = O(\eta^3)$ uniformly, giving

$$D_0(\text{LocalFull}, K2) = O(\eta^2),$$
$$D_0(\text{TRUE}, K2) \le D_0(\text{TRUE}, \text{LocalFull}) + D_0(\text{LocalFull}, K2) = O(\eta^2).$$

These bounds allow higher-order or exact special cases; they do not assert sufficiency of an arbitrary two-dimensional space.

A.6 SHARP FULL-HORIZON COMMON-LINE OPTIMIZATION

Set $v_h = \Delta_h / \|\Delta_h\|_2$. For a fixed line ℓ, rescale each prediction by its positive TRUE norm. Orthogonal projection is the Euclidean best approximation, so

$$\inf_{\widehat{\Delta}_h \in \ell} \frac{\|\Delta_h - \widehat{\Delta}_h\|_2}{\|\Delta_h\|_2} = \min_{x_h \in \ell} \|v_h - x_h\|_2 = \|(I - \Pi_\ell) v_h\|_2.$$

The coefficients are independent across horizons. Therefore

$$\inf_{\mathcal{A}_1(\eta)} E = \min_{\ell \in \mathrm{Gr}(1,d)} J_\eta(\ell), \qquad J_\eta(\ell) = \frac{1}{H} \sum_{h=1}^{H} \|(I - \Pi_\ell) v_h\|_2. \tag{36}$$

The minimum exists because the real projective space $\mathrm{Gr}(1, d)$ is compact and J_η is continuous in its orthogonal projector.

To normalize Equation 28, write $H_a q_a = \alpha_a q_a + r_a$. After division by $h\eta G_a$, the norm is $1 - \eta(h-1)\alpha_a/2 + O(\eta^2)$. Multiplication by its reciprocal cancels the parallel curvature term, yielding

$$v_h = q_a - \eta \frac{h-1}{2} r_a + O(\eta^2) \tag{37}$$

uniformly over the fixed horizon. The division is valid uniformly because $G_a \to G_0 > 0$ and the strict denominators are nonzero.

The admissible line $\operatorname{span}(q_a)$ has $J_\eta = O(\eta)$. If ℓ_η^* is any exact minimizer, the first summand and $v_1 = q_a$ give the localization bound

$$\operatorname{dist}(q_a, \ell_\eta^*) \le H J_\eta(\ell_\eta^*) = O(\eta). \tag{38}$$

The same argument covers every sequence with $J_\eta(\ell_\eta) = O(\eta)$, including asymptotically near-optimal sequences. It assumes neither continuity in η nor advance locality of the global optimum.

Choose the unit representative u_η of a localized line so that $u_\eta^\top q_a > 0$, and write $u_\eta = c_\eta q_a + s_\eta$ with $s_\eta \perp q_a$. Then $\|s_\eta\| = O(\eta)$, $c_\eta = 1 + O(\eta^2)$, and, exactly,

$$u_\eta = \frac{q_a + \eta w_\eta}{\|q_a + \eta w_\eta\|_2}, \qquad w_\eta = \frac{s_\eta}{\eta c_\eta} \perp q_a, \qquad \|w_\eta\|_2 = O(1).$$

No convergence of w_η is needed. For every bounded such sequence, Equations 37 and the projector formula give, uniformly over $h \le H$,

$$(I - u_\eta u_\eta^\top) v_h = -\eta \left(w_\eta + \frac{h-1}{2} r_a \right) + O(\eta^2).$$

The reverse triangle inequality handles even a vanishing leading vector, so

$$J_\eta(\ell_\eta) = \frac{\eta}{H} \sum_{j=0}^{H-1} \|w_\eta + j r_a/2\|_2 + O(\eta^2), \tag{39}$$

with a remainder uniform on the bounded set supplied by localization.

Let $\widehat{r}_a = r_a/\beta_a$ and decompose $w = \alpha \widehat{r}_a + w_\perp$, where $w_\perp \perp r_a$. Termwise,

$$\|w + j r_a/2\|_2 = \sqrt{(\alpha + j\beta_a/2)^2 + \|w_\perp\|_2^2} \geq |\alpha + j\beta_a/2|.$$

Thus $w_\perp$ cannot improve the objective and equality is possible with $w_\perp = 0$. Setting $m = -2\alpha/\beta_a$ reduces the minimum to

$$\frac{\beta_a}{2} \min_{m \in \mathbb{R}} \sum_{j=0}^{H-1} |j - m|.$$

The minimizers are the medians of $\{0, \ldots, H-1\}$, and direct summation on the two sides of a median gives

$$\min_m \sum_{j=0}^{H-1} |j - m| = \left\lfloor \frac{H^2}{4} \right\rfloor. \tag{40}$$

For even H, the entire interval $[H/2 - 1, H/2]$ consists of medians; uniqueness is not asserted.

Applying Equations 39– 40 to the localized exact minimizer gives the lower bound

$$\min_\ell J_\eta(\ell) \geq \frac{\lfloor H^2/4 \rfloor}{2H} \beta_a \eta - O(\eta^2).$$

Conversely, for any median m_H, the explicit line represented by

$$u_\eta^* = \frac{q_a - \eta(m_H/2) r_a}{\|q_a - \eta(m_H/2) r_a\|_2}$$

gives the matching upper bound. Since $\beta_a = \beta_0 + O(\eta)$, this proves Equation 10, including its uniform $O(\eta^2)$ remainder and ratio limit. In particular, an oracle line changes K1's constant but cannot change its $\Theta(\eta)$ order.

A.7 Two-step lower bound used in the synthetic checks

A simpler nonsharp bound supplies the diagnostic L_1 in Appendix A.13.1. Since $v_1 = q_a$ exactly and $v_2 = q_a - \eta r_a/2 + O(\eta^2)$, let $V_{12} = [v_1^\top; v_2^\top]$ and $\rho = v_1^\top v_2 > 0$ for small η. The eigenvalues of $V_{12} V_{12}^\top$ are $1 \pm \rho$, while $\|v_1 - v_2\|^2 = 2(1 - \rho)$. Thus, exactly before expansion,

$$\sigma_2(V_{12}) = \frac{\|v_1 - v_2\|_2}{\sqrt{2}} = \frac{\beta_0 \eta}{2\sqrt{2}} + O(\eta^2). \tag{41}$$

Every $\widehat{\Delta}_h \in \ell$ has relative error at least $\|(I - \Pi_\ell) v_h\|$. Keeping two horizons, using the ℓ_1/ℓ_2 inequality and then Eckart–Young (Eckart & Young, 1936) gives

$$E(\widehat{\Delta}; \eta) \geq \frac{\|V_{12}(I - \Pi_\ell)\|_F}{H} \geq \frac{\sigma_2(V_{12})}{H} =: L_1. \tag{42}$$

The remainder is independent of ℓ, so for sufficiently small η, $E \geq \beta_0 \eta/(4\sqrt{2}H)$ uniformly over arbitrary, possibly future-informed or discontinuous line choices and unrestricted coefficients. The sharp full-horizon result above strengthens this bound.

A.8 Historical top-two sufficiency

We prove Corollary 3 directly. Let $\mathcal{S}_{\mathrm{hist},2}(\eta)$ be the top-two right singular subspace of the actual prefix update matrix, with projector $\Pi_{\mathrm{hist},2}$. Scale that matrix by ηG_0 and index the rows by $j = 0, \ldots, P-1$:

$$A_j = q_0 - \eta j H_0 q_0 + O(\eta^2), \qquad A_0 = q_0 \text{ exactly.}$$

The rank-two comparison matrix with rows $q_0 - \eta j H_0 q_0$ is within $O(\eta^2)$ in Frobenius norm. Combining this fact with the rank-one residual calculation in Appendix A.2 gives

$$\sigma_1(A) = \sqrt{P} + O(\eta),$$
$$\sigma_2(A) = \eta\beta_0\sqrt{P(P^2-1)/12} + O(\eta^2),$$
$$\sigma_3(A) = O(\eta^2),$$

where absent singular values are treated as zero. Since $P \geq 2$ and $\beta_0 > 0$, the top-two projector is uniquely defined for small η, and Eckart–Young gives

$$\|A(I - \Pi_{\mathrm{hist},2})\|_F = \left(\sum_{k\geq 3} \sigma_k(A)^2\right)^{1/2} = O(\eta^2). \tag{43}$$

The exact first row therefore yields $\|(I-\Pi_{\mathrm{hist},2})q_0\| = O(\eta^2)$. Moreover, $A_0 - A_1 = \eta H_0 q_0 + O(\eta^2)$; projecting this identity and using Equation 43 rowwise gives

$$\|(I - \Pi_{\mathrm{hist},2})H_0 q_0\| = O(\eta).$$

The moving-anchor expansions

$$q_a = q_0 - P\eta r_0 + O(\eta^2), \qquad H_a q_a = H_0 q_0 + O(\eta)$$

transfer the capture rates to the anchor: since $(I - \Pi_{\mathrm{hist},2})r_0 = (I - \Pi_{\mathrm{hist},2})H_0 q_0 - (q_0^\top H_0 q_0)(I - \Pi_{\mathrm{hist},2})q_0 = O(\eta)$, they give

$$\|(I - \Pi_{\mathrm{hist},2})q_a\| = O(\eta^2), \qquad \|(I - \Pi_{\mathrm{hist},2})H_a q_a\| = O(\eta). \tag{44}$$

For the matched recurrence, set $\Pi = \Pi_{\mathrm{hist},2}$ in Equation 34. The leakage identity

$$H_a q_a - \Pi H_a \Pi q_a = (I - \Pi)H_a q_a + \Pi H_a (I - \Pi)q_a = O(\eta)$$

shows that its first- and second-order differences from LocalFull are both $O(\eta^3)$. Thus $z_h - y_h^{\mathrm{hist}} = O(\eta^3)$ uniformly, and

$$D_0(\mathrm{LocalFull}, \mathrm{HistSVD2}) = O(\eta^2), \qquad D_0(\mathrm{TRUE}, \mathrm{HistSVD2}) = O(\eta^2).$$

History alone selects this space, but its matched recurrence still uses the anchor gradient and Hessian. Noisy, finite-precision, changing-driver and AdamW histories remain outside this exact-history result.

A.9 Fixed and scale-dependent energy thresholds

Fixed threshold. For fixed $\tau \in (0,1)$, Equation 12 gives $M(\eta) = 1 - O_1(\eta) \to 0$. Choose η_τ so that $M(\eta) \leq 1 - \tau$ for $0 < \eta < \eta_\tau$. Then the leading squared-singular-value fraction of the actual prefix matrix satisfies $O_1(\eta) \geq \tau$. Because the selector in Corollary 6 minimizes over $r \geq 1$, it equals one. Equation 10 supplies the stated inadequacy for the current, future-strengthened $\mathcal{A}_1(\eta)$ class.

Constructive exact-history threshold. Under the nondegenerate assumptions of Theorem 4, define

$$T_r(\eta) = \frac{\sum_{j>r} \sigma_j(U_\eta)^2}{\|U_\eta\|_F^2}.$$

The motion theorem gives $T_1 = M = \Theta(\eta^2)$. For the normalized prefix $A = U_\eta/(\eta G_0)$, Equation 43 gives an $O(\eta^2)$ rank-two residual amplitude and $\|A\|_F^2 = P + O(\eta)$. Therefore $T_2 = O(\eta^4)$, including the case of a zero rank-two tail. Fix $\kappa \in (0,1)$ and set $\tau_\kappa(\eta) = 1 - \kappa M(\eta)$. For sufficiently small η, this threshold lies in $(0,1)$ and

$$T_2 \leq \kappa M < T_1.$$

Rank one fails the captured-energy threshold and rank two meets it, proving $r_{\tau_\kappa} = 2$. The matched recurrence on that historical space then has $O(\eta^2)$ error by Appendix A.8.

Merely imposing $1 - \tau = O(\eta^2)$ is insufficient: its coefficient can exceed the actual rank-one tail coefficient. The choice above uses only the prefix and makes the allowed tail smaller than measured M. It does not establish noisy-history stability, a finite-tolerance certificate or computational advantage.

A.10 MINIMUM DIMENSION AND SCOPE

A zero-dimensional anchored response predicts a and has normalized error one. The sharp one-dimensional result rules out $O(\eta^2)$ for every allowed common line, while both K2 and the exact historical top-two space attain that order. This proves Theorem 2. Combining it with Equation 12, whose leading coefficient is positive for $P \geq 2$ and $\beta_0 > 0$, gives

$$M = \Theta(\eta^2), \qquad \inf_{\mathcal{A}_1(\eta)} E = \Theta(\sqrt{M}), \qquad E_{K2} = O(M).$$

Free-offset affine lines are not covered (two points always lie on an affine line); nor are rotating lines, nonlinear manifolds, refresh, changing drivers, AdamW, growing windows, or changing metrics.

A.11 VARYING-DRIVER EXPANSION

This derivation proves the proposition in Section 6. Let the finitely many losses share a neighborhood of θ_0, with uniformly bounded Hessians and a common local Lipschitz constant. For $\lambda \in [0, 1]$ and $\eta > 0$, put $\rho = \lambda + \eta$ and $x_t = \theta_t - \theta_0$. Here $H_0 = \nabla^2 L_{\mathrm{ref}}(\theta_0)$ and $q_0 = -g_0/\|g_0\|_2$. With $G_0 = \|g_0\|_2$ and $d_t = \nabla L_t(\theta_0) - g_0$, define the row matrices

$$D[t,:] = -d_t^\top / G_0, \qquad K[t,:] = -(t-1)(H_0 q_0)^\top, \qquad t = 1, \ldots, P. \tag{45}$$

Let $C^{\mathrm{row}} = I_P - \mathbf{1}\mathbf{1}^\top/P$ center the temporal rows and $Q_0 = I - q_0 q_0^\top$ remove the leading motion direction. The matrices in Equation 18 are $D_\perp = C^{\mathrm{row}} D Q_0$ and $K_\perp = C^{\mathrm{row}} K Q_0$. A fixed finite induction gives $x_{t-1} = (t-1)\eta G_0 q_0 + O(\eta\rho)$. Taylor expansion of the mixed loss then yields, uniformly over $t \leq P$,

$$\frac{u_t^\top}{\eta G_0} = q_0^\top + \lambda D[t,:] + \eta K[t,:] + O(\rho^2). \tag{46}$$

Stacking the rows gives $U/(\eta G_0) = \mathbf{1} q_0^\top + E + O(\rho^2)$ with $E = \lambda D + \eta K$. The rank-one projection calculation in Appendix A.2 applies with Frobenius normal projector

$$N(E) = (I_P - \mathbf{1}\mathbf{1}^\top/P) E (I - q_0 q_0^\top). \tag{47}$$

Consequently the squared tail energy is $\|N(E)\|_F^2 + O(\rho^3)$, while $\|U/(\eta G_0)\|_F^2 = P + O(\rho)$. Division gives Equation 17 and the definitions of A, B, C in Equation 18. Cauchy–Schwarz gives $B^2 \leq AC$.

For $\lambda = 0$, write $r_0 = (I - q_0 q_0^\top) H_0 q_0$ and $j = (0, 1, \ldots, P-1)^\top$. Then $K_\perp = -(j - (P-1)\mathbf{1}/2) r_0^\top$, hence

$$\frac{1}{P}\|K_\perp\|_F^2 = \frac{\|r_0\|^2}{P} \sum_{k=0}^{P-1} \left(k - \frac{P-1}{2}\right)^2 = \frac{P^2 - 1}{12}\beta_0^2.$$

This establishes the fixed-driver specialization. The proof is joint-local in (λ, η) and its remainder constant may depend on the fixed losses, dimension, and window, but not on sufficiently small λ, η.

