When Historical Energy Is Insufficient for Local Training Response

Min Wu

Shanghai Key Lab of Trustworthy Computing
East China Normal University

Ongoing work jointly with Yuqiang Zhou *(ECNU, China)*

October 21, 2026

Table of Contents

01 LLM Reasoning and Training Dynamics

02 Historical Energy vs. Local Response

03 Gradient and Curvature Geometry

04 Experimental Results

05 Concluding Remarks

From LLM Reasoning to Local Training Dynamics

01 **LLM Mathematical Reasoning**

Improving complex reasoning capabilities requires effective post-training of large language models.

↓

02 **Reinforcement Learning with Verifiable Rewards (RLVR)**

RLVR improves reasoning through verifiable outcomes, such as mathematical correctness, motivating the study of training dynamics and efficiency.

↓

03 **The Challenge: Low-Dimensional Representation**

Can dominant historical update directions adequately represent future training behavior?

↓

04 **This Work: Historical Energy vs. Local Response**

We investigate whether high historical SVD energy retention guarantees accurate future local training response under fixed-loss gradient descent.

Problem Statement: Does preserving historical update energy preserve the directions needed for accurate future local training response?

The local forecasting task

Historical updates

Observed history

Use a finite sequence of recent updates.

One anchor

Shared starting point

Start every predictor at the same parameters.

Future response

Forecast window

Predict the local parameter response.

Only the allowed response space changes.

Anchor, local operator, step size, horizon, and error metric stay matched.

Related work: three research directions

Historical compression Sirovich (1987); Ilak & Rowley (2008)
POD/SVD retains high-energy modes. Weak-energy modes can still shape transient response.

Krylov response models Grimme (1997); Vinyals & Povey (2011)
Gradient and curvature actions build response spaces. These constructions have a classical foundation.

Training subspaces Li et al. (2023); Wei et al. (2026, RELEX)
Compact trajectories support training and extrapolation. Task gains differ from matching a local parameter trajectory.

Research gap and contributions

Does capturing almost all past motion preserve local prediction accuracy?

Capture conditions
Specify how accurately a space must retain the gradient and its first curvature action.

Best-line limit
Even the best fixed line loses one accuracy order relative to a suitable plane.

Exact-history recovery
Exact history contains a sufficient plane, but a fixed energy cutoff can discard its weak direction.

Historical energy and forecast error

Problem Setup and Methods: What exactly do we predict, compare, and measure?

Historical energy

$$M = 1 - \sigma_1^2 / \| U \|_F^2$$

U: observed updates; $\| U \|_F$: Frobenius norm. M measures squared motion outside the leading historical direction.

Forecast error

$E =$ mean relative forecast error. At each horizon, divide the error distance by the true displacement, then average over horizons.

Three trajectories, two comparisons

TRUE

Actual nonlinear gradient-descent continuation.

LocalFull

Linearized gradient dynamics in the full parameter space.

Restricted

The same local dynamics confined to a chosen subspace.

TRUE vs LocalFull: local-model mismatch.

LocalFull vs Restricted: cost of subspace restriction.

The candidate response spaces

Method	Dimension	Direction source
K1	1	Current gradient
K2	2	Gradient and Hessian–gradient action
HistSVD1	1	Leading direction of past updates
HistSVD2	2	Leading two directions of past updates
LocalFull	Full	No subspace restriction

The same local dynamics, with or without projection

LocalFull

$$z_{h+1} = z_h - \eta(g_a + H_a z_h), \qquad z_0 = 0.$$

Restricted

$$\hat{z}_{h+1} = \hat{z}_h - \eta QQ^\top(g_a + H_a \hat{z}_h), \qquad \hat{z}_0 = 0.$$

Why preserve second-order relative accuracy?

Theoretical Results: What must a response space retain for the target local accuracy?

A common accuracy target

For fixed $P, H \geq 2$, fixed dimension, smooth fixed loss and $g_0 \neq 0$,

$$\|\Delta_h\| = \Theta(\eta), \qquad \|\Delta_h - z_h\| = O(\eta^3).$$

Thus LocalFull guarantees $O(\eta^2)$ normalized error.

One denominator for every comparison

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

Compression should preserve this guarantee. Special losses can make LocalFull higher order or exact.

The theorem concerns a short, fixed-loss window

Dynamics	Exact gradient descent on one fixed smooth loss.
Regularity	A locally Lipschitz Hessian. Convexity is unnecessary.
Local geometry	A nonzero gradient and transverse curvature component.
Limit	Fixed dimension and finite window lengths as $\eta \to 0$.

