THE GEOMETRY OF WEIGHT-TYING SENSITIVITY IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transpose tying changes both a learned operator and the residual trajectory on which later computation acts. We develop a finite geometric account of these two effects in pre-normalized Transformers. Positive channel scaling yields a unique balanced SwiGLU representation and a weighted-adjoint tying operation that is invariant under function-preserving channel rescaling. Exact gated derivatives identify a rank-two contribution to normalization-induced rotation, while radial–directional coordinates distinguish residual magnitude from the angular change produced by an update. For a fixed pre-RMSNorm suffix at input scale c, we derive explicit coefficients for its c^{-1} logit and c^{-2} output-KL response, together with finite-radius bounds and direction-resolved endpoint certificates. Public-model interventions show large propagation differences after local edit-magnitude matching. A complete two-model, six-text strength panel tests the balanced edit itself: its coordinate consistency coexists with substantial functional changes, including a low-amplification regime. A separate two-model reference-precision experiment computes the suffix coefficients without fitting output responses: at $c = 32768$, the six cut-level forward-KL ratios to prediction range from 0.99134 to 1.00022. The theory links a well-defined editing operation to residual geometry and quantitatively testable suffix effects.

1 INTRODUCTION

A parameter constraint and an intervention on a trained model pose different questions. A Transformer can adapt while its feed-forward projections are tied, whereas replacing an independently learned projection imposes a new operation on an existing configuration (Liang, 2026). Our public-model controls show that calibrated transpose tying propagates far more strongly than Gaussian edits matched in either weight-change norm or original-input branch perturbation. Yet residual amplification does not order their language-model loss damage.

The geometric structure of the computation provides two complementary points of entry. At the operator level, the same branch function admits different channel-scale allocations, so ordinary transpose replacement need not denote the same edit in equivalent coordinates. At the state level, normalization separates residual radius from direction: radius modulates the relative step of later computation, while the final normalized readout is approximately insensitive to positive rescaling. Both effects can be analyzed directly on the finite attention–FFN computation, without a continuous-depth approximation.

We establish three results. First, the unique row/column-balanced representative of a bias-free SwiGLU branch defines a scale-compatible weighted-adjoint reference and an invariant affine editing path. Second, exact gate and normalization derivatives identify a rank-two rotation component and lead to precise radial–directional residual updates. Third, a fixed-suffix expansion computes logit, NLL and KL coefficients from branch responses at one frozen input direction. Complete finite-radius bounds and endpoint certificates accompany the asymptotic coefficients.

Two fixed-checkpoint experiments test the suffix results. Same-scale native float32 continuation, with a common float64 head, measures distributional attenuation. A separate float64 reference of the stored checkpoints compares independently calculated coefficients with measured responses, retaining all cuts and passages, including early-Qwen transition regimes. A third, native-precision panel tests the balanced operator directly over the complete strength grid, with separate matched

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.

controls for every reference. Earlier structural edits, oracle state controls, diagnostic comparisons and historical training records remain available with their original numerical outcomes.

2 FINITE ARCHITECTURE AND RESPONSE SETTING

For column inputs the bias-free SwiGLU branch is

$$f(x) = Dz_U(x), \qquad z_U(x) = \phi(Gx) \odot Ux, \quad \phi(s) = s\ \mathrm{sigmoid}(s), \tag{1}$$

with $U, G \in \mathbb{R}^{m\times d}$ and $D \in \mathbb{R}^{d\times m}$. A pre-normalized attention–FFN block maps the entire sequence by

$$X \mapsto X + \mathcal{T}(X) \mapsto X + \mathcal{T}(X) + \mathcal{R}(X + \mathcal{T}(X)). \tag{2}$$

Masks, positions, weights and sequence length are fixed. The maps are C^3 near the trajectories in the theorems; at zero RMS epsilon all relevant token states are nonzero. Token weights $\mu_i > 0$ on the declared target set sum to one. The output law at a target is the softmax of finite logits. All response statements are on this fixed evaluation set.