A.12 WHY THE FIXED DENOMINATOR GUARD CHANGES THE STRICT LIMIT

With a fixed $\gamma > 0$, the literal implemented quantity is

$$D_\gamma(\mathrm{LocalFull}, K1) = \frac{1}{H} \sum_{h=1}^{H} \frac{\binom{h}{2}\eta^2 G_a \beta_a + O(\eta^3)}{\gamma + h\eta G_a + O(\eta^2)}.$$

As $\eta \to 0^+$ at this fixed guard,

$$D_\gamma(\mathrm{LocalFull}, K1) = \frac{H^2 - 1}{6\gamma}\eta^2 G_a \beta_a + O(\eta^3), \tag{48}$$

where remainder constants can depend on γ. In particular, $D_\gamma/(\eta\beta_a) \to 0$; the first-order theorem cannot be stated for literal fixed-γ D_γ.

For a finite run with all TRUE displacement norms at least $m > 0$, each summand changes by relative amount at most γ/m, so $0 \leq D_0 - D_\gamma \leq (\gamma/m) D_0$. The retained implementation uses $\gamma = 10^{-30}$ and the guard-clean rule $m > 10^6\gamma$. All conditions pass, and separate FP64 evaluation of D_0 and D_γ gives identical reported values. Numerical agreement in this regime does not identify their different strict limits.

A.13 SYMMETRIC QUADRATIC AND FINITE-HORIZON PROPAGATION

For $0 < \varepsilon < 1/2$, take $\eta = 1$, $\theta_0 = 0$ and

$$L_\varepsilon(\theta) = \tfrac{1}{2}\theta^\top H_\varepsilon \theta + b^\top \theta, \quad H_\varepsilon = \varepsilon \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad b = (-1, 0)^\top. \tag{49}$$

At anchor zero, $q_1 = e_1$, $q_2 = e_2$. For a fixed integer $T \geq 2$, the prefix and forecast calculations give

$$1 - O_1 = \varepsilon^2 \frac{T^2 - 1}{12} + O(\varepsilon^3), \qquad D_0(\mathrm{TRUE}, K1) = \varepsilon \frac{T - 1}{4} + O(\varepsilon^2). \tag{50}$$

This varies the loss coupling with η fixed; it is a separate illustrative family from the fixed-loss small-step theorem. For Equation 49, H_ε has eigenvalues 0 and 2ε. Write $r = 1 - 2\varepsilon$. Diagonalizing $I - H_\varepsilon$ gives

$$u_j = \theta_{j+1} - \theta_j = \tfrac{1}{2}(1 + r^j, r^j - 1)^\top, \qquad \theta_t = \left(\frac{t}{2} + \frac{1 - r^t}{4\varepsilon}, -\frac{t}{2} + \frac{1 - r^t}{4\varepsilon} \right)^\top. \tag{51}$$

The matrix is positive semidefinite but the loss has a linear component in its nullspace and is not bounded below. The example concerns finite-horizon prediction, not convergence to a minimizer. All expansions below hold for fixed T; they are not uniform statements as $T \to \infty$.

For $j = 0, \ldots, T - 1$, $u_j = (1 - j\varepsilon, -j\varepsilon)^\top + O(\varepsilon^2)$. At $\varepsilon = 0$ the first temporal column is constant. Removing its leading component from the second column leaves $-\varepsilon(j - (T-1)/2) + O(\varepsilon^2)$. The small singular value therefore satisfies

$$\sigma_2(U_T)^2 = \varepsilon^2 \sum_{j=0}^{T-1} (j - (T-1)/2)^2 + O(\varepsilon^3) \tag{52}$$

$$= \varepsilon^2 T(T^2 - 1)/12 + O(\varepsilon^3). \tag{53}$$

Division by $\|U_T\|_F^2 = T + O(\varepsilon)$ proves the occupancy rate in Equation 50. The K1 recurrence is $c_{t+1} = (1 - \varepsilon)c_t + 1$, so $c_t = [1 - (1 - \varepsilon)^t]/\varepsilon$. Its first-coordinate error is $O(\varepsilon^2)$ and the omitted second coordinate is $-\varepsilon t(t-1)/2 + O(\varepsilon^2)$. Using $\|\theta_t\| = t + O(\varepsilon)$ gives relative error $\varepsilon(t-1)/2 + O(\varepsilon^2)$. Averaging $t = 1, \ldots, T$ proves Equation 50. Here D_0 uses the nonzero $\|\theta_t\|$; at fixed T these norms tend to the positive integers t. In particular,

$$\frac{D_0(\mathrm{TRUE}, K1)}{\sqrt{1 - O_1}} \longrightarrow \frac{\sqrt{12}(T-1)}{4\sqrt{T^2 - 1}},$$

a finite positive constant. K1 error itself tends to zero. The saved deterministic sweep covers $\varepsilon = 10^{-4}$ through 0.02. For $\varepsilon = 0.002$, $T = 24$, it gives update occupancy 0.9998080994, K1 AUE 0.0114530885, and endpoint error 0.0228462069; K2 and LocalFull are exact because their span is $\mathbb{R}^2$.

A.13.1 SYNTHETIC MINIMUM-DIMENSION VALIDATION

A separate deterministic CPU suite checks the complete theorem pattern using analytic FP64 gradients and Hessians. It uses $P \in \{4, 8, 16\}$, $H \in \{2, 8, 24\}$, and the fixed five-scale grid $\eta \in \{0.02, 0.01, 0.005, 0.0025, 0.00125\}$; every slope holds P, H and the loss fixed and uses all five scales. Motion nonoccupancy is computed from direct singular-value tail energy, and the two-step lower bound $L_1 = \sigma_2([v_1^\top; v_2^\top])/H$ uses direct SVD rather than subtracting nearly equal cosines.

The two-dimensional quadratic closes exactly under K2. The three-dimensional quadratic uses the reference matrix $[[2, 0.7, 0.2]; [0.7, 1.4, 0.1]; [0.2, 0.1, 3]]$ and $b = (-1, -0.3, 0.4)^\top$; its two-term Krylov space is not closed, so K2 is second-order-compatible but not exact. Adding the analytic cubic $(0.7/6)(u^\top x)^3$ and quartic $(0.4/4)(x^\top x)^2$ terms, with $u = (0.6, -0.2, 0.5)^\top$, gives the smooth nonlinear row. Across all 205 conditions, the direct identity $\sigma_2(V_{12}) = \|v_1 - v_2\|/\sqrt{2}$ agrees to 6.1×10^{-17} and K1 never violates L_1 at 10^{-12} tolerance.

Boundary cases behave as the theorem scope requires: a linear loss with $\beta_0 = 0$ is captured by one direction; $H = 1$ is captured exactly by the gradient line; a free-offset affine line fits the first two future points; and a rotating line can fit every point. The last two are deliberately outside the common anchored window-fixed class. Supplement S4 provides the complete condition and result tables and analytic code.

Table 1: Synthetic analytic ALL-5 slopes at $P = 8, H = 24$. "Exact" denotes a structural zero, which is not log-floored or fit. These are CPU mathematical checks, not LLM evidence.

Case	p_M	p_{E_F}	$p_{E_{K1}}$	$p_{E_{K2}}$	p_{L_1}
2D quadratic	2.013	exact	0.937	exact	0.988
3D quadratic	2.067	exact	1.065	1.912	1.070
Smooth nonlinear	2.064	2.463	1.060	2.370	1.067

Propagation within a common frozen model. Let $P = QQ^\top$, $y_h = Qc_h$ and $r_h = z_h - y_h$. When $g \in \mathrm{span}(Q)$, the two recurrences imply

$$r_{h+1} = (I - \eta H_a) r_h - \eta (I - P) H_a y_h, \qquad r_0 = 0.$$

Thus powers of $I - \eta H_a$ propagate omitted response; instantaneous energy alone does not bound forecast error. Separately, Equation 6 is exact without an affine-trajectory assumption. FP64 evaluation gives zero maximum vector residual and nonnegative scalar triangle slack in every condition. These algebraic checks do not verify local-model accuracy.

B Protocols and Information Budgets

Table 2: Information budgets for the fixed-driver predictors. Every method uses the common anchor and has a 24-step forecast horizon. History means the eight-update prefix; HVP denotes Hessian-vector products, and d is the selected B dimension. Future TRUE is used only for scoring.

Predictor	History	Anchor gradient	HVP	Response dimension
RELEX-style	yes	no	no	one per tensor
DMD2	yes	no	no	two
K1 / K2	no	yes	yes	one / two
Random-R2	no	yes	yes	two
HistSVD1 / HistSVD2	yes	yes	yes	one / two
LocalFull	no	yes	yes	d

B.1 Ranked objects and their interpretation

For the adjacent-update matrix U, define

$$O_k^{\mathrm{update}} = \frac{\sum_{i=1}^{k} \sigma_i(U)^2}{\|U\|_F^2}.$$

Replacing its rows by $\theta_t - \theta_0$ defines cumulative-displacement occupancy O_k^{disp}. Both depend on the window and coordinate metric and differ from the spatial matrix rank of an individual tensor. The main text's O_1 always means adjacent-update occupancy. Retaining k historical directions specifies a space; response dimension counts the coordinates evolved by the chosen recurrence. Neither dimension nor occupancy alone certifies that recurrence's accuracy. Forecast error must be evaluated for a declared driver, optimizer state, anchor, horizon and observable, and does not by itself establish task utility.

B.2 Fixed-driver SFT and numerical implementation

The loss is assistant-only causal language-model cross-entropy with shifted targets; non-assistant and padding tokens are masked. The fixed cohort contains 32 mathematical problem–solution examples from `siyanzhao/Openthoughts_math_30k_opsd`, with no ID or normalized-hash overlap with the development split used for model and protocol development. Each model uses the same eight batches, four examples per batch, with maximum sequence length 512 and a fixed model-specific serializer. The batch objective normalizes the summed supervised-token loss by the number of

supervised tokens. The same serialized inputs and normalization are used for the nonlinear trajectory and local derivatives.

The primary fixed-driver panel uses FP32 model forward and selected parameters, evaluation mode, zero LoRA dropout, fixed buffers, no resampling, no clipping, no scheduler, and no optimizer state. The adapter initialization seed is 314159; the numerical random-direction namespace uses seed 271828. Only two B tensors, final-block `q_proj` and `v_proj`, are trainable. Their concatenated dimensions are 24,576 (Qwen3), 32,768 (SmolLM2), 20,480 (Llama), and 32,768 (OLMo). The LoRA initialization has random A and zero B, $r = 8$, $\alpha = 16$; $\alpha/r = 2$. All other parameters, including A, are frozen.

The training $\eta = 0.03/\|g(\theta_0)\|$ remains fixed after the prefix. The prediction basis uses $g(\theta_8)$; it does not reuse the initial trajectory gradient. Matrix-free automatic differentiation supplies Hessian actions in the selected B space. K1 requires the action on q_1; K2 additionally uses the action on q_2. LocalFull applies the same frozen operator to its evolving selected-space displacement over the horizon. No dense Hessian is required. Numerical validation records finite-difference windows and bilinear Hessian-symmetry checks per condition. Thin row-Gram spectra and FP64 analysis of stored trajectories are used for energy and error summaries. LocalFull is computationally more expensive and serves as a diagnostic reference, not a proposed acceleration method.

Qualitative Qwen3-8B robustness study. This study uses the same selected-space SGD recurrence, four-example fixed drivers, prefix 8, horizon 24, and 0.03 initial-step target. Its batch roster comes from the separate fixed-driver robustness panel (internally denoted `Round12` in the data manifest) rather than the primary fixed-driver roster. The 8B backbone is stored and loaded in BF16, while selected B parameters, gradient/HVP accumulation, trajectory and forecast artifacts are FP32. Supervised-token numerators are streamed over examples with one common batch denominator. Eager attention and non-reentrant gradient checkpointing are used; execution enables training mode for checkpointing with dropout disabled. Thus it matches the scientific protocol at one step scale, with a different precision/memory implementation. It is not the separate full-FP32 exact-HVP scaling confirmation in Appendix C.5. Forecasts were generated before $\theta_9, \ldots, \theta_{32}$. The reported median prefix occupancy is 0.999828, with displayed condition values ranging from 0.999760 to 0.999907. Supplement S2 indexes all eight displayed occupancies and their original-precision source; Appendix H.2 states the separate exact-HVP 8B confirmation's reproduction limits.

History controls. The dynamic mode decomposition control DMD2 takes the top-two left singular vectors of the column-stacked prefix updates. In their coordinates x_t, it fits $x_{t+1} = Ax_t + b$ by least squares on the seven observed transitions, then integrates forecasts from θ_8. RELEX-style takes a rank-one temporal SVD of each tensor's eight cumulative deltas from θ_0 and fits its coefficient linearly in time, excluding the zero reference row from regression. Neither uses future TRUE states or derivative queries. Their coefficient dimensions count dynamical coordinates, not parameter storage.

HistSVD/K1/K2 fix the gradient, operator and recurrence to isolate space selection; DMD2 and RELEX-style have the different budgets in Table 2. Auxiliary nonlinear predictors add fitting choices and training costs. We therefore leave learned predictors such as NExt to a separate comparison with explicit training-data and computational budgets.

Random-R2. The isotropic Random-R2 control (internally denoted `P3-H1` in the released artifacts) uses 32 fixed seeds, integers 3100001 through 3100032, for its additional operator queries. The roster specifies the exact Gaussian vectors, orthogonalization, and condition assignment in dimension 40,960. Each random transverse basis uses $Q = [q_1, r_2]$ with $r_2 \perp q_1$, and exactly Equation 3. The same reference anchor/operator, η, and target trajectories are used. Future truth enters only scoring; the random-control design was specified after the corresponding structured-control outcomes were known. It is therefore a matched control rather than an independent discovery cohort. Projected operators are reconstructed from sufficient statistics and Hessian symmetry. The 32-direction distribution is not a search over all possible subspaces.

B.3 CHANGING-DRIVER PROTOCOL

The changing-driver response diagnostic evaluates a response-input-only formulation. It is distinct from the projected-state study in Appendix E.2. The main-text fixed-driver K1/K2 instantiate

the *Fixed-Driver Projected-State Local Predictor* (FD-PS-Kk), with the same-state unprojected selected-LoRA-B reference FD-LocalFull. The distinct schedule-conditioned response predictor is the *Changing-Driver Response-Restricted Predictor* (CD-RR-Kk). Its full schedule-conditioned reference is CD-SC-Full. With full batch-specific forcing and projection only on the response input,

$$\delta_{t+1} = \delta_t - \eta\big(g_{B_t,0} + H_{B_t,0} Q_k Q_k^\top \delta_t\big), \qquad \delta_0 = 0, \qquad \widehat{\theta}_t = \theta_0 + \delta_t.$$

Thus the displacement generally leaves the retained subspace, unlike the projected-state recurrence in Equation 3.

Table 3: Fixed- versus changing-driver definitions. Each local reference removes the corresponding restriction within selected B space.

Property	FD-PS-Kk	CD-RR-Kk
State	$c_h \in \mathbb{R}^k$, $\widehat{\theta} = a + Qc_h$	$\delta_t \in \mathbb{R}^d$; generally outside $\mathrm{span}(Q)$
Response	$Q^\top H_a Q$	$H_{B_t,0} Q Q^\top$
Forcing	projected $Q^\top g(a)$	full batch-specific $g_{B_t,0}$
Anchor	θ_8	θ_0
Truth / horizon	$\theta_9, \ldots, \theta_{32}$; 24	$\theta_1, \ldots, \theta_{32}$; 32
Step size	$0.03/\|g_B(\theta_0)\|_2$	$0.03/\|\,\mathrm{mean}_b\, g_b(\theta_0)\|_2$
Driver access	one known repeated batch	fixed future batch-ID schedule and cached anchor fields/actions
Future truth	none in construction	none in construction
Reference	FD-LocalFull	CD-SC-Full
Normalization	true displacement from θ_8	true displacement from θ_0

The changing-driver schedule study (internally denoted `B2` in released artifacts) contains eight balanced schedules per family, each with 32 updates over eight four-example batch units. Supplement S3 indexes the eight ordered schedule seeds and the source manifest specifying the complete batch order. Replacing the projected input by δ_t gives CD-SC-Full. The projection basis does not restrict the full forcing/output state, which is why CD-RR-Kk is not the same method as FD-PS-Kk. Cumulative-displacement occupancy in the prefix remains 0.965934 and 0.967174 for Llama and OLMo, respectively; adjacent-update medians are 0.529318 and 0.518759. These values refer to the response-input-only changing-driver protocol, not the common-start driver-regime study.