Forward motion and the first local turn

$$\Delta_h = h\eta G q - \frac{1}{2}h(h-1)\eta^2 G H_a q + O(\eta^3), \qquad q = -g_a/G, \quad G = \|g_a\|.$$

Forward displacement
Order η. The gradient line captures this term.

Turning displacement
Order η^2. Relative error divides by a displacement of order η.

The sharp optimum over every anchored fixed line

Theorem 2 exact leading coefficient

$$\inf_{\mathcal{A}_1(\eta)} E = c_H^* \beta_0 \eta + O(\eta^2), \qquad c_H^* = \frac{\lfloor H^2/4 \rfloor}{2H}.$$

Here $q_0 = -g_0/\|g_0\|$ and $\beta_0 = \|(I - q_0 q_0^\top) H_0 q_0\| > 0$.

The line passes through the anchor and stays fixed across the window. Its orientation may use the complete future; coefficients are unrestricted at each horizon.

For $H = 24$

Best line: $3\beta_0\eta + O(\eta^2)$; K1: $5.75\beta_0\eta + O(\eta^2)$. Orientation improves the constant while the error remains first order.

Proof idea: a geometric problem becomes a median problem

Normalized future directions

With $r_a = (I - q_a q_a^\top) H_a q_a$,

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

Every optimal line localizes near q_a.

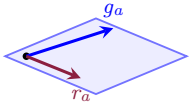
Leading-order optimization

The transverse tilt reduces the problem to

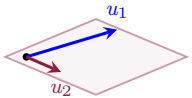
$$\frac{\eta\beta_0}{2H} \min_{t\in\mathbb{R}} \sum_{h=1}^{H} |(h-1) - t|.$$

A median of $\{0, \dots, H-1\}$ minimizes the sum, whose minimum is $\lfloor H^2/4 \rfloor$.

A suitable plane captures the first turn



K2
Gradient + curvature



HistSVD2
Top two directions of exact history

$$E = O(\eta^2)$$

K2

$$K_2 = \mathrm{span}\{g_a, H_a g_a\}$$

Gradient and Hessian at the anchor

HistSVD2

Top two singular directions of exact history

Different constructions; both preserve the second-order local response.

Two capture conditions characterize the target order

Theorem 1 necessary and sufficient capture

Under the fixed-loss assumptions, an orthogonal projector Π_η, fixed across the forecast window, preserves $E = O(\eta^2)$ if and only if

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

The space may otherwise depend arbitrarily on η.

Why both directions matter

The first two horizons give necessity; the local expansion gives sufficiency. The same conditions characterize best representation with unrestricted coefficients at each horizon.

The needed plane is already present in exact history

Construction	Information	Local relative error
K2	Gradient and Hessian action	$O(\eta^2)$
HistSVD2	Top two directions of exact past updates	$O(\eta^2)$
HistSVD1	Leading historical direction alone	Generally first order

A fixed energy cutoff can discard the local turn

1D

selected by a fixed energy rule

Historical tail: $M = \Theta(\eta^2)$. Any fixed retained-energy threshold below 100% eventually selects rank one.

2D

needed for the second-order guarantee

Best fixed line: $E = \Theta(\eta)$. Suitable plane: $E = O(\eta^2)$.

Historical energy quantitatively predicts omitted response

Theorem 4 historical energy and omitted response

$$M = C_P(\eta\beta_0)^2 + O(\eta^3), \qquad C_P = (P^2-1)/12,$$
$$R := D_0(\text{LocalFull}, K1) = C_H\eta\beta_a + O(\eta^2), \qquad C_H = (H-1)/4,$$

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

Energy calibrates the omitted response

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

The retention criterion must match the accuracy target.

A scale-dependent threshold can retain the sufficient plane

An asymptotic selection rule

In exact noiseless history, choose

$$\tau(\eta) = 1 - \kappa M(\eta), \qquad 0 < \kappa < 1.$$

The relative energy tail after rank two is $T_2 = O(\eta^4)$.

Why it selects exactly two directions

Eventually $T_2 \le \kappa M < M$: rank one fails the threshold, while rank two meets it. Merely choosing $1 - \tau = O(\eta^2)$ is insufficient without a small enough coefficient.