For empirical residual block entries H_k and complete last output H_L, we retain

$$A_k = \frac{\mathrm{std}(H_k)}{\max\{\mathrm{std}(H_0), 10^{-9}\}}, \qquad \mathcal{E}_k = \frac{A_{k,\mathrm{edited}}}{\max\{A_{k,\mathrm{original}}, 10^{-9}\}}. \tag{3}$$

The standard deviation pools tokens and features within a passage. The per-token radii used below are distinct observables. Next-token NLL is target-weighted within a model; absolute losses across tokenizers are not pooled.

3 SCALE-COMPATIBLE EDIT SOURCES

Positive diagonal channel scaling preserves the gate-fixed branch:

$$U_C = CU, \quad D_C = DC^{-1}, \quad z_{U_C} = Cz_U, \quad D_C z_{U_C} = Dz_U. \tag{4}$$

Ordinary transpose replacement instead gives $U^\top C^2 z_U$ in the new representation. A feature metric transported by

$$M_C = C^{-\top} M C^{-1}, \qquad U_C^\top M_C z_{U_C} = U^\top M z_U \tag{5}$$

keeps the edited function unchanged.

Proposition 1 (Tying in the balanced representative)**.** *For every nonzero row $u_j^\top$ of U and nonzero column d_j of D, define*

$$(C_*)_{jj} = \sqrt{\|d_j\|/\|u_j\|}, \quad M = C_*^2, \quad T_M = U^\top M, \qquad \|(T_M)_{:j}\| = \|d_j\|. \tag{6}$$

The pair (C_*U, DC_*^{-1}) is the unique point on this positive-scaling orbit with equal corresponding row/column norms. Ordinary transpose tying there gives precisely $T_M z_U$ in the original branch coordinates. Both the reference and the entire affine edit $D(t) = (1-t)D + tT_M$ are functionally invariant under Eq. (4)*.*

Balancing solves $c_j\|u_j\| = \|d_j\|/c_j$ uniquely. The metric then transforms by Eq. (5); it also preserves $\|D\|_F$ and fixes an already tied branch. The active-channel and zero-channel cases are treated in Appendix E. Section 7 tests this reference and its affine path directly on the two fixed models, separately from the earlier scalar-reference observations.

For the scalar subgroup, $\alpha = \|D\|_F/\|U\|_F$ and the archived leading-mode statistic g_ℓ satisfy

$$f_c = f, \quad f_{\mathrm{tied,c}} = c^2 f_{\mathrm{tied}}, \quad \alpha_c = \alpha/c^2, \quad g_{\ell,c} = c^2 g_\ell. \tag{7}$$

Thus αg_ℓ, not raw g_ℓ, respects this scalar symmetry. This separates the choice of an editing operation from the scale of its coordinates.

4 NORMALIZATION AND RESIDUAL-STATE GEOMETRY

The exact gated derivative is

$$K(x) = D[\mathrm{diag}(\phi(Gx))U + \mathrm{diag}(Ux)\,\mathrm{diag}(\phi'(Gx))G]. \tag{8}$$

Both terms and the normalization derivative enter the residual Jacobian. For a block of Eq. (2), $J_{\mathrm{block}} = (I + D\mathcal{R}|_{X+\mathcal{T}(X)})(I + D\mathcal{T}|_X)$. Causality makes these sequence Jacobians block lower triangular. Appendix C retains the normalization-induced rotation plane and exact radial–directional update; no symmetry of a weight pair is substituted for this full derivative.

How normalization changes the geometry. A normalized gradient branch $\nabla U \circ \rho$ differs from the Euclidean gradient pullback $D\rho^\top \nabla U \circ \rho$. The architecture evaluates the former composition, not the latter adjoint-corrected field. Consequently even a symmetric effective derivative can acquire vorticity on composition with radial normalization. Write $h = ru$, $A = K(N(h))\Gamma = S + B$, $S^\top = S$, $B^\top = -B$, and $J_\rho = aI - buu^\top$. The inherited identity and its symmetric-normalization component give

$$\begin{aligned} \Omega &= -[S, J_\rho] - \{B, J_\rho\}, \\ \Omega_S &= b(v_S u^\top - u v_S^\top), \quad v_S = (I - uu^\top)Su, \quad b = \frac{\sqrt{d}\, r^2}{(r^2 + d\epsilon)^{3/2}}. \end{aligned} \tag{9}$$

Thus Ω_S has rank at most two, operator norm $b\|v_S\|$, and acts only on $\mathrm{span}\{u, v_S\}$. The full skew contribution remains in B. This identifies a local rotation mechanism without treating skewness as automatic norm stabilization.