B.4 AdamW with causal optimizer state

Let $t_a = 8$ be the observed anchor step and $a = \theta_{t_a}$ its parameter vector. At state $s_{t_a} = (a, m_{t_a}, v_{t_a})$, let Ψ_{t_a+1} be one AdamW transition, and let E_θ extract the parameter component. Define

$$q_1 = \frac{E_\theta \Psi_{t_a+1}(s_{t_a}) - a}{\|E_\theta \Psi_{t_a+1}(s_{t_a}) - a\|}, \qquad q_2 \propto (I - q_1 q_1^\top) E_\theta D\Psi_{t_a+1}(s_{t_a})(q_1, 0, 0).$$

The memory lift is zero for basis construction, not for forecast propagation. Initialize $c_0 = 0$, $\widehat{\theta}_0 = a$, $\widehat{m}_0 = m_{t_a}$, $\widehat{v}_0 = v_{t_a}$. At $t = t_a + j + 1$:

$$\widehat{g}_j = g_a + H_a(\widehat{\theta}_j - a), \tag{54}$$

$$\widehat{m}_{j+1} = \beta_1 \widehat{m}_j + (1-\beta_1)\widehat{g}_j, \qquad \widehat{v}_{j+1} = \beta_2 \widehat{v}_j + (1-\beta_2)\widehat{g}_j^{\odot 2}, \tag{55}$$

$$\theta^{\mathrm{raw}}_{j+1} = (1 - \eta\lambda)\widehat{\theta}_j - \eta\, \frac{\widehat{m}_{j+1}/(1-\beta_1^t)}{\sqrt{\widehat{v}_{j+1}/(1-\beta_2^t)} + 10^{-8}}, \tag{56}$$

$$c_{j+1} = c_j + Q_k^\top(\theta^{\mathrm{raw}}_{j+1} - \widehat{\theta}_j), \qquad \widehat{\theta}_{j+1} = a + Q_k c_{j+1}. \tag{57}$$

Here $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda = 0$. LocalFull sets $\widehat{\theta}_{j+1} = \theta^{\mathrm{raw}}_{j+1}$ and otherwise uses the identical local gradient and optimizer transition. The full predicted m, v never use future realized moments or gradients. The condition-specific constant learning rates are fixed by matching the initial selected-parameter displacement norm to the corresponding SGD step, with no outcome-based retuning. The

exact rates and configuration hash are provided in the supplementary manifests. Despite the AdamW implementation, this panel has zero weight decay and a repeated fixed batch. Appendix E.4 combines this state-aware AdamW transition with a known shuffled-minibatch schedule in a same-anchor comparison.

B.5 ALL-LAYER NUMERICAL SCOPE

The all-layer numerical-scope study (internally denoted `C2` in released artifacts) expands selection to q/v B in every transformer block with A frozen: 327,680 coordinates for Llama and 524,288 for OLMo. The centered finite-difference operator is $[g(a+\epsilon v) - g(a-\epsilon v)]/(2\epsilon)$. Before all-layer forecasts, $\epsilon = 0.1$ was selected from an ordered set of candidates passing final-block validation against exact HVPs. Across the 16 final-block validation conditions, the median HVP cosine is 0.9999999515 and the maximum relative error is 0.00074193. This validates the chosen numerical window at final-block scope. The weaker all-layer transverse residual was not directly certified across finite-difference scales, so this study provides only a partial scope boundary.

B.6 CONDITION-LEVEL EXPORTS AND NOTATION

The supplement's evidence index lists the condition tables and their source/precision descriptions (Appendix H.2), retaining the fixed-driver, historical 8B, Random-R2, AdamW, changing-driver and all-layer cohorts separately. Scalar exports do not imply availability of trajectory or operator arrays (Appendix H). AUE is dimensionless; secondary columns use $\Delta_k = \mathrm{AUE}(Kk) - \mathrm{AUE}(\mathrm{LocalFull})$ and $\rho_k = \mathrm{AUE}(Kk)/\,\mathrm{AUE}(\mathrm{LocalFull})$ with the reference for the stated protocol. Tables exported from displayed sources retain their displayed precision; existing full-precision artifacts are identified separately in the index.

C SCALING EVIDENCE AND NUMERICAL RESOLUTION

The primary exact-HVP study tests accuracy orders and sharp constants. The separate finite-difference, exact-HVP 8B, and precision cohorts below provide supporting coefficient checks and numerical boundaries.

C.1 MULTI-SCALE EXACT-HVP ORDER VALIDATION

The validation uses Qwen3-1.7B and Llama-3.2-1B with four independently rostered four-example blocks per model, drawn from the same frozen `siyanzhao/Openthoughts_math_30k_opsd` files as the primary fixed-driver panel, with historical prompt hashes excluded before selection. The frozen roster and file hashes are indexed in Appendix H. Each model–block is reset to a common adapter initialization at each of five scales $\delta \in \{0.03, 0.015, 0.0075, 0.00375, 0.001875\}$, with $\delta = \eta\|g_0\|_2$. The driver is fixed within a condition; the optimizer is exact stateless fixed-step GD; $P = 8$, $H = 24$; model and selected-coordinate arithmetic are FP32; analysis is FP64; and each HVP is an exact autograd Jacobian-vector product (JVP) of the gradient. Only final-block query/value LoRA-B coordinates are selected, with LoRA-A and the backbone frozen.

The 40 conditions are nested scale observations within eight independent model–blocks. For a condition with anchor $a = \theta_P$, let $\Delta_h = \theta_{P+h}^{\mathrm{TRUE}} - a$ and $v_h = \Delta_h/\|\Delta_h\|_2$, $h = 1, \ldots, H$, in the selected-coordinate Euclidean norm. For a unit direction q, the error of the anchored, window-fixed line $a + \mathrm{span}(q)$ after optimizing its scalar coefficient separately at each horizon is

$$\mathcal{E}(q) = \frac{1}{H}\sum_{h=1}^{H}\sqrt{1-(v_h^\top q)^2}, \qquad E_{1\mathrm{D}}^* = \min_{\|q\|_2=1}\mathcal{E}(q). \tag{58}$$

The retained implementation uses

$$L_{\mathrm{pair}} = \frac{1}{H}\max_{1\le i<j\le H}\sqrt{1-(v_i^\top v_j)^2} \;\le\; E_{1\mathrm{D}}^* \;\le\; U_{\mathrm{med}} = \mathcal{E}(v_{12}), \tag{59}$$

Here $H = 24$, and v_{12} is the lower median-time direction. Both bounds use future TRUE displacements; the upper line is fixed through the common anchor while its scalar coefficient varies with h,

exactly as allowed by Theorem 2. A diagnostic numerical optimum lies between these bounds in every condition but is not used for slope fits. Figure 3 uses the tighter disjoint-pair lower bound in Appendix C.2; the frozen slope analysis below uses L_{pair} and the same U_{med}.

Slopes are fit by OLS in $(\log \eta, \log y)$ space within each block, followed by descriptive medians. Scale inclusion is applied independently to each of the seven fitted quantities $M, L_{\mathrm{pair}}, U_{\mathrm{med}}, E_{K1}, E_{K2}, E_{\mathrm{HistSVD2}}, E_{\mathrm{LocalFull}}$: a scale is included iff its scalar value is finite and strictly greater than 5×10^{-6}. Values at or below this resolution floor are neither replaced by the floor nor fitted; a fit additionally requires at least three included scales spanning a factor of at least four in δ. The threshold therefore acts on the dimensionless fitted ordinate (the normalized error/bound for the six error quantities, and M itself), not on a raw error numerator. It is a numerical fit-quality criterion, not a theoretical or predictor-acceptance threshold. It does not apply to the diagnostic numerical best-line optimum, raw displacement norms, or the $r_{0.99}$ rank calculation. HistSVD2 denotes the response recurrence in the prefix top-two right-singular space. The numerical histories use exact autograd gradients in floating point, while the sufficiency theorem concerns exact, noiseless history.

Table 4: Per-block OLS slopes for the multi-scale exact-HVP validation. The n tuple is ordered as M/lower/upper/K1/K2/HistSVD2/LocalFull and gives the number of scales actually included in each fit. Model and all-block summary rows are descriptive medians of block slopes, not fits pooling scale conditions; the independent unit is the model–block ($n = 8$).

Model	Block	n scales	p_M	$p_{\mathrm{1D-L}}$	$p_{\mathrm{1D-U}}$	p_{K1}	p_{K2}	p_{Hist2}	p_{Full}
Qwen3-1.7B	1	5/5/5/5/5/5/5	1.894	0.704	0.699	0.749	1.689	1.737	1.670
	2	5/5/5/5/5/5/5	1.886	0.731	0.727	0.766	1.682	1.766	1.658
	3	5/5/5/5/5/5/5	1.909	0.729	0.724	0.770	1.850	1.874	1.833
	4	5/5/5/5/5/5/5	1.903	0.750	0.746	0.784	1.819	1.863	1.792
	median	—	1.898	0.730	0.726	0.768	1.754	1.814	1.731
Llama-3.2-1B	1	3/5/5/5/5/5/5	1.990	0.977	0.977	0.981	1.805	1.848	1.749
	2	3/5/5/5/5/5/5	1.999	0.997	0.998	0.998	1.967	1.968	1.975
	3	3/5/5/5/5/5/5	1.979	0.965	0.965	0.970	1.881	1.928	1.902
	4	3/5/5/5/5/5/5	1.967	0.952	0.952	0.959	1.868	1.921	1.855
	median	—	1.985	0.971	0.971	0.976	1.875	1.924	1.879
All	8-block median	—	1.938	0.851	0.849	0.872	1.835	1.868	1.813

All eight blocks have $r_{0.99} = 1$ at every scale. Table 4 retains the model difference: Llama approaches the predicted orders more closely over this range. These finite-scale slopes do not recover limiting exponents, and an $O(\eta^2)$ guarantee requires neither a nonzero quadratic leading term nor an empirical slope exactly equal to two.

C.2 Sharp-constant calibration from retained vectors

For the same 40 conditions, normalize by $T = c_H^* \eta \beta_0 = 3\eta\beta_0$ (Theorem 2). A tighter lower bound is

$$L_{\mathrm{disj}} = \frac{1}{H} \sum_{i=1}^{H/2} \sqrt{1 - (v_i^\top v_{H+1-i})^2} \;\le E_{\mathrm{1D}}^* \le U_{\mathrm{med}}, \qquad H = 24.$$

Indeed, for unit v, w, the projective distance $d(v, w) = \|vv^\top - ww^\top\|_2 = \sqrt{1 - (v^\top w)^2}$ satisfies the triangle inequality; sum $d(v_i, v_{H+1-i}) \le d(v_i, q) + d(q, v_{H+1-i})$ over the fixed pairs. Thus $[L_{\mathrm{disj}}, U_{\mathrm{med}}]$ bounds the same common-line optimum without numerical optimization, as plotted in Figure 3.

All five scales have finite, positive denominators and both original bounds above 5×10^{-6}, so all 40 conditions enter this calibration. Table 5 gives every block at the smallest scale; its maximum normalized bracket width is below 2.71×10^{-5}. As δ decreases from 0.03 to 0.001875, the Qwen model-median bracket moves from $[0.434046, 0.435319]$ to $[0.951423, 0.951445]$, and Llama's from $[0.913789, 0.914292]$ to $[0.994074, 0.994077]$. These are bounds rather than exact optima. Qwen retains a 4.9% deviation from one; five scales cannot establish asymptotic convergence.

Analysis uses FP64 on the saved FP32 vectors. Evaluating the same sine distances as explicit projection residuals avoids cancellation; extended-precision post-processing changes the bounds by

at most 2.4×10^{-17} and does not assess the original FP32 trajectory error. No new trajectory or HVP is used.

Table 5: Sharp-constant calibration from the eight retained model–blocks at $\delta = 0.001875$. All bounds are divided by $c_H^* \eta\beta_0 = 3\eta\beta_0$. The last column is $(U_{\mathrm{med}} - L_{\mathrm{disj}})/(3\eta\beta_0)$; values are rounded, and model medians summarize four blocks, not independent scales.

Model	Block	L_{pair}	L_{disj}	U_{med}	Gap
Qwen3-1.7B	1	0.151567	0.948981	0.949008	2.70×10^{-5}
	2	0.150915	0.944859	0.944879	1.96×10^{-5}
	3	0.153293	0.959843	0.959867	2.43×10^{-5}
	4	0.152358	0.953865	0.953882	1.69×10^{-5}
	median	0.151962	0.951423	0.951445	—
Llama-3.2-1B	1	0.159038	0.995639	0.995641	2.33×10^{-6}
	2	0.159462	0.998317	0.998325	8.34×10^{-6}
	3	0.158542	0.992509	0.992513	3.71×10^{-6}
	4	0.158006	0.989444	0.989448	3.63×10^{-6}
	median	0.158790	0.994074	0.994077	—

C.3 Cross-model protocol and source identities

This separate supporting study has 12 four-example blocks per model, Qwen3-1.7B and Llama-3.2-1B, and uses finite-difference Hessian actions (Appendix C.7). It uses the pinned model/tokenizer revisions in Appendix H, seed 314159, rank-eight LoRA with $\alpha = 16$, random frozen A and zero-initialized trainable final-block q/v B. Model computation and updates are FP32, evaluation mode disables dropout, and saved arrays are analyzed in FP64.

The source is the frozen `sxiong/synthetic-math` snapshot in the manifest. The fixed block IDs are Qwen $\{03, 06, 10, 05, 07, 02, 12, 14, 09, 16, 11, 08\}$ and Llama $\{10, 04, 11, 02, 03, 12, 01, 08, 07, 16, 14, 13\}$; each uses only batch zero, its first four examples. The fixed system prompt, assistant-only serializer, maximum length 512, and supervised-token-normalized cross-entropy follow Appendix B. These blocks were unused in previous model executions, and estimands were fixed before outcome evaluation.

Each of the same five δ scales listed in Appendix C.1 resets to the common initialization. Repeated serialized batches use fixed $\eta = \delta/\|g_0\|_2$, $P = 8$, $H = 24$, without clipping, momentum, weight decay, scheduling, or step renormalization. Initial and anchor transverse curvatures are computed separately. The driver, coordinates and implementation are fixed within each block; prediction specifications and scales precede TRUE continuation, with condition identifiers binding the prediction and TRUE records.

C.4 Supporting estimands and descriptive summaries

ALL-5 fits OLS in log space within each model–block, using all five scales; all ordinates are positive and finite without log-floor substitution. Model exponents are medians of 12 block fits, with p_{RM} fitted directly from $\log R$ against $\log M$, not a ratio of aggregated exponents. TAIL-3 uses the three smallest scales as a descriptive sensitivity analysis, not an independently prespecified result. Interquartile ranges describe block spread, not confidence intervals. Within blocks δ is proportional to η; adjacent-scale halving ratios give 48 nested pairs per model. The coefficient diagnostics

$$C_M = \frac{M}{(\eta\beta_0)^2}, \qquad C_R = \frac{R}{\eta\beta_a}, \qquad C_{RM} = \frac{R}{\sqrt{M}}\frac{\beta_0}{\beta_a} \tag{60}$$

are evaluated per condition and then summarized over the 60 conditions per model, without fitting a multiplicative constant.