A Controlled Experiment to Isolate Local Response Geometry

Core experimental setup

Dimension	Setting	Purpose
Models	Qwen3-1.7B, SmolLM2-1.7B, Llama-3.2-1B, OLMo-2-0425-1B	Check consistency across models
Trainable coordinates	Final-block query/value LoRA-B; backbone and LoRA-A frozen	Control computational scale
Optimizer	Stateless gradient descent	Exclude optimizer history
Precision	FP32	Resolve weak directions
Stochasticity	Dropout disabled	Avoid dropout randomness
Step scale	$\eta\|g_0\| = 0.03$	Match initial update norms

Core experimental pipeline

8 Historical Steps
Updates $u_1, \ldots, u_8$
Construct historical SVD space

$\longrightarrow$

Anchor a
Same starting point
Construct gradient- and Hessian-based spaces

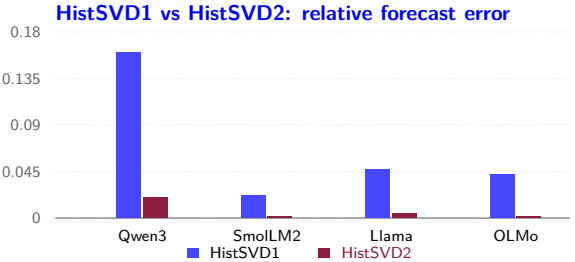
$\longrightarrow$

24 Future Steps
Forecast the response
Compare TRUE, LocalFull, K1, K2, HistSVD1, HistSVD2

A second historical direction improves every condition

Family median relative forecast error E (lower is better)

Model	HistSVD1	**HistSVD2**	Improvement ratio
Qwen3	0.160	0.0203	7.9×
SmolLM2	0.0214	0.00119	18.1×
Llama	0.0469	0.00401	11.7×
OLMo	0.0422	0.00124	34.0×



32/32

HistSVD2 improves on HistSVD1 in every condition.

Same local operator and recurrence; only the retained historical space changes.

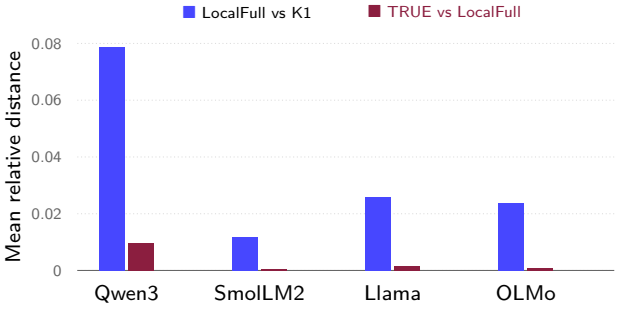
Restriction exceeds local-model mismatch

32/32

Restriction distance exceeds local-model mismatch.

TRUE vs LocalFull
Local-model approximation mismatch.

LocalFull vs K1
One-dimensional restriction error.

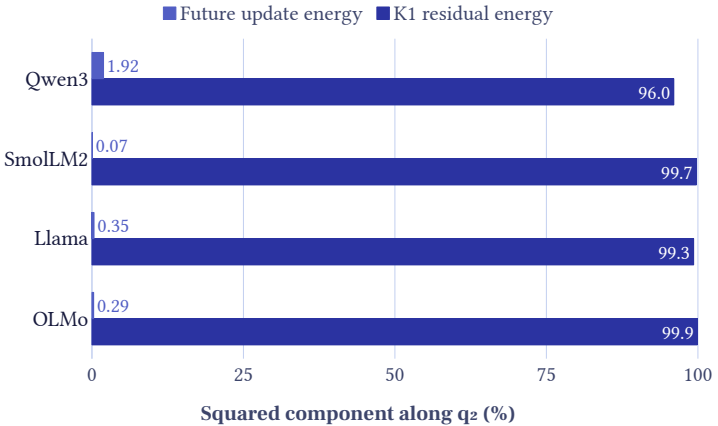


A weak direction explains most of the missed response

$< 2\%$
of **total future update energy**

$> 96\%$
of **K1 squared residual energy**

Different denominators:
low contribution to total motion does not imply low contribution to prediction residual.



A separate study tests the predicted scaling

Why vary the scale?

Single-scale comparisons test improvement; multi-scale comparisons test accuracy orders.