For $h \neq 0$, $h + q \neq 0$ and $q = \gamma u + v$ with $v \perp u$, the state variables obey

$$r^+ = \sqrt{(r+\gamma)^2 + \|v\|^2}, \qquad u^+ = ((r+\gamma)u + v)/r^+. \tag{10}$$

To first order in $\|q\|/r$, direction changes by v/r. At the final readout, $\mathcal{R}_\epsilon(ru) = \sqrt{d}\, W\Gamma_f u/\sqrt{1 + d\epsilon/r^2}$. Radius therefore controls angular step size inside the network, whereas positive-ray direction controls its zero-epsilon final readout. These two roles explain why scale invariance of the final head does not make a preceding suffix scale invariant.

Propagation of a specified small edit. For $X_{\ell+1}(t) = \Phi_\ell(X_\ell(t), t)$ with fixed input, differentiation at $t = 0$ gives $\dot{X}_{\ell+1} = J_\ell \dot{X}_\ell + e_\ell$, where $J_\ell = D_X \Phi_\ell$ and $e_\ell = \partial_t \Phi_\ell$ on the base trajectory. Thus induction yields

$$\dot{X}_L = \sum_\ell J_{L-1} \cdots J_{\ell+1} e_\ell. \tag{11}$$

For a down-projection edit $e_\ell = \dot{D}_\ell z_{U_\ell}$ at its original inputs. The complete ordered derivatives, rather than a local edit norm alone, therefore determine the first-order state change. This is a finite chain-rule identity, not an approximation for an arbitrarily large edit.

5 EXPLICIT COEFFICIENTS OF FIXED-SUFFIX INFLUENCE

Fix a finite pre-RMSNorm suffix, all its weights and epsilons, and a sequence $H_i = r_i u_i \neq 0$. Define $q_j^0 = F_j \circ N_{j,0}$, $a_j = q_j^0(H)$ and $L_j = Dq_j^0(H)$. Every branch is evaluated at the *same* H, with its own weights, gains and causal context. This is different from an ordinary clean suffix trajectory. Set

$$E_0 = \sum_j a_j, \qquad E_1 = \sum_{j<\ell} L_\ell a_j, \quad P_i = I - u_i u_i^\top, \quad A_f = W\Gamma_f. \tag{12}$$

Theorem 1 (Frozen-ray coefficient of suffix influence)**.** *For fixed finite smooth branches and $c \to \infty$,*

$$\mathcal{S}(cH) = cH + E_0 + c^{-1}E_1 + O(c^{-2}), \qquad v_i = \frac{\sqrt{d}}{r_i} A_f P_i E_{0,i}. \tag{13}$$

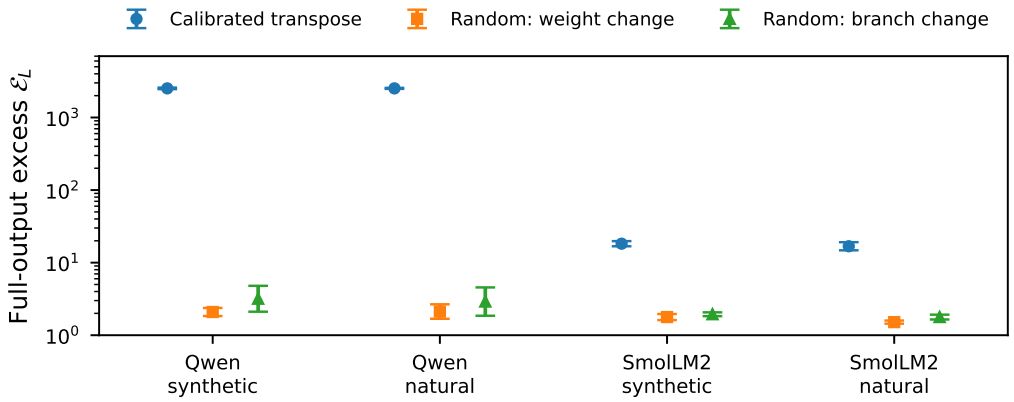


Figure 1: Local magnitude matching leaves a large propagation gap. Bars span the observed all-layer excess across texts and perturbation seeds; symbols sit at the geometric midpoint of each range for display, not at an estimated mean. Both text panels and both models use the same two local matching rules. Ranges are descriptive, not confidence intervals; see Table 1 for loss damage and exact ranges.