Condition-level results. The machine-readable supplement contains all 120 model/block/scale rows, including prediction/TRUE bindings, restriction values, and normalization diagnostics.

Table 6: Complete exponent summary for the three nonlinear scaling studies. These are the two supporting finite-difference model cohorts and the full-FP32 exact-HVP Qwen3-8B confirmation, separate from Table 4. Rows are medians of within-block fits. ALL-5 is primary within these studies; TAIL-3 is the fixed three-smallest-scale sensitivity summary.

Model	Scales	p_M	p_R	p_{RM}
Qwen3-1.7B	ALL-5	1.894	0.813	0.429
Qwen3-1.7B	TAIL-3	1.956	0.920	0.470
Llama-3.2-1B	ALL-5	1.995	0.928	0.466
Llama-3.2-1B	TAIL-3	1.998	0.902	0.451
Qwen3-8B (confirmation)	ALL-5	2.010	1.005	0.501
Qwen3-8B (confirmation)	TAIL-3	2.005	1.003	0.501
Theory	$\eta \to 0$	2	1	$1/2$

C.5 QWEN3-8B FULL-FP32 EXACT-HVP CONFIRMATION

The evaluation roster contains 32 unique examples in eight sequential four-example blocks, excluding the earlier 8B evaluation and design examples and with zero overlap with either earlier roster. The model revision, serializer, fixed LoRA-A tensor, selected final-block q/v LoRA-B names, five scales, $P = 8$, $H = 24$, and analysis estimands are held constant. The selected dimension is $40{,}960$, and the LoRA-A hash matches the historical protocol.

The memory-managed full-FP32 execution keeps checkpoint values in FP32 while invariant prefix layers are streamed through the GPU, and the final block, final norm, LM head, and adapters remain FP32. TF32, autocast, BF16/FP16 arithmetic, and quantization are disabled. The primary operator is the exact batch-four selected-space autograd HVP. A normalized central-FD arm with $\epsilon = 0.003$ is secondary and cannot replace the exact result. All 40 exact and FD predictions were saved before the corresponding TRUE continuations. Every condition is reported.

Table 7 retains every block fit, and Table 6 gives the model medians. The separate consistency diagnostic p_{ratio} fits $\log(R/\sqrt{M})$ against $\log \delta$ with theoretical reference zero; its medians are 0.0023 (ALL-5) and 0.0017 (TAIL-3). It differs from p_{RM}, the direct $\log R$ versus $\log M$ slope.

Table 7: Every Qwen3-8B confirmation block fit. Unadorned columns use ALL-5; "tail" uses the three smallest scales. p_{ratio} is defined in the text and has theoretical reference zero.

Block	p_M	p_R	p_{RM}	p_{ratio}	p_M^{tail}	p_R^{tail}	p_{RM}^{tail}	$p_{\text{ratio}}^{\text{tail}}$
01	2.0120	1.0160	0.5050	0.0100	2.0046	1.0053	0.5015	0.0030
02	2.0166	1.0138	0.5027	0.0055	2.0066	1.0055	0.5011	0.0022
03	2.0088	0.9919	0.4938	−0.0125	2.0053	1.0001	0.4987	−0.0026
04	1.9922	0.9878	0.4958	−0.0084	1.9968	0.9951	0.4984	−0.0033
05	2.0189	1.0238	0.5071	0.0144	2.0069	1.0083	0.5024	0.0049
06	2.0213	1.0006	0.4950	−0.0101	2.0128	1.0081	0.5008	0.0017
07	2.0029	1.0097	0.5041	0.0083	1.9996	1.0014	0.5008	0.0016
08	1.9877	0.9930	0.4995	−0.0009	1.9941	0.9950	0.4990	−0.0021

Table 8 tests the unfitted prefactors across all 40 conditions, retaining the full ranges. Median relative deviations from theory are $+0.535\%$, $+0.017\%$, and -0.230% for (C_M, C_R, C_{RM}).

Table 8: Qwen3-8B confirmation geometry-normalized coefficients over all 40 conditions. Theory is fixed by $P = 8, H = 24$ and is not fit to these outcomes.

Quantity	Median	Minimum	Maximum	Theory
C_M	5.278	5.0550	5.6043	5.2500
C_R	5.751	5.5929	5.8283	5.7500
C_{RM}	2.504	2.3742	2.5923	2.5095

Condition-level results. The machine-readable supplement contains all 40 exact-HVP condition summaries, including each scale, coefficient, LocalFull error, and paired FD value. K2 was outside this frozen estimand; this cohort supports neither K2 nor future-aware geometric-bound scale claims. Its artifact limitations and the resulting reproduction boundary are stated in Appendix H.2.

Exact-versus-FD operator comparison. The secondary FP32 finite-difference arm closely tracks the exact operator (Table 9). Median $R^{\mathrm{FD}}/R^{\mathrm{exact}}$ is 0.999975; median ALL-5 shifts are 1.44×10^{-5} for p_R and 7.17×10^{-6} for p_{RM}. Exact autograd remains primary regardless of this agreement.

Table 9: FP32 central-FD comparison with exact autograd HVP over all 40 confirmation conditions. Relative errors are vector ℓ_2 errors.

FD/exact diagnostic	Median	Minimum	Maximum
Hq cosine	0.999999659	0.999998921	0.999999799
Hq relative ℓ_2	0.000844268	0.000636840	0.001523416
$H_a q_a$ cosine	0.999999621	0.999999264	0.999999887
$H_a q_a$ relative ℓ_2	0.000907755	0.000475665	0.001248258
$R^{\mathrm{FD}}/R^{\mathrm{exact}}$	0.999974744	0.999369217	1.000298448

C.6 BF16 SMALL-SIGNAL PRECISION BOUNDARY

A separate BF16 8B scaling evaluation exhibited a small-signal precision floor. Direct FP64 SVD of the saved prefixes reproduced the measured motion values, while changing only the forward-compute precision on a fixed design cohort recovered quadratic motion under two FP32 paths (Table 10). The BF16 path flattened further on two smaller predeclared diagnostic scales. This is the signature of a small-signal precision floor, not evidence for a different 8B asymptotic law.

Table 10: Post-hoc fixed precision-diagnostic cohort results for motion. All three rows use the same examples and declared scales; the seven-scale column adds two smaller diagnostic scales. These are diagnostic, not confirmatory, fits.

Precision-diagnostic path	ALL-5 p_M	TAIL-3 p_M	Seven-scale p_M
Original BF16	1.2115	1.0547	0.9171
Final block FP32	1.9893	1.9955	1.9932
Full-FP32 staged	1.9889	1.9953	1.9932

BF16 also fails to resolve the finite-difference Hessian action on the predeclared epsilon grid (Table 11). Increasing epsilon crosses more quantization bins and reduces the spurious amplitude without validating the operator. In contrast, FP32 central differences at $\epsilon = 0.003$ have median relative error below 10^{-3} in both higher-precision diagnostic paths and in the full-FP32 exact-HVP confirmation. The diagnosis therefore supports a limited boundary statement: BF16 can obscure second-order motion signals and bias finite-difference Hessian actions in this small-step regime. It does not supply an additional estimate of the 8B scaling rates.

C.7 CROSS-MODEL NUMERICAL VALIDATION AND FINITE-STEP DEVIATIONS

The finite-difference Hessian-action approximation used in this cross-model study is

$$H_x^{\mathrm{FD}}(v) = \frac{\nabla L(x+\epsilon v) - \nabla L(x-\epsilon v)}{2\epsilon}, \qquad \epsilon = 10^{-3}. \tag{61}$$

Directions supplied to the finite-difference operator are normalized, and their original norms are reinstated in the resulting actions; LocalFull uses its raw displacement. Parameters and gradients in these calls are FP32. A finite difference approximates a linear Hessian action but is not literally the exact operator in Theorem 4; truncation, cancellation and FP32 update rounding can matter as the signal shrinks. These diagnostics do not supply a uniform exact-Hessian error bound for all evaluation conditions.

Table 11: Post-hoc BF16 central-FD HVP sweep against exact autograd on the fixed precision-diagnostic cohort. No epsilon on the declared grid validates the BF16 operator; $\epsilon = 0.012$ is merely the least inaccurate listed point.

ϵ	HVP cosine	Relative ℓ_2 error	$\beta_{\mathrm{FD}}/\beta_{\mathrm{exact}}$
0.0005	0.251	14.620	18.909
0.0010	0.059	6.136	8.026
0.0015	0.078	4.855	6.343
0.0030	−0.040	3.390	4.195
0.0060	0.277	1.984	2.602
0.0120	0.645	1.004	1.347

All states and spectra are finite. Across the 120 conditions, the largest relative first-step deviation is 4.143×10^{-8}; direct tail energy agrees with $1 - O_1$ within $\max(10^{-12}, 10^{-4}M)$; and every TRUE norm exceeds $10^6\gamma$. Direct singular-tail energy remains the primary motion quantity.

For the exact frozen operator, the horizon-two ratio

$$Q^{(2)} = \frac{\|\widehat{\theta}^{\mathrm{LocalFull}}_{P+2} - \widehat{\theta}^{K1}_{P+2}\|_2}{\eta^2 \|g_a\|_2 \beta_a} \tag{62}$$

equals one. The pooled median is approximately 1.064, but the model-specific medians are 1.007 (Qwen) and 1.535 (Llama), and the maximum is 5.154. The full by-scale summary below shows the increasing small-step Llama deviations. A pooled near-one median must not be taken as uniform validation. These short-horizon deviations and the Llama TAIL-3 exponent delimit finite-step agreement with the asymptotic result.

Table 12: Two-step normalized restriction, whose exact-operator value is one. All 12 blocks per model and scale are included. Large small-step deviations, particularly for Llama, are retained.

Model	δ	Median	Q25	Q75	Maximum
Llama	0.001875	3.7796	3.2915	4.5342	5.1540
Llama	0.003750	2.1635	1.9767	2.4090	2.7356
Llama	0.007500	1.5349	1.3805	1.6742	1.9703
Llama	0.015000	1.1902	1.1323	1.2676	1.4213
Llama	0.030000	1.0837	1.0635	1.1258	1.1856
Qwen3	0.001875	1.0160	1.0066	1.0614	1.1099
Qwen3	0.003750	1.0228	1.0106	1.0601	1.0727
Qwen3	0.007500	1.0038	0.9957	1.0118	1.0283
Qwen3	0.015000	1.0034	0.9984	1.0085	1.0149
Qwen3	0.030000	1.0023	0.9938	1.0058	1.0231

D Matched Fixed-Driver Evidence

The analyses in this section reuse the four-family, 32-condition fixed-driver panel unless a separate cohort is stated. Shared protocols are in Appendix B; each comparison keeps its stated information budget.

D.1 Exact-gradient matched-history control

The matched-history control uses the same 32 primary fixed-driver anchors, histories, operators, step sizes and nonlinear TRUE continuations. The exact saved gradient direction q_1 is reused without redefinition. Form $U_\perp = U(I - q_1 q_1^\top)$ and take its leading right singular vector r_H, with the frozen orthogonalization and sign convention. The comparison basis is $[q_1, r_H]$. Its transverse direction

maximizes uncentered prefix-update energy after matching the current forcing. It is fixed before its additional Hessian actions are queried and is propagated through Equation 3 for 24 steps.

K2 has lower total AUE in all 32 conditions (Table 14). Basis Gram checks and four prespecified $q_1^\top H_a q_1$ checks pass; direct ambient distances use $h = 1, 2, 4, 8, 16, 24$. K2's two-step containment follows from construction; its finite-horizon advantage beyond gradient capture is measured here. The supplement retains all 32 condition-level results. Table 13 separates direct vector restriction from total-AUE differences.

A future-aware two-step geometric reanalysis normalizes the first two TRUE displacements, forms $V_{12} = [v_1^\top; v_2^\top]$, and computes the future-aware lower bound $L_1 = \sigma_2(V_{12})/24$. It is positive in all 32 conditions, ranges from 2.37×10^{-5} to 2.78×10^{-4}, and does not exceed observed K1 error in 32/32. This is a geometric diagnostic using future TRUE information, not an online score or practical certificate. The optional full-horizon Frobenius lower bound ranges from 0.00112 to 0.00909; because SVD minimizes squared Frobenius residual, it is not presented as the exact optimizer of the average relative norm objective. The primary fixed-driver panel has one step scale and supplies no eta exponent.

Table 13: Family medians of normalized distances in the matched panel. The restriction column is LocalFull–K1, mismatch is TRUE–LocalFull, and the last column is the paired total-AUE excess of K1 over LocalFull.

Family	TRUE–LocalFull	LocalFull–K1	LocalFull–K2	AUE(K1)−AUE(Full)
Qwen3	0.009663	0.078631	0.005331	0.060900
SmolLM2	0.000593	0.011956	0.000216	0.011462
Llama	0.001732	0.026016	0.001194	0.023674
OLMo	0.000731	0.023679	0.000372	0.022679

Table 14: Family summaries for the exact-gradient matched-history panel. Entries are medians of total AUE against nonlinear TRUE; the final column gives the full range of the paired History−K2 AUE gap across the eight conditions in each family.

Family	n	History	K2	LocalFull	History−K2 range
Qwen3	8	0.015807	0.011624	0.009663	0.003483–0.005595
SmolLM2	8	0.000929	0.000601	0.000593	0.000151–0.003192
Llama	8	0.003027	0.001854	0.001732	0.000694–0.001923
OLMo	8	0.000989	0.000704	0.000731	0.000248–0.000573

D.2 Localization, history, and parameter mapping

Separate diagnostic denominators. Let $r_h = \widehat{\theta}^{K1}_{8+h} - \theta_{8+h}$ and $u_{8+h} = \theta_{8+h} - \theta_{8+h-1}$, with V_2 containing the top two right singular vectors of prefix updates as columns. The three diagnostics are

$$E_{q_2} = \frac{\sum_h (q_2^\top u_{8+h})^2}{\sum_h \|u_{8+h}\|^2}, \qquad R_{q_2} = \frac{\sum_h (q_2^\top r_h)^2}{\sum_h \|r_h\|^2}, \qquad C_{q_2} = \|V_2^\top q_2\|^2.$$

The first two sums run over the 24 future steps. Their denominators differ from each other and from AUE; prefix capture is a unit-direction projection. The numerical energy denominators use the same negligible guard as the analysis implementation. The family medians below compare q_2's realized future-update energy, share of K1's squared residual and prefix capture using these distinct denominators.

Family	Future energy (%)	Residual share (%)	Prefix capture (%)
Qwen3	1.9181	96.0220	98.7487
SmolLM2	0.0746	99.7410	99.8764
Llama-3.2	0.3470	99.2957	99.5617
OLMo-2	0.2874	99.8927	99.9549

These primary fixed-driver summaries apply to all four families. The supporting horizon/localization analysis reuses the separate fixed-driver robustness panel; it is not an additional independent replication.

Unequal-information comparisons. The complete primary fixed-driver family medians are:

Family	RELEX-style	DMD2	K1	K2	LocalFull
Qwen3	0.20106	0.018983	0.070563	0.011624	0.0096630
SmolLM2	0.027883	0.0011410	0.011834	0.00060100	0.00059300
Llama-3.2	0.061081	0.0039490	0.025937	0.0018540	0.0017320
OLMo-2	0.053041	0.0011600	0.023328	0.00070400	0.00073100

A separate fixed-driver robustness panel used eight separately sampled fixed batches in each of Llama and OLMo and likewise found K2 better than K1 and DMD2 in 8/8 conditions per family. Its median K2/LocalFull ratios were 0.992675 and 1.064016. This is historical corroboration in a related protocol, not an additional set of initialization seeds for the four-family panel. Neither its DMD2 nor RELEX-style comparison isolates basis quality from operator/propagation information.