8 independent blocks $\times$ 5 nested scales

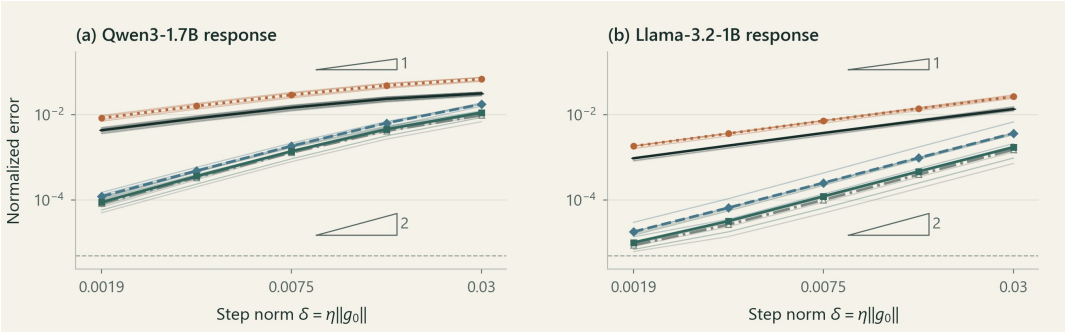
Qwen: 4 model–data blocks. Llama: 4.

Initial step norms $\delta = \eta \|g_0\|$:
0.03, 0.015, 0.0075, 0.00375, 0.001875.

Same initialization within each block; fixed loss, $P = 8$, $H = 24$.
Directions from autograd Hessian–vector products in FP32.

Measured scaling separates lines from planes

If $E \propto \eta^p$, the log–log slope is p.



Step norm = $\eta\|g_0\|$. Thin curves: blocks. Thick curves: medians.
Triangles: order guides. Black curves: best-line bounds, not confidence intervals.

Line-based response: approximately first order.
K2, HistSVD2, LocalFull: closer to second order.

Finite-scale agreement and deviations

Slopes over five tested scales

K1 median slopes: **Qwen** 0.77**, Llama** 0.98.
K2, HistSVD2 and LocalFull: approximately 1.73–1.92.
The five tested scales support, but do not prove, the limiting asymptotic orders.

A separate check of the sharp coefficient

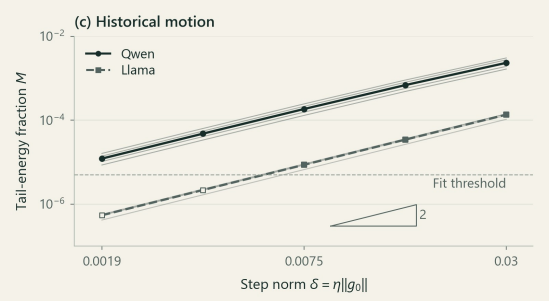
For $H = 24$, normalize the geometric best-line bounds by $3\eta\beta_0$.
At the smallest scale, lower and upper bounds nearly coincide:
Qwen 0.951**; Llama** 0.994. Qwen retains about 4.9% finite-scale deviation.

Historical energy shrinks faster than line error

99%

retained-energy threshold selects rank one in every tested block and scale.

Median tail slopes:
Qwen: 1.898
Llama: 1.985



Historical energy becomes increasingly one-dimensional while second-order forecast accuracy still requires the curvature direction.

Changing drivers: a local law for historical motion

Interpolating fixed and changing losses

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

Fixed finite P, $g_0 \neq 0$; uniformly bounded, locally Lipschitz Hessians near θ_0.

Proposition 1: joint local expansion

As $(\lambda, \eta) \to (0,0)$ and $\rho = \lambda + \eta$,

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

$D_\perp, K_\perp$: normalized driver variation and curvature correction, centered across steps and projected transverse to g_0.

Driver variation and curvature can reinforce or cancel

Coefficients and geometry

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

$A, C \geq 0$, $B^2 \leq AC$, but B can have either sign. At $\lambda = 0$, the fixed-loss curvature law is recovered.

Evidence and scope

480 grid conditions within 16 model–data blocks. In the primary region, predicted-versus-measured log–log slopes are 1.009 (Qwen), 1.023 (Llama). This joint-local law proves neither a fixed-λ positive floor nor monotonicity, and does not extend the response theorem to AdamW.

The AdamW control supports the selection concern

Comparison	Observed outcome
Same dimension	Control wins 32/32 windows
Include current update	Control wins 29/32 windows
Replication	2 models; 8 trajectories; 4 windows each
Extra control information	Scheduled gradients; HVPs; local predictions

Concluding Remarks

Historical energy

Describes past motion and, in the fixed-loss protocol, calibrates omitted response through $R \propto \sqrt{M}$.

Local prediction

Also depends on capturing the next turn.

Practical utility

Requires evidence beyond better parameter forecasts.

Thank you for your attention!