Use the same final RMSNorm/head for direct and suffix outputs, and put $p_{\infty,i} = \mathrm{softmax}(\sqrt{d}A_f u_i)$. *Then*

$$\Delta z_i(c) = c^{-1} v_i + O(c^{-2}), \tag{14}$$

$$\lim_{c\to\infty} c^2 \mathrm{KL}(p_i^{\mathrm{dir}}(c) \| p_i^{\mathrm{suf}}(c)) = \tfrac{1}{2} \mathrm{Var}_{p_{\infty,i}}(v_i), \tag{15}$$

$$\lim_{c\to\infty} c\Delta \mathrm{NLL}_{y,i}(c) = (p_{\infty,i} - e_y)^\top v_i. \tag{16}$$

Reverse KL has the same limit. A nonconstant v_i *gives* $\Theta(c^{-2})$ *KL. If* v_i *is a common shift, both divergences are* $O(c^{-4})$.

Set $t = 1/c$ and $Y_j = tX_j$. Exactly, $Y_{j+1} = Y_j + tF_j(N_{j,\epsilon_j t^2}(Y_j))$. Finite recursion gives $Y_s = H + tE_0 + t^2 E_1 + O(t^3)$. Expanding the common readout and the output KL proves the theorem. Appendix B gives the full proof and explicit second logit coefficient. The finite-radius norm bound and direction-resolved endpoint certificate remain in Appendices D and E.5.

6 Local magnitude, propagation and task function

Qwen3-0.6B and SmolLM2-360M have 28 and 32 blocks respectively (Qwen Team, 2025; Ben Allal et al., 2025). All experiments fix checkpoint/tokenizer revisions. The earlier intervention studies use three synthetic passages (382/383 targets) and three separate 1024-token WikiText excerpts per model. They evaluate native float32 forwards without retraining.

On original calibration inputs z_ℓ, the reference is $T_\ell = \beta_\ell U_\ell^\top$ with

$$\beta_\ell = \frac{\mathrm{RMS}(D_\ell z_\ell)}{\max\{\mathrm{RMS}(U_\ell^\top z_\ell), 10^{-9}\}}. \tag{17}$$

Gaussian edits $D_\ell + \lambda_\ell R_\ell$ separately match $\|T_\ell - D_\ell\|_F$ or $\mathrm{RMS}((T_\ell - D_\ell)z_\ell)$. Three fixed perturbation seeds are used; calibration does not read evaluation losses. The seven edit matrices remain unchanged on the natural-text panel.

Table 1 and Figure 1 show that local magnitude matching leaves large propagation differences on both text panels. All-layer transpose excess on natural text is 2479–2560 for Qwen and 14.79–19.13 for SmolLM2, versus 1.69–4.56 and 1.44–1.92 for ordinary edits. Both families substantially increase loss. All original calibration and per-passage conditions remain in Appendix X.

Propagation and function separate. A quiet recorded entry can miss the edited layer: editing Qwen layer 27 leaves $\mathcal{E}_{27} = 1$, although loss increases by 19.54–22.01 nats and complete-output

Table 1: All-layer reference and ordinary edits. Short uses three synthetic passages; Natural uses three 1024-token WikiText excerpts per model. Original calibration coefficients are frozen across texts. Loss ranges span three seed-level target-weighted means for random edits; output-excess ranges span seeds and texts. Transpose has one weight configuration. Ranges are descriptive, not confidence intervals; loss is compared within each model/text panel.