D.2.1 Exact frozen-A B-to-W map

For every selected tensor, $\Delta W = (\alpha/r)BA = 2BA$. Apply this same linear map to TRUE, K1, and K2 and recompute the corresponding Euclidean quantities in concatenated selected effective-q/v weight space. Primary fixed-driver medians are:

Family	O_1 in B	O_1 in W	W K1 AUE	W K2 AUE
Qwen3	0.997547	0.997570	0.070153	0.011574
SmolLM2	0.999975	0.999975	0.011876	0.000606
Llama-3.2	0.999885	0.999886	0.025828	0.001856
OLMo-2	0.999907	0.999908	0.023210	0.000700

K2 has lower AUE in all 32 mapped conditions. This is not an isometry claim: the frozen A induces a different metric, and neither the selected effective weights nor the selected B tensors constitute all LLM parameters.

D.3 Complete Structured HistSVD Comparison

Table 15: Family medians of total AUE for the matched local predictors, with the median condition-level HistSVD2/K2 ratio. All eight conditions per family are included.

Family	K1	K2	HistSVD1	HistSVD2	LocalFull	SVD2/K2
Qwen3	0.070563	0.011624	0.15969	0.020276	0.0096629	1.7412
SmolLM2	0.011834	0.00060070	0.021395	0.0011851	0.00059290	2.0324
Llama	0.025937	0.0018541	0.046867	0.0040052	0.0017318	2.0667
OLMo	0.023328	0.00070370	0.042183	0.0012419	0.00073050	1.7980

HistSVD1 and HistSVD2 use the first one and two right singular vectors, respectively, of the uncentered matrix of eight adjacent prefix updates. Singular order is fixed; the largest absolute component in each vector has positive sign, with the first index breaking ties. Only prefix history enters basis selection. The comparisons use all 32 conditions of the primary fixed-driver cohort, not independent replication. Both subspaces receive the same gradient, operator, step size, anchor, 24-step recurrence, and true-displacement normalization.

Every HistSVD2 AUE is below K1 and HistSVD1, and above K2 and LocalFull; every HistSVD1 AUE exceeds K1 and LocalFull. Paired HistSVD2/LocalFull has pooled median 2.04, range 1.57–2.53; the median signed excess over LocalFull is $+1.36 \times 10^{-3}$. No equivalence margin, outcome threshold, or condition exclusion was used. SmolLM2 batch 8 has top-2 capture 0.9272 and HistSVD2 AUE 0.01187, and remains in every summary. Paired HistSVD2/K2 has pooled median 1.94 and

range 1.52–2.26. The signed excess AUE over K2 has pooled median $+1.30 \times 10^{-3}$, ranging from $+2.71 \times 10^{-4}$ to $+1.08 \times 10^{-2}$. These finite-scale differences do not contradict historical top-two sufficiency: Corollary 3 concerns the asymptotic error order, not equal finite-scale constants or equality with K2.

The supplement retains the full-precision condition-level results, paired gaps, and direction diagnostics for all 32 conditions; displayed values are rounded.

Table 16: Family medians of signed paired gaps $\mathrm{AUE}(\mathrm{HistSVD2}) - \mathrm{AUE}(X)$, absolute alignment of the first gradient and history directions, and squared top-2 projection energy of q_2. Here v_1 is the first prefix right singular vector, and V_2 contains the first two as columns. Every family contains eight conditions.

Family	$X = K1$	$X = K2$	$X = \mathrm{LocalFull}$	$\lvert q_1^\top v_1 \rvert$	$\lVert V_2^\top q_2 \rVert^2$
Qwen3	−0.05052	+0.008188	+0.009965	0.99598	0.9875
SmolLM2	−0.01015	+0.000565	+0.000592	0.99995	0.9988
Llama	−0.02198	+0.002151	+0.002282	0.99978	0.9956
OLMo	−0.02209	+0.000576	+0.000660	0.99982	0.9995

D.4 Geometric Comparison of Response Subspaces

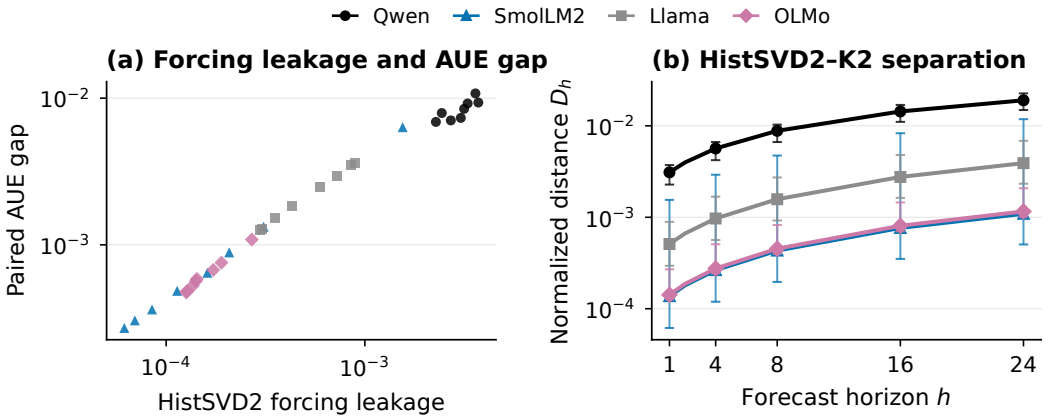


Figure 5: Geometric comparison across all 32 conditions. Left: historical forcing leakage versus the paired gap $\mathrm{AUE}(\mathrm{HistSVD2}) - \mathrm{AUE}(\mathrm{K2})$. Right: median normalized distances D_h between HistSVD2 and K2 predictions at every horizon, with thin min–max lines at $h = 1, 4, 8, 16, 24$. Ranges are not confidence intervals; associations do not uniquely explain causation.

D.4.1 Rotation invariance and finite-horizon identity

Fix a, g, H, η, an orthonormal Q, and an orthogonal R. Write $b = Q^\top g$ and $A = Q^\top HQ$. For $Q' = QR$, $b' = R^\top b$ and $A' = R^\top AR$. Starting at zero,

$$c'_{h+1} = R^\top c_h - \eta(R^\top b + R^\top ARR^\top c_h) = R^\top c_{h+1}, \qquad Q'c'_h = Qc_h.$$

Thus an orthogonal basis rotation preserves the exact ambient prediction; the comparison concerns the Krylov and historical subspaces, not their coordinate orientation.

With $P = QQ^\top$, the ambient recurrence is $z_{h+1}^P = z_h^P - \eta P(g + Hz_h^P)$, $z_0^P = 0$. For another orthogonal projector S, set $e_h = z_h^P - z_h^S$ and $A_P = I - \eta PH$. Exactly,

$$e_{h+1} = A_P e_h - \eta(P - S)(g + Hz_h^S), \tag{63}$$

$$e_h = -\eta \sum_{j=0}^{h-1} A_P^{h-1-j}(P - S)(g + Hz_j^S). \tag{64}$$

The assumptions are a common fixed operator, forcing, step size, and zero anchored state; symmetry of H is unnecessary. The identity separates propagated mismatch from new projected-field mismatch. It supplies no numerical amplification guarantee without stability and field information, and does not assume globally affine neural-network training.

D.4.2 Geometry and numerical validity

All 32 conditions include q_1, q_2, v_1, v_2, the anchor gradient, Hv_1, Hv_2, reduced actions, and all 24 forecasts of both methods.

Raw maximum-entry Gram errors are below 1.23×10^{-7} for K2 and 1.12×10^{-15} for SVD2. FP64 QR measures the saved spans. Rotation of the measured coordinates reproduces ambient vectors to maximum relative error 1.65×10^{-16}. Write $P_K = Q_K Q_K^\top$ and $P_S = Q_S Q_S^\top$. For orthonormal Q_K, Q_S, singular values $\sigma_1 \geq \sigma_2$ of $Q_K^\top Q_S$ define

$$\theta_i = \arccos(\mathrm{clamp}(\sigma_i, -1, 1)), \qquad d_{\mathrm{spec}} = \sin\theta_{\max}, \qquad d_F = \sqrt{2\sum_i \sin^2\theta_i}.$$

Normalized $d_F/2$ is the root-mean-square (RMS) sine of the two angles. An independent four-dimensional projector representation checks both distances. The smaller angle has pooled median 2.25×10^{-5} radians ($0.00129°$); the maximum angle has median $2.86°$ and reaches $15.65°$ in SmolLM2 batch 8. One principal direction is very close, but the complete subspaces need not be.

D.4.3 Forcing containment and closure

Since $g \parallel q_1$ ideally, forcing leakage is $\|(I - P_S)q_1\|$. Although SVD2 captures more than 99.998% of q_1 energy in every condition, forcing leakage remains nonzero, ranging from 6.16×10^{-5} to 3.73×10^{-3}. Using the separately saved gradient, the first-step vector difference agrees with $-\eta(P_S - P_K)g$ to maximum true-displacement-normalized residual 1.22×10^{-7}. That residual is far below the smallest first-step distance 6.16×10^{-5}. This identifies the initial forcing term to saved precision, not the fraction of the terminal AUE gap due to it.

For orthonormal Q, closure is

$$C(Q) = \frac{\|(I - QQ^\top)HQ\|_F}{\|HQ\|_F + 10^{-30}}.$$

With saved $V_{\mathrm{raw}} = Q_S R_S$, already saved actions transform as $HQ_S = (HV_{\mathrm{raw}})R_S^{-1}$; an orthogonal rotation checks invariance. The actions have FP32 evaluation precision; subsequent algebra accumulates in FP64. Reprojection reproduces the saved reduced SVD action to maximum relative discrepancy 1.64×10^{-8}. $C(Q_S)$ has pooled median 0.512, range 0.210–0.823. Qwen's lowest family-median $C(Q_S)$ coexists with the largest AUE gap: an unweighted column norm does not account for excitation. The supplement retains individual column leakages.

Table 17: Subspace geometry: median (full range) across 32 related conditions pooled or eight per family. Leakage is $\|(I - P_S)q_1\|$; its displayed values are multiplied by 10^3.

Scope	$\theta_{\max}$ (degrees)	Leakage (10^{-3})	$C(Q_S)$
Pooled	2.861 (1.048–15.65)	0.3066 (0.06157–3.725)	0.512 (0.2101–0.8226)
Qwen3	6.418 (5.137–7.504)	3.101 (2.278–3.725)	0.2572 (0.2101–0.4266)
SmolLM2	1.996 (1.205–15.65)	0.1374 (0.06157–1.55)	0.7383 (0.5546–0.7963)
Llama	3.783 (2.425–6.111)	0.5104 (0.2951–0.8914)	0.717 (0.5796–0.8226)
OLMo	1.217 (1.048–2.26)	0.142 (0.1263–0.2698)	0.4551 (0.4048–0.4694)

D.4.4 SAVED FINITE-HORIZON DIFFERENCES

Define

$$D_h = \frac{\|\widehat{\theta}^{\mathrm{HistSVD2}}_{8+h} - \widehat{\theta}^{K2}_{8+h}\|_2}{\|\theta_{8+h} - \theta_8\|_2 + 10^{-30}}.$$

Only saved forecasts are used. D_h increases at each of the 24 saved horizons in all 32 conditions; none was selected for monotonicity. It is a vector distance, not an AUE difference. Its pooled median rises from 3.07×10^{-4} at $h = 1$ to 2.46×10^{-3} at $h = 24$. Figure 5 shows family medians and full ranges.

Associations and what remains unexplained. Spearman is Pearson correlation of average ranks. Family centering subtracts each family's arithmetic mean from each raw variable before ranking the pooled residuals. No iid p-values are computed. Let $\Delta_{S2,K2} = \mathrm{AUE}(\mathrm{HistSVD2}) - \mathrm{AUE}(K2)$. The associations below support a partial interpretation involving mismatch and forcing noncontainment, not a causal decomposition. Both directions change; finite-precision actions, excitation, and propagation prevent attribution of the complete gap to a single term.

Table 18: Descriptive Spearman associations, with eight conditions per family. The d_{spec} correlations coincide with those for $\theta_{\max}$ in these data and are omitted here.

Metric	Target	Pooled	Centered	Qwen3	SmolLM2	Llama	OLMo
Forcing	$\Delta_{S2,K2}$	0.997	0.924	0.905	1.000	1.000	1.000
Forcing	D_{24}	0.999	0.991	0.952	1.000	1.000	1.000
$\theta_{\max}$	$\Delta_{S2,K2}$	0.918	0.874	0.810	1.000	0.976	0.952
$\theta_{\max}$	D_{24}	0.905	0.808	0.690	1.000	0.976	0.952

All original geometry, closure, horizon and correlation summaries, with full-precision condition data and metric definitions, are indexed as Supplement S1.

D.5 OCCUPANCY ASSOCIATION AND RESPONSE LEAKAGE

Let $M = 1 - O_1^{\mathrm{update}}$ and $G = D(\mathrm{LocalFull}, K1)$ use the common true-displacement denominator. The analysis gives M versus G Spearman 0.999 pooled, 0.941 raw family-centered, and 0.976–1 within families. The ordering is strong within this protocol and persists after family centering; it does not calibrate an error tolerance.

Saved reduced-action entries $a = q_1^\top H q_1$, $b = q_2^\top H q_1$ recover the first transverse norm to saved precision using the construction of q_2. Define $L_{\mathrm{abs}} = \eta b$ and $L_{\mathrm{angle}} = b/(\sqrt{a^2 + b^2} + 10^{-30})$. These scalar recoveries are not full H-actions and cannot supply K2 closure. The ideal two-step identity is

$$\mathrm{LocalFull}_2 - K1_2 = -\eta^2 \|g\| (I - q_1 q_1^\top) H q_1.$$

The saved-vector check has maximum relative residual 1.02×10^{-4}. L_{abs} uses first-step scale; it is not the normalized second-step error.

Absolute leakage adds some within-protocol information: for G, its family-centered Spearman correlation is 0.992, compared with 0.941 for M. For the signed K1–LocalFull AUE gap, the corresponding correlations are 0.957 and 0.890. The advantage is not uniform: pooled AUE ordering and Llama's within-family ordering slightly favor M. Angular leakage is weakly associated with both targets, with mixed signs across families. Thus magnitude and step size matter, but incremental value is partial. In the changing-driver schedule study, the M–G association becomes weak and negative (pooled -0.135, centered -0.288). That study changes several contract components, so it identifies no isolated driver or optimizer effect. No iid tests are used.

D.6 ADDITIONAL INITIALIZATION ROBUSTNESS

The original seed 314159 and two additional seeds are shown separately. The same eight batches are reused. All 32 additional cells have K1>LocalFull and K2<K1. Llama/OLMo each have three tested initializations; Qwen/SmolLM2 retain one main initialization. The indexed supplement retains

all 32 additional initialization cells; the table reports family–seed medians with the original seed for comparison, including medians where K2 exceeds LocalFull.

Table 19: Family/seed medians with the original seed for context. Δ_1 is the median paired K1-minus-LocalFull AUE gap.

Family	Seed	O_1	K1	K2	LocalFull	Δ_1
Llama	314159	0.999885	0.025937	0.0018541	0.0017318	0.023674
Llama	314160	0.999774	0.033412	0.0025218	0.0022321	0.030854
Llama	314161	0.999827	0.031487	0.0025279	0.0027400	0.028361
OLMo	314159	0.999907	0.023328	0.00070375	0.00073055	0.022679
OLMo	314160	0.999795	0.033687	0.0013826	0.0011752	0.032558
OLMo	314161	0.999793	0.034025	0.0014588	0.0014996	0.032305

D.7 Isotropic Random-Direction Control

Random-R2 is a matched isotropic generic-extra-coordinate control in the 40,960-dimensional space of the historical qualitative Qwen3-8B study. Conditional on q_1, expected squared alignment with any prescribed transverse unit direction is $1/(40960-1)$, making this a weak high-dimensional isotropic baseline. K2 beats all 256 condition-matched random controls, but these are 32 basis draws on each of eight existing trajectories, not 256 independent LLM runs. Random directions slightly improve K1 in 228/256 cases, yet their median relative AUE change is only -0.000355%, range -0.010399%–$+0.000483\%$. Pooled Random-R2 median AUE is 0.0322823, versus K2's condition median 0.00694490. Full condition ranges remain in the machine-readable supplement indexed from Appendix B.6. The sampled isotropic extensions yield near-K1 errors; HistSVD provides the main structured comparison.

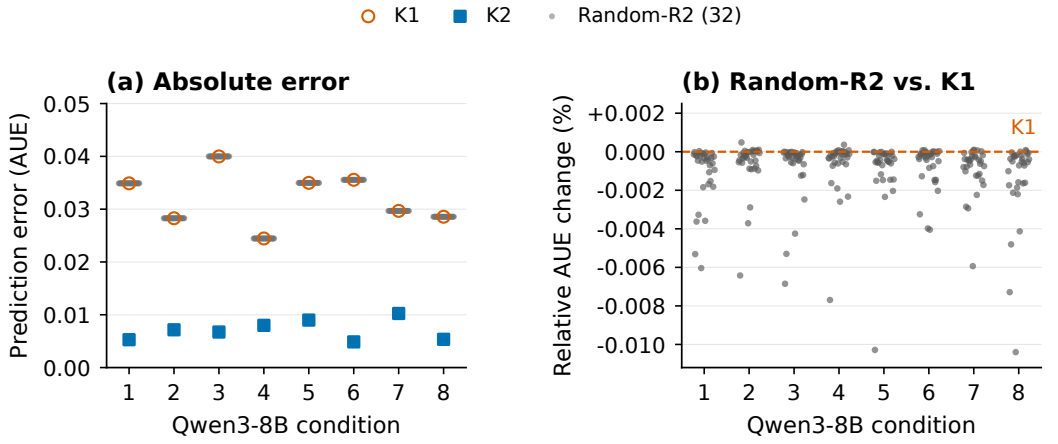


Figure 6: Isotropic controls on eight historical Qwen3-8B conditions. (a) Absolute AUE for K1, K2, and all 32 Random-R2 directions per condition; thin lines show the full random ranges at their actual scale. (b) Each direction's paired change $100[\mathrm{AUE}(\mathrm{Random\text{-}R2}) - \mathrm{AUE}(\mathrm{K1})]/\,\mathrm{AUE}(\mathrm{K1})$. The zero line denotes K1; negative values are improvements. Only horizontal positions are spread for visibility. The 256 directions reuse eight trajectories, not 256 independent runs.

In all eight historical Qwen3-8B conditions, K2 improves on K1 and remains close to LocalFull. Supplement S2 retains the complete eight-condition K1/K2/LocalFull comparison, original figure and occupancy list. This is a qualitative fixed-driver scale check; Appendix H.2 states the separate 8B artifact limitations.

E Driver and Optimizer Controls

E.1 Nonlinear varying-driver validation protocol

The validation has 240 conditions for each of Qwen3-1.7B and Llama-3.2-1B. Each model contributes eight schedule/data blocks. Within each block, 32 unique examples are partitioned into eight four-example batches; the first scheduled batch defines L_{ref}, and the remaining schedule defines L_t. Every block uses the complete grid

$$\lambda \in \{0, 0.0625, 0.125, 0.25, 0.5, 1\}, \qquad \eta = \delta/\|g_0\|,$$
$$\delta \in \{0.03, 0.015, 0.0075, 0.00375, 0.001875\}.$$

Thus the 480 grid cells are repeated conditions within 16 blocks, not iid replicates. The nonlinear update uses the exact mixed loss $(1-\lambda)L_{\mathrm{ref}} + \lambda L_t$ at its current state. Computation is FP32 with FP64 analysis, plain fixed-step SGD, and final-block q/v rank-eight LoRA-B coordinates; there is no momentum, clipping, decay, scheduling, or within-condition renormalization.

The data, schedules, grid, theory coefficients, primary subset, and analysis rules were hash-locked before outcomes. The primary region is $\lambda \le 0.25$ and the three smallest δ values: 12 cells per block, or 96 points per model. Within each model, the reported slope is ordinary least squares of $\log_{10} M_{\mathrm{true}}$ on $\log_{10} M_{\mathrm{quad}}$ over those 96 points. The log-ratio summary is the median of $|\log_{10}(M_{\mathrm{true}}/M_{\mathrm{quad}})|$ over the same points. All 480 cells remain reported, but $\lambda = 0.5$ and 1 are empirical boundaries and are excluded from the primary fit. Independent recomputation passes for every condition.

The saved block geometry also permits direct coefficient recovery. Across the 16 blocks, the medians are $A = 0.294364$, $C = 0.00193422$, and $A/C = 205.45$; model-median A/C is 25.29 for Qwen and 937.57 for Llama. Because these values use the declared λ/η parameterization, their scale is not compared across unrelated protocols. The cross coefficient satisfies $B < 0$ in 13/16 blocks and $B > 0$ in 3/16, so the signed cross term $2B\lambda\eta$ produces both partial cancellation and reinforcement. Comparing the magnitudes of the two pure contributions $A\lambda^2$ and $C\eta^2$, the driver contribution is larger in 47 of the 60 model/grid-cell medians and the curvature contribution in 13; at the smallest δ, the driver contribution is larger in every nonzero-λ cell. This supports the observed plateaus as a transverse-driver effect in these blocks, without a monotonicity claim.

Equating the driver and curvature magnitudes gives the leading predicted crossover

$$\delta_\times = \lambda G_0 \sqrt{A/C}.$$

Across blocks its median is 0.04268, 0.08535, and 0.17071 for $\lambda = 0.0625, 0.125, 0.25$, respectively. These are coefficient-based asymptotic crossover predictions, not fitted transition locations.

E.2 Driver-regime sensitivity

The study uses 16 Llama trajectories of 128 continuous updates and four predeclared starts $s \in \{0, 32, 64, 96\}$ per trajectory. At each start the repeated arm uses the first scheduled batch for the entire 32-step branch, while the shuffled arm follows its specified schedule. Both begin at the same saved θ_s and then take their own eight-update prefixes, producing different anchors. Each uses the trajectory's fixed initial-gradient step size, satisfying $\eta\|g_{\mathrm{first}}(\theta_0)\|_2 = 0.03$. The two arms therefore form common-start paired driver regimes, not a same-anchor intervention.

For each arm, let a be its own prefix endpoint. Evaluate all future batch-specific fields there and hold them fixed: $g_h = \nabla L_{B_h}(a)$ and $H_h = \nabla^2 L_{B_h}(a)$ for $h = 0, \dots, 23$. The basis Q_k uses the first future batch's negative gradient and, for K2, its first transverse curvature direction. The schedule-conditioned projected-state and full responses are

$$c_{h+1} = c_h - \eta(Q_k^\top g_h + Q_k^\top H_h Q_k c_h), \qquad c_0 = 0, \tag{65}$$
$$z_{h+1} = z_h - \eta(g_h + H_h z_h), \qquad z_0 = 0. \tag{66}$$

Predictions are interpreted as displacements from each arm's own anchor, and each arm uses its nonlinear TRUE displacement norm for scoring. Future batch IDs are known; future TRUE vectors enter scoring only. This projected-state contract differs from the separate response-input-only changing-driver study in Appendix B.3.

All 128 condition rows are included in the machine-readable supplement. The following summaries are means over the 64 predeclared windows per arm, preserving their nested trajectory/window interpretation. Repeated/shuffled mean update occupancies are $0.999881/0.429235$; direct K1 restrictions are $0.025628/0.714981$, and K2 restrictions are $0.001167/0.708701$. Direct distances and total AUE both separate restricted-response error from LocalFull mismatch. The supplement retains condition-level results and schedule metadata, including the trajectory/window hierarchy for this common-start comparison.

E.3 Controlled driver-by-parameterization occurrence study

This separate 2×2 study crosses repeated versus shuffled minibatches with final-block q/v LoRA-B-only versus A+B trainability under assistant-only AdamW LoRA-SFT. It uses Qwen3-1.7B and Llama-3.2-1B, four paired schedule replicates per model, and all four arms, for 32 trajectories. Each trajectory has 128 continuous AdamW updates with constant learning rate 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, and no warmup. LoRA has rank eight, $\alpha = 16$, and zero dropout; base computation is BF16, LoRA parameters and optimizer state are FP32, and occupancy analysis is FP64.

Within each model/replicate, arms share the initial model, adapter seed, data roster, and paired schedule. Repeated arms reuse the first four-example batch; shuffled arms follow the frozen eight-batch permutations. The four primary windows are steps 1–8, 33–40, 65–72, and 97–104. They yield 128 descriptive window observations, 16 per model/arm, but the trajectory is the replication unit. Later windows arise from separately evolved branches, not same-anchor interventions. Protocol, roster, schedules, windows, classification rules, and the effective-weight check were fixed before outcomes; recomputation matches all 32 trajectories exactly.

Table 20: Controlled 2×2 occurrence study. Each row contains four trajectories and 16 predeclared eight-update windows. Entries are window medians and descriptive counts; windows are not treated as iid replicates.

Model	Driver	Trainable LoRA	Median O_1	$O_1 \geq 0.99$
Qwen3-1.7B	Repeated	B only	0.999344	16/16
Qwen3-1.7B	Shuffled	B only	0.987376	0/16
Qwen3-1.7B	Repeated	A+B	0.986479	4/16
Qwen3-1.7B	Shuffled	A+B	0.978213	0/16
Llama-3.2-1B	Repeated	B only	0.999423	16/16
Llama-3.2-1B	Shuffled	B only	0.964854	0/16
Llama-3.2-1B	Repeated	A+B	0.985407	3/16
Llama-3.2-1B	Shuffled	A+B	0.942198	0/16

Both factors change compression prevalence. Paired effects use one median compression score per trajectory. The supplement retains primary-window observations and trajectory-level factorial summaries. Below, S-AB denotes the shuffled A+B cell.

E.4 Joint shuffled-AdamW same-anchor comparison

This schedule-conditioned study covers two models, four trajectories, and 32 non-overlapping windows with prefix $P = 8$ and forecast horizon $H = 24$. It uses changing minibatches and real AdamW optimizer state. Within each window, all predictors start from a common post-prefix anchor, use the same future schedule, and evaluate the batch-specific local fields at that common anchor. K1, K2, HistSVD1, and LocalFull are compared against the same nonlinear TRUE continuation. Appendix B.4 defines the state-aware AdamW transition.

LocalFull mismatch stays tiny, so the much larger restricted-response errors reflect response-space restriction rather than local-model failure. K2 improves K1 in 31/32 windows, but its median improvement is small. This supports persistence of the motion/response separation under the joint change, not rank-two sufficiency.

Table 21: Same-anchor shuffled-AdamW comparison summaries. Counts are over 16 windows per model; AUE entries are medians.

Model	median O_1	K1>Full	SVD1>Full	K2<K1	AUE K1	AUE K2	AUE Full
Llama-3.2-1B	0.961466	16/16	16/16	16/16	0.266230	0.266007	0.000278889
OLMo-2-0425-1B	0.970269	16/16	16/16	15/16	0.238193	0.237861	0.0000784799

E.5 Energy-selected history subspaces under shuffled A+B AdamW

Question and cohort. This study (artifact alias `B1`) tests whether retaining 99% of prefix energy also gives an adequate prospective response space. It separately re-executes the S-AB arm of Appendix E.3, whose original summaries lack parameter/moment states and full TRUE continuations. All 32 original prefix occupancies are reproduced within 5.64×10^{-12}; these are not new independent occurrence replications.

The S-AB optimizer, adapter, four-example roster, shuffled schedules, 128-update length, and four prefix windows are as defined there. Qwen3-1.7B and Llama-3.2-1B each contribute four trajectories: eight replication units and 32 nested windows, with $P = 8$, $H = 24$. The protocol preceded prediction and scoring, and no window was excluded after outcomes.

Rank and predictor definitions. For each prefix, $r_{99} = \min\{r : \sum_{j \le r} \sigma_j^2 / \sum_j \sigma_j^2 \ge 0.99\}$, using prefix history only. HistSVD-r_{99} is the top-r_{99} singular direction basis of that update history. MatchedHist-r_{99} has the same total dimension, explicitly includes the current anchor AdamW update, and fills the remaining dimensions from residual history directions. Prospective-r_{99} has the same total dimension and is constructed from linearly independent schedule-conditioned LocalFull local-response directions. LocalFull is the unrestricted schedule-conditioned local response and is a diagnostic reference. All three restricted bases use the same projected state-aware AdamW recurrence.

Prospective is a *response-informed positive control*: it uses the common anchor $\theta/m/v$, global step, known future minibatch IDs, anchor gradients, exact autograd HVPs and LocalFull predictions. It uses no nonlinear future parameters, realized gradients, losses or TRUE states. The secondary TRUE-TWO-STEP instead fits a secant through two real future AdamW states. Primary predictions were hash-locked before those states were opened; TRUE-TWO-STEP was then locked before opening the complete continuation for scoring. Per-model non-primary checks cover finite values, first-step agreement, HVP bilinearity, reduced/full-state agreement, and blindness. Table 22 summarizes all 16 windows per model; the supplement retains all 32 rows and eight trajectory medians.

Numerical and operator details. Responses use per-batch anchor gradients and exact reverse-over-reverse autograd HVPs on the training model. Bases use FP64 modified Gram–Schmidt with relative residual tolerance 10^{-10} and a canonical sign rule. The 32 windows require 2,944 HVPs and 768 shared gradient calls; construction timing synchronizes CUDA and excludes shared model loading.

Table 22: Energy-selected history-subspace results under shuffled A+B AdamW. Entries are medians over 16 windows per model. Mismatch is $D_\gamma(\mathrm{TRUE}, \mathrm{LocalFull})$; restriction is $D_\gamma(\mathrm{LocalFull}, K)$; AUE is $D_\gamma(\mathrm{TRUE}, K)$. All three use the common TRUE displacement denominator $\|\Delta_h\|_2 + 10^{-30}$ at each horizon, as in Section 2. TRUE-2 denotes TRUE-TWO-STEP. r_{99} entries give nonzero count distributions. Windows are nested and are not iid replicates.

	Prefix history		Local-model	Restriction to LocalFull			Prediction
Model	r_{99} counts	median O_1	mismatch	HistSVD	MatchedHist	Prospective	TRUE-2 AUE
Qwen3-1.7B	2:8, 3:4, 5:4	0.978213	0.066708	0.251739	0.218272	0.202570	0.234266
Llama-3.2-1B	3:4, 5:8, 6:3, 7:1	0.942198	0.044584	0.253649	0.220189	0.163216	0.278945

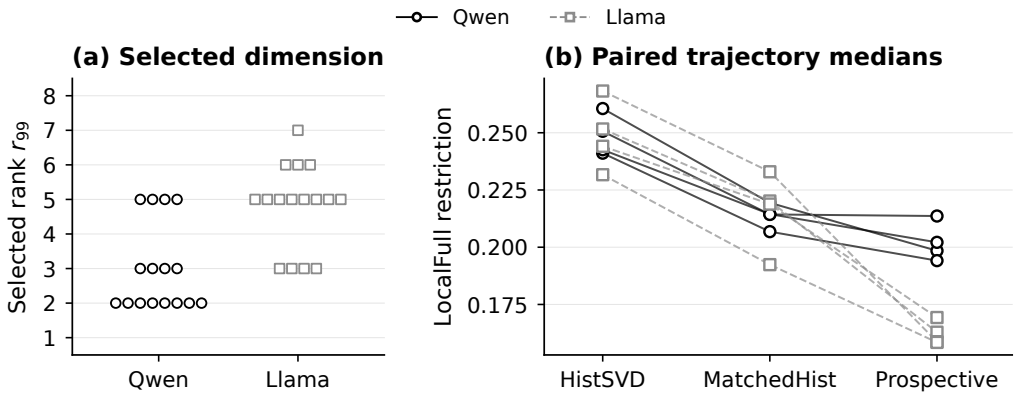


Figure 7: Energy-selected history spaces: all 32 windows from eight trajectories. (a) Prefix-only r_{99}; dots spread horizontally at each integer rank, without changing ranks. (b) Paired trajectory restriction medians; circles/solid lines denote Qwen, squares/dashed lines Llama. Prospective reduces restriction in every trajectory median, including after matching the current forcing, but uses additional operator information and compute.

HistSVD–Prospective restriction is positive in 32/32 windows and all eight trajectory medians, with median trajectory gap 0.061188. MatchedHist–Prospective is positive in 29/32 windows and all eight trajectory medians, with median gap 0.024178. Matching the current forcing improves HistSVD in every window; the three Qwen windows not favoring Prospective over MatchedHist are retained.

A separate FP64 check of total TRUE-relative errors, using nonzero TRUE norms without the implementation guard, gives the same win counts. Its median paired trajectory gaps are 0.062087 and 0.026356, respectively. These total-error gaps differ from the direct vector restrictions above; neither comparison treats nested windows as iid replicates.

Table 22 also shows that restriction exceeds LocalFull mismatch and that compression prevalence differs by model: 12/16 Qwen windows have $r_{99} \in \{2, 3\}$, whereas 12/16 Llama windows have $r_{99} \in \{5, 6, 7\}$. These paired trajectory-level results support better response alignment of the informed positive control at the selected dimensions. The released recomputation artifact checks ranks, errors, paired gaps, and timing.

E.6 Two-point local information: constructive evidence and limits

Definition, information contract, and algebra. For the frozen RELEX path R_h and the schedule-conditioned state-aware LocalFull path L_h, let $d_j = L_j - R_j$, $j \in \{1, 2\}$. The tested controls were

$$R_h^{\mathrm{corr1}} = R_h + hd_1, \qquad R_h^{\mathrm{corr2}} = R_h + d_1 + (h-1)(d_2 - d_1), \quad h = 1, \ldots, 24. \tag{67}$$

The discrepancies are used directly without normalization, projection, clipping, tuning, or learning. With the future schedule fixed, L_1 uses the anchor parameters, AdamW moments/global step, and first-batch anchor gradient; L_2 uses the second-batch anchor gradient and one batch-2 HVP in direction $L_1 - \mathrm{anchor}$. Neither uses future nonlinear states or outcomes.

The frozen RELEX forecast is affine in horizon: $R_h = R_1 + (h-1)(R_2 - R_1)$. Under exactly this condition,

$$\begin{aligned} R_h^{\mathrm{corr2}} &= R_h + (L_1 - R_1) + (h-1)\{(L_2 - R_2) - (L_1 - R_1)\} && (68)\\ &= L_1 + (h-1)(L_2 - L_1). && (69) \end{aligned}$$

CORR2 is therefore the two-point affine secant through L_1, L_2 in this study, not an inherently higher-dimensional predictor. The HVP is one way the second local point was constructed; direct controls are needed to attribute the gain.

Parameter-space direct controls. The direct-control cohort uses two models, 16 shuffled-AdamW trajectories, eight non-overlapping windows per trajectory, $P = 8$, and $H = 24$. Its original comparison retained the unfavorable CORR1 arm, which worsened RAW, while the two-point construction improved the tested history-only extrapolator. The unified $h = 3{:}24$ comparison uses all 128 windows and five methods: RAW; ONE-STEP, extrapolated from one prospective local point; NO-HVP TWO-STEP, which uses the two scheduled anchor gradients without the batch-2 HVP; FULL CORR2; and TRUE-TWO-STEP, the affine secant through the first two real schedule-conditioned AdamW steps. Future TRUE states enter none of the first four constructions. TRUE-TWO-STEP uses only its first two real steps for construction; steps 3–24 enter scoring only.

Table 23: Direct controls for AUE over $h = 3{:}24$. The first three numeric columns pool 128 windows; trajectory median is the median of 16 within-trajectory medians. Time is the same-run median described below. AUE values are rounded to four decimals; small paired differences are reported in the text.

Method	Median	Mean	p90	Traj. median	Time (ms)
RAW	0.4358	0.4851	0.7941	0.4345	13.7
ONE-STEP	0.3211	0.3297	0.4036	0.3205	33.2
NO-HVP TWO-STEP	0.2787	0.2858	0.3359	0.2823	60.3
FULL CORR2	0.2787	0.2858	0.3359	0.2823	101.1
TRUE-TWO-STEP	0.2787	0.2858	0.3359	0.2823	60.1

ONE-STEP reduces the RAW pooled median by 26.3%, and the second prospective point reduces the ONE-STEP median by a further 13.2%. The incremental HVP effect is negligible: FULL improves the NO-HVP pooled median by only 0.003%. The median paired trajectory difference (FULL minus NO-HVP) is -1.19×10^{-5}, with FULL winning 13/16 pairs. FULL and TRUE-TWO-STEP are still closer: their median paired trajectory difference (FULL minus TRUE) is $+3.40 \times 10^{-8}$, with FULL winning only 2/16 pairs. These small differences explain the apparent ties at the table's display precision. The supplement retains full-precision window results and pooled and trajectory-paired summaries.

Timing is a single RTX 4090 D microbenchmark on schedule-1/window-00 for each model, not 128 scientific windows. Method order rotates; one warmup per method/model is excluded, CUDA is synchronized, and 40 samples per method are retained. Cache creation is excluded; cache loading and CPU-to-GPU transfer are included. TRUE-TWO-STEP and NO-HVP are respectively 40.6% and 40.3% cheaper than FULL at the median. These controls support prospective two-point information, not a unique HVP or curvature advantage.

Functional evaluation. The separate functional protocol, data, and code were frozen before reading outcomes. All 128 windows are evaluated at $h = 8, 16, 24$ on 64 deterministic, unique OpenThoughts examples disjoint from the correction cohort. The primary endpoint is assistant-token-averaged Kullback–Leibler (KL) divergence, $\mathrm{KL}(p_{\mathrm{TRUE}} \| p_{\mathrm{PRED}})$, under the same teacher-forced contexts. Two exact TRUE states per model and the corresponding RAW/CORR2 injections give 12/12 state-reconstruction checks with zero logit difference.

Table 24: Pooled window medians of KL and absolute negative log-likelihood (NLL) gaps; parentheses give trajectory-level CORR2 wins.

h	RAW KL	CORR2 KL	Reduction (wins)	RAW NLL gap	CORR2 NLL gap
8	8.1695×10^{-7}	1.9758×10^{-7}	75.81% (16/16)	0.000544	0.000315
16	3.3151×10^{-6}	1.7376×10^{-6}	47.58% (16/16)	0.001313	0.001059
24	7.9027×10^{-6}	5.3097×10^{-6}	32.81% (15/16)	0.002231	0.001710

The absolute KL scale is small, 10^{-7}–10^{-5} nats/token in pooled medians. This frozen check shows closer fixed-context distributions and NLL for CORR2 than RAW; it is not a task-accuracy or free-running generation test. NO-HVP and TRUE-TWO-STEP did not receive the same functional evaluation, so the functional comparison is restricted to RAW versus CORR2.

E.7 PROSPECTIVE RESPONSE-AWARE SELECTIVE EXTRAPOLATION

This selective-extrapolation study uses a separate 1,024-example cohort, Llama-3.2-1B and OLMo-2-0425-1B, and four shuffled schedules per model. Its eight trajectories each contribute eight non-overlapping windows, giving 64 conditions with $P = 8$ and $H = 24$. Predictions and selection rules were locked before TRUE continuation; no future TRUE quantity enters either score. Motion uses $S_{\text{motion}} = 1 - O_1$. The truth-free two-step response score is

$$S_{\text{response}} = \frac{1}{2} \sum_{h=1}^{2} \frac{\|\text{LocalFull}_h - \text{HistSVD1}_h\|_2}{\|\text{LocalFull}_h - \text{anchor}\|_2 + 10^{-30}}, \tag{70}$$

where lower scores mean safer selections. Each score ranks windows separately within each trajectory at fixed 25%, 50%, 75%, and 100% coverages, with no fitted threshold. The primary target is 24-step HistSVD1 AUE.

Table 25: Pooled selective risk for the frozen motion and response rankings.

Coverage	Motion median	Response median	Motion p90	Response p90
25%	0.359644	0.327622	0.407856	0.371932
50%	0.359354	0.337852	0.399978	0.377084
75%	0.359509	0.353442	0.413596	0.391667
100%	0.365175	0.365175	0.586137	0.586137

At 50% coverage, response wins 7/8 trajectory pairs and lowers the pooled median by 5.98351%; area under the risk-coverage curve (AURC) also favors response in 7/8 trajectories. Reusing the frozen rankings gives descriptive medians of 0.438177 versus 0.429643 for the motion versus response selections on the RELEX-style target, and 0.540199 versus 0.529006 on DMD2.

Median score times are 0.00815894 seconds per window for motion and 0.120685 seconds for response, versus 0.403119 seconds for the 24-step TRUE continuation. The response score uses two gradient evaluations, three exact selected-space HVPs, and zero JVPs. The added operator information and compute provide a modest selective-risk improvement, largest at lower coverage and narrowing as coverage increases. This is neither an equal-compute/equal-information comparison nor a general extrapolation certificate or acceleration result.

F PRACTICAL FORECASTING BOUNDARIES

These diagnostics reuse frozen shuffled A+B AdamW trajectories or protocol-fixed anchors outside the fixed-loss exact-GD theorem. They test forecasting limitations.

F.1 FINITE-BUDGET HORIZON BOUNDARY

For each retained window, let t_a be its anchor step and $a = \theta_{t_a}$ its anchor parameter vector. The shuffled A+B AdamW finite-budget metric for LocalFull is the endpoint quantity

$$E_h^{\text{LF}} = \frac{\|\widehat{\theta}_{t_a+h}^{\text{LF}} - \theta_{t_a+h}^{\text{TRUE}}\|_2}{\|\theta_{t_a+h}^{\text{TRUE}} - a\|_2}. \tag{71}$$

The Euclidean norm uses flattened, sorted final-block query/value LoRA A+B factor coordinates. There is no denominator guard; all retained denominators are finite and nonzero. The 0.10 budget applies to E_{24}^{LF}, not AUE or a horizon average. A window passes iff $E_{24}^{\text{LF}} \leq 0.10$, and a trajectory passes iff at least three of its four fixed windows pass. There are eight independent trajectories (two models times four trajectories). LocalFull satisfied this predeclared trajectory-level criterion in 0/8 trajectories. In this regime, rank-attribution failures cannot be cleanly separated from local-model error. Accordingly, this analysis does not report a functional endpoint.

Table 26: LocalFull horizon sensitivity. "Median error" is the pooled median of the 32 window-level endpoint errors E_h^{LF}, not AUE or an average over a trajectory. At a displayed horizon, a window passes iff $E_h^{\mathrm{LF}} \leq 0.10$ and a trajectory passes iff at least three of its four windows pass. Dashes mark horizons for which the trajectory pass count is not used here; only $h = 24$ is the predeclared primary endpoint.

Horizon	Median endpoint error	Trajectories passing 0.10
1	7.76×10^{-7}	—
2	0.00216	—
8	0.01561	8/8
16	0.06519	7/8
24	0.16007	0/8

The increase is horizon sensitivity, not a formal crossover; $h = 24$ remains the primary endpoint. Despite sharing coordinates and windows with Appendix E.5, these numbers are not directly comparable: that study averages 24 guarded horizon components (AUE), summarizes 16 windows per model, and includes LocalFull–restricted-predictor errors. Here the metric is unguarded TRUE–LocalFull endpoint error, pooled over 32 windows, with LocalFull alone tested against the trajectory pass rule. Different horizons, targets, and aggregation remain even if the 10^{-30} AUE guard is numerically negligible.

F.2 Finite-budget feasibility and candidate sets

The post-hoc feasibility analysis tested causal HistSVD and MatchedHist predictors at ranks 1–8 and horizons 8, 16, and 24. None reached the 0.10 parameter budget over the tested windows. At $h = 8$, LocalFull passed 32/32 windows while every tested causal reduced rank failed. The least-error fixed reduced entry, MatchedHist rank 8 at $h = 8$, had median error 0.1425, passed 0/32 windows, and had median net saving -1.508, i.e., cost about $2.51\times$ direct continuation under the measured convention. Thus no practical accuracy–cost window was found for this predictor family and protocol; this does not rule out all low-rank predictors.

The TRUE-labelled causal oracle searched LocalFull, HistSVD, and MatchedHist (but not the unequal-information Prospective control), and its 32/32 $h = 8$ passes were obtained by permitting fallback to LocalFull. The best fixed *reduced* comparison searched only HistSVD and MatchedHist. Because these candidate sets differ, the reported 0.004 difference in saving is not interpreted as headroom for an adaptive reduced predictor.

F.3 Same-batch secant diagnostic

The following secant, response-replay, teacher-forcing, and future-driver diagnostics share four protocol-fixed anchors (Qwen and Llama replicates 1 and 2), with horizon $h = 8$. With $s = a - \theta_{t_a-1}$ and the first future batch as reference, the same-batch secant uses $y_{\mathrm{same}} = g_{b_0}(a) - g_{b_0}(\theta_{t_a-1})$; the cross-batch control also changes the batch. Against the diagnostic action $H_{b_0}s$, the median cosine improves from 0.0489 cross-batch to 0.8878 same-batch, and the relative residual decreases from 33.96 to 0.5269. The remaining residual is substantial, so this is an improved local directional estimate, not exact response recovery.

Table 27: Median $h = 8$ terminal TRUE-relative parameter error over the four diagnostic anchors. The same-batch rank-one gradient surrogate recursively evolves AdamW state.

Driver	Constant gradient	Same-batch secant
Fixed	0.08323	0.09295
Shuffled	0.74788	0.76209

Thus matching the batch improves the local secant estimate but does not improve median checkpoint rollout error over the constant-gradient baseline in either intervention.

F.4 Oracle response-change replay and response/update geometry

For the fixed-reference continuation, form the $8 \times d$ gradient-response matrix with rows $G_j = g_{b_0}(\theta_j^{\mathrm{fixed}}) - g_0$, $j = 1, \ldots, 8$, and let U_k contain its top k right singular vectors. The oracle replay uses

$$\widehat{g}_j = g_0 + U_k U_k^\top \left(g_{b_0}(\theta_j^{\mathrm{fixed}}) - g_0 \right) \tag{72}$$

inside the exact AdamW recursion. Consequently, k counts retained *response-change* directions in addition to the affine baseline g_0; it is not the total gradient-representation dimension used by the main theorem. The basis and coefficients use the complete future fixed-reference response sequence, so the replay is an oracle diagnostic.

Median $h = 8$ terminal errors over the four anchors are 0.01639, 0.002619, 0.001222, and 0.001060 for $k = 1, 2, 3, 4$. The largest error reduction is from one to two retained response-change directions, and $k = 2$ improves all 4/4 anchors. The extra direction helps in this window, but this does not re-verify the theorem's minimum dimension or establish two-direction necessity under the 0.10 budget: $k = 1$ already satisfies it.

For comparison with G, the effective-update-response matrix has rows $\Delta\theta_j - \Delta\theta_1$ along the same fixed-reference AdamW path. For either matrix, effective rank is

$$r_{99} = \min \left\{ r : \frac{\sum_{i \le r} \sigma_i^2}{\sum_i \sigma_i^2} \ge 0.99 \right\},$$

using the Frobenius/SVD squared-energy metric. The median r_{99} across the four anchors is 2 for gradient response and 1 for effective-update response. Parameter updates are therefore more compressible under this descriptive AdamW metric; it is not the main theorem's motion rank.

F.5 Teacher-forcing ladder

The same rank-one secant surrogate is evaluated with a three-level ladder. TF0 uses the true parameter and true AdamW moment/variance state for the one-step map; TF1 uses the true parameter at each step but recursively evolved surrogate moment/variance states; TF2 recursively evolves parameters and both optimizer states. At step 8, median instantaneous update-relative errors are 0.06667, 0.19632, and 0.19629 for TF0, TF1, and TF2.

For this surrogate, four-anchor protocol, and horizon, releasing recursive AdamW moment/state evolution causes the larger increase; subsequently releasing parameter free-running adds essentially no further step-8 increase. The ladder does not separately identify the first moment, second moment, or preconditioner.

F.6 Oracle future-driver intervention

Let b_0 be the first future reference batch, b_j the scheduled shuffled batch, and θ_j^{true} the future true shuffled state. The retained decomposition is

$$R_j = g_{b_0}(\theta_j^{\mathrm{true}}) - g_{b_0}(a), \tag{73}$$

$$F_j = g_{b_j}(\theta_j^{\mathrm{true}}) - g_{b_0}(\theta_j^{\mathrm{true}}), \tag{74}$$

$$g_{b_j}(\theta_j^{\mathrm{true}}) - g_{b_0}(a) = R_j + F_j. \tag{75}$$

Across four anchors and eight future steps, the median $\|F_j\|/\|R_j\|$ is 9.8006 and median $\cos(F_j, R_j)$ is 0.0876.

The oracle intervention adds the exact F_j sequence to the same-batch secant rollout. Because each F_j is evaluated at the unavailable future true state, the intervention uses future information and is not deployable. It reduces median shuffled $h = 8$ terminal error from 0.76209 to 0.09262. Unmodeled driver forcing is a substantial error source here, but its magnitude does not establish intrinsic unpredictability or irreducible noise.

F.7 Supporting horizon and locality evidence

These historical panels use separate cohorts and, where stated below, a different objective. They are not pooled with the primary fixed-driver panel.

Horizon accumulation. In the supporting horizon/localization diagnostic, K1 and K2 are nearly identical at the first forecast step, as expected because both include the anchor gradient. The median endpoint-error difference $e_{K1}(h) - e_{K2}(h)$ grows to approximately 0.0456 (Llama) and 0.0452 (OLMo) at $h = 24$. These are endpoint errors at that horizon, not 24-step AUE differences.

Step scale and refresh. The supporting step-scale and refresh diagnostic used one Qwen checkpoint and a static four-prompt driver under its historical fixed-teacher and recomputed-teacher gradient-field conventions, rather than the later assistant-only primary fixed-driver SFT objective. For fixed-teacher mode, frozen-local step-32 trajectory errors increased from 0.036% to 0.338% to 2.144% as $\eta\|G_0\|$ increased from 0.003 to 0.01 to 0.03. At the largest scale, the respective step-32 errors were 2.144% for the frozen predictor, 1.751% for refreshed forcing with the old Jacobian, and 0.825% for refreshed forcing and Jacobian. Refresh also resets at the true step-8 checkpoint. Only the two refreshed arms hold that state reset fixed while comparing old versus refreshed response. The numbers support local drift dependence, not a unique causal explanation or a new forecast sufficiency threshold.

Driver-conditioned local fields. The supporting driver-conditioned local-field diagnostic uses that historical Qwen setup to forecast the same B-driven trajectory with (G_A, J_A), (G_A, J_B), (G_B, J_A), and (G_B, J_B). At the first step the Jacobian acts on zero displacement, so forcing mismatch alone produces the A-to-B error (approximately 0.9634). The B-forcing arms match the first step and satisfy that study's declared criterion through step 32; substituting only J_B fails from step 1. The result concerns one checkpoint and two drivers and does not identify population-level batch geometry.

Nearby anchors. A saved nearby-anchor reconstruction at mean-field, batch-1, and batch-8 reference choices gave prefix update occupancies 0.9996, 0.9997, 0.9998 for Llama and 0.9998, 0.9999, 0.9999 for OLMo. This is limited sensitivity evidence for that historical reconstruction, not a new anchor sweep of the primary fixed-driver panel or an invariant-dimension claim.

G Output-Space Functional Evaluation

This evaluation compares the forecast endpoints at $h = 24$ (training step $P + H = 32$), not an average over forecast horizons. For an evaluation example i and shifted assistant-token position t, let p_{it} be the TRUE distribution and q_{it}^K the candidate distribution at the same reference context. Define the condition-level KL sum and functional recovery (FR):

$$S_K = \sum_{i=1}^{128} \sum_{t \in \mathcal{A}_i} \sum_{v \in \mathcal{V}} p_{it}(v) \big[\log p_{it}(v) - \log q_{it}^K(v)\big], \qquad \mathrm{FR}(K) = 1 - \frac{S_K}{S_{\mathrm{ANCHOR}}}. \tag{76}$$

The numerator and denominator have identical aggregation domains: full vocabulary, shifted assistant-token positions, and the same 128 examples. FR is not an average of per-example ratios. It requires a positive, numerically resolved anchor denominator. Small candidate KL additionally requires its own numerical sensitivity check; a large anchor KL or deterministic TRUE-repeat does not resolve close-distribution cancellation.

Table 28: Family totals against TRUE. S_{K1}, S_{K2} sum over eight conditions per family; N counts assistant tokens across repeated condition evaluations of the same 128 examples. KL totals are nats. The final column is the median paired condition difference. SmolLM2's small absolute K2 values are sensitive to forward precision.

Family	S_{K1}	S_{K2}	S_{K1}/N	S_{K2}/N	med. $(S_{K1} - S_{K2})$
Qwen3	0.483241	0.0149	2.60×10^{-6}	8.01×10^{-8}	0.05919
SmolLM2	0.00661062	0.000132	3.50×10^{-8}	7.01×10^{-10}	0.0007239
Llama-3.2	0.315958	0.00615	1.80×10^{-6}	3.50×10^{-8}	0.03451
OLMo-2	0.125292	0.000224	7.11×10^{-7}	1.27×10^{-9}	0.01562

Logits are produced by the canonical FP32 forward pass with eager attention and TF32 disabled. To resolve close-distribution KL values reliably, all post-logit log-softmax, exponentiation, $p(\log p -$

$\log q$), and reduction operations are evaluated in FP64. Every raw KL is reported without clipping. All candidates use the same teacher-forced contexts and 128-example evaluation pool.

The table below shows family medians of condition-wise FR(K1) and $1 - \mathrm{FR}(K2)$, the latter to avoid rounding tiny residuals to zero. Both methods already recover much of the anchor-to-TRUE distributional change. Canonical SmolLM2 residual values are descriptive and not resolved to fine precision across forward configurations.

Family	med. S_{anchor}	med. FR(K1)	med. $1 - \mathrm{FR}(K2)$	min. raw KL	negatives
Qwen3	9.4949	0.9937	1.78×10^{-4}	8.19×10^{-4}	0
SmolLM2	0.8868	0.9992	5.26×10^{-6}	8.64×10^{-7}	0
Llama-3.2	12.1995	0.9971	3.82×10^{-5}	1.33×10^{-4}	0
OLMo-2	5.7796	0.9974	4.82×10^{-6}	1.98×10^{-5}	0

TRUE-repeat KL and maximum repeated-logit difference are both zero, which establishes repeatability only. The SmolLM2 sensitivity analysis uses batches 2, 3, 4, and 7 and the first eight entries of the evaluation pool. Both arms use FP64 post-logit arithmetic; only the forward precision differs (FP32 versus FP64). Both configurations yield nonnegative KL and K2<K1 in all four conditions. Absolute K2 differences span 5.35%–103.80% of the smaller-arm KL (median 36.75%), whereas paired-gap differences span 0.0346%–0.1101% (median 0.0755%). We therefore emphasize paired ordering and give no near-perfect output recovery interpretation.

Across all 32 conditions, K2<K1. The K1–K2 KL sum is 0.910 nats; the paired median is 2.00×10^{-2} and the paired minimum 2.91×10^{-4} nats. The 726,240 tokens count repeated condition evaluations of the same pool and model-specific tokenizations. They are not iid scientific replicates. Teacher-forced evaluation measures fixed-context distributional agreement, not free-running task performance. The machine-readable supplement provides all 32 condition-level KL totals and token counts.

G.1 Weak-direction functional and continuation effects

A separate evaluation tests whether the weak second historical direction improves held-out NLL. It uses the eight independent Qwen3-1.7B and Llama-3.2-1B model–data blocks from the multi-scale study, with five step scales and construction horizons 8, 16, and 24. An eight-example held-out set, disjoint from the training examples, is fixed across candidates. For each condition, let $d = \theta^{\mathrm{TRUE}} - a$ and let V_k contain the first k historical update singular vectors. The candidate checkpoint OracleHistProj$k$ is $a + V_k V_k^\top d$, for $k = 1, 2$. The coefficients use the true future displacement, so this is an oracle representation diagnostic, not a rollout of the HistSVD1/2 matched recurrence in the same historical spaces. Evaluation uses FP32 forward passes and held-out assistant-token mean NLL. Reusing the multi-scale blocks does not provide eight additional independent parameter-space tests.

Within each block, results are summarized over its 15 scale–horizon conditions. OracleHistProj2 improves on OracleHistProj1 in six of eight block medians; the median NLL gain is 4.25×10^{-6} nats/token. The median of the valid block-level recovery ratios is 0.208% of the anchor-to-TRUE NLL improvement, below the predeclared 10% materiality threshold. Each ratio divides the block's median OracleHistProj1–OracleHistProj2 gain by its median anchor-to-TRUE improvement; this statistic is not a ratio of pooled medians. The gain is resolved above the evaluator's 10^{-8} nats/token floor, but small relative to the learning progress available in this protocol.

A continuation intervention uses the same eight blocks at step scale 0.0075 and construction horizon 24. Starting from OracleHistProj1, OracleHistProj2, the sign-reversed second-component control, and TRUE, it performs 24 real GD updates on the same fixed training loss. All starts retain the full original selected LoRA-B training coordinates. If $V_s(t)$ is held-out NLL for start s, define progress as $P_s = V_s(0) - V_s(24)$. The median $P_{\mathrm{OracleHistProj2}} - P_{\mathrm{OracleHistProj1}}$ is 2.30×10^{-5} nats/token; all eight block effects are positive, but the magnitude is below the predeclared 2.42×10^{-4} threshold. Median TRUE progress is 0.00342 nats/token. These tests establish a small functional effect in the evaluated regime; they do not establish a material learning benefit or test adaptation to a new task. Both evaluations used protocols fixed before their outcomes were read and retained all eight blocks.

H Reproducibility Details

H.1 Model and dataset identities

Tokenizers use the pinned model snapshots below. Packaged identity manifests record full revisions, configuration/tokenizer/weight hashes, and the frozen datasets' identities and parquet hashes. The compact supplement contains these manifests and source rosters; it does not contain model weights or the raw dataset snapshots.

Table 29: Pinned model repositories and revision prefixes. The selected-space results refer to these exact snapshots, not an unspecified current release; full revisions are indexed in the portable supplement.

Repository	License	Revision prefix
`Qwen/Qwen3-1.7B`	Apache-2.0	`70d244cc86cc`
`HuggingFaceTB/SmolLM2-1.7B`	Apache-2.0	`effd688a1292`
`meta-llama/Llama-3.2-1B`	Llama 3.2 Community License	`4e20de362430`
`allenai/OLMo-2-0425-1B`	Apache-2.0	`a1847dff3500`
`Qwen/Qwen3-8B`	Apache-2.0	`b968826d9c46`

The primary fixed-driver and exact-HVP multi-scale studies both use mathematical problem–solution examples from `siyanzhao/Openthoughts_math_30k_opsd`. Their frozen identity is the pair of parquet content hashes in `metadata/scope_extension/fresh_cohort_manifest.json`; the recorded revision is `local-frozen-equivalent`, not a pinned Hub commit. The separate supporting finite-difference study uses `sxiong/synthetic-math` at revision prefix `fbcdfb863622`. The following map distinguishes dataset reuse from reuse of experimental units. Paths here are relative to the supplement's `submission_supplement/` directory.

Study / purpose	Source	Cohort and reuse
Primary fixed-driver / mechanism	OpenThoughts math	32 examples in eight batches, shared across four model families.
Primary exact-HVP / orders	OpenThoughts math	Eight model–blocks; historical prompt hashes excluded. Five scales reuse each block.
Supporting finite-difference / coefficients	Synthetic math	Separate 12 blocks per model; five scales reuse each block (Appendix C.3).
Weak-direction NLL / continuation	OpenThoughts math	Reuses the eight primary exact-HVP blocks; a separate eight-example evaluation set (Appendix G.1).

OpenThoughts math and Synthetic math above abbreviate the two repositories just named. The primary multi-scale roster is `metadata/protocol/PHASE_A1_ROSTER_LOCK.json`; it records source parquet hashes, selected example hashes, exclusions and model–block membership. The primary fixed-driver cohort manifest contains source IDs, normalized hashes, batch membership, model-specific tokenization metadata, and exclusions. The saved primary `siyanzhao/Openthoughts_math_30k_opsd` snapshot and repository card do not declare a license. We record that absence rather than infer a license from its name; the manuscript exports aggregates and source identities, not the source text. The output-space evaluation reuses a fixed evaluation pool (internally denoted `Round13-FS` in the released artifacts), with zero source-ID and normalized-hash overlap with the primary fixed-driver panel. Its pool manifest is cryptographically identified in the supplementary manifest. The pool is pre-existing research material, not a newly held-out evaluation set.

Shared adapter and fixed-driver settings are specified once in Appendix B; the full-FP32 exact-HVP 8B exception is in Appendix C.5.

The released trajectory is `relex-rlvr/RLVR-Qwen2.5-Math-1.5B`, steps 1–32, with base `Qwen/Qwen2.5-Math-1.5B` at `4a83ca6e4526a4f2da3aa259ec36c259f66b2ab2`. All 33 revision identities and extracted tensor hashes are provided in the supplementary manifest.

The analysis uses the first eight adjacent checkpoint differences for its primary occupancy, and labels cumulative displacements separately. Under the selected effective-q/v metric, their rank-one occupancies are 0.299 for adjacent checkpoint differences and 0.817 for cumulative displacements. The public release therefore represents a different motion regime from the controlled primary fixed-driver panel. Released checkpoint labels do not establish consecutive optimizer-step spacing, and this analysis supplies no matched local-response comparison. It neither re-executes source training nor assumes the source optimizer matches SGD. The released trajectory card does not provide a separate license declaration verified here; it remains a referenced external artifact.

H.2 Code, artifacts, and reproduction

The supplement provides condition-level results, protocol metadata, analysis code, and model-free numerical checks. Its entry point, `submission_supplement/README.md`, maps scientific questions to evidence and verification procedures. Display-precision summaries are distinguished from full-precision sources. The supplied checks verify reported results; they do not constitute a re-execution of model training, and FP64 analysis does not remove rounding from saved FP32 trajectories.

For the Qwen3-8B confirmation, the supplement provides 40 display-precision condition summaries and saved consistency audits; raw trajectory vectors are not included.