Model	Text	Edit	$\Delta\mathcal{L}$ (nats)	$\mathcal{E}_L$ range
Qwen	Short	Transpose	31.56	2459.01–2579.78
Qwen	Short	Random: weight match	9.49–10.66	1.84–2.38
Qwen	Short	Random: branch match	10.19–11.86	2.11–4.79
Qwen	Natural	Transpose	33.46	2479.31–2559.51
Qwen	Natural	Random: weight match	10.71–11.94	1.69–2.67
Qwen	Natural	Random: branch match	11.15–13.08	1.85–4.56
SmolLM2	Short	Transpose	40.43	16.81–19.80
SmolLM2	Short	Random: weight match	11.05–11.53	1.62–1.96
SmolLM2	Short	Random: branch match	11.08–12.36	1.83–2.07
SmolLM2	Natural	Transpose	42.48	14.79–19.13
SmolLM2	Natural	Random: weight match	12.09–12.26	1.44–1.59
SmolLM2	Natural	Random: branch match	12.21–12.67	1.65–1.92

excess is 2.44–2.60. On the repair passage, editing layer 26 instead gives excess 19.94 and loss damage 0.81. These examples motivate measuring the complete output law rather than using amplification as its proxy (Appendix T).

State interventions. At a cut, compare the edited state h_e with $(\|h_o\|/\|h_e\|)h_e$ and $(\|h_e\|/\|h_o\|)h_o$, then run the original suffix. They restore clean length or clean direction. Direction restoration yields lower NLL in all 45 original transpose comparisons per model and all 252 ordinary-edit internal comparisons across both text panels. These oracle patches use the complete clean direction. Their shared contrast motivates the radial–directional variables, not a claim that a single singular vector mediates recovery. Full contrasts, unit/Frobenius controls and all diagnostic baselines remain in Appendices U–X.

7 Scale-compatible tying from weak to complete replacement

We now apply the balanced construction itself, rather than treating the earlier scalar-calibrated transpose as its measurement. The fixed panel compares T_M, $T_F = \alpha U^\top$ and $T_B = \beta U^\top$ along $D(t) = D + t(T - D)$ at eight strengths from 0 to 1. Each reference has three seeds under each of the two local matching rules. Both models and all six texts are retained: 2016 primary cells and 288 coordinate checks, with repeated zero-strength cells sharing measured baselines. Appendix H gives the exact inventory and arithmetic.

The edit is the same in equivalent channel coordinates. The fixed nonuniform rescaling test recomputes each reference from the transformed parameters and compares its full path to the untransformed path. The balanced reference has maximum relative state discrepancy 0 across all tested strengths, texts and depths. The scalar references generally change under this operation. This directly tests the coordinate-consistency property of Proposition 1; its empirical consequence for language-model loss is a separate question.

Function can change before large amplification. At $t = 0.05$, Qwen's natural-text NLL increases by 0.1650 nats while complete-output excess spans only 0.9806–1.0029; SmolLM2 gives 0.2536 nats and 1.0000–1.0146. The corresponding synthetic-panel increases are 0.2744 and 0.4726 nats, also with excess near one. Table 2 compares each result with its own locally matched random controls. These responses occur before large amplification in the pooled standard-deviation measure, rather than only at complete replacement. The entire fixed grid, including nonlinear transitions, signed changes and all control outcomes, is retained.

Table 2: Balanced-reference edits at $t = 0.05$. NLL and KL average prediction targets within each model/text panel. Excess ranges span the three texts; random-control NLL ranges span three seed-level target-weighted means. The controls are calibrated to the balanced reference. This is one displayed point from the complete eight-strength grid, not a selected functional-safety threshold.

Model	Text	ΔL	KL	$\mathcal{E}_L$ range	Random ΔL Weight match	Random ΔL Branch match
Qwen	Natural	0.1650	0.1211	0.9806–1.0029	0.0204–0.0413	0.0680–0.0962
Qwen	Short	0.2744	0.1795	0.9815–1.0033	-0.0065–0.0369	0.0227–0.0882
SmolLM2	Natural	0.2536	0.2494	1.0000–1.0146	0.0365–0.0504	0.0776–0.0949
SmolLM2	Short	0.4726	0.2777	1.0047–1.0164	0.0763–0.1167	0.1391–0.1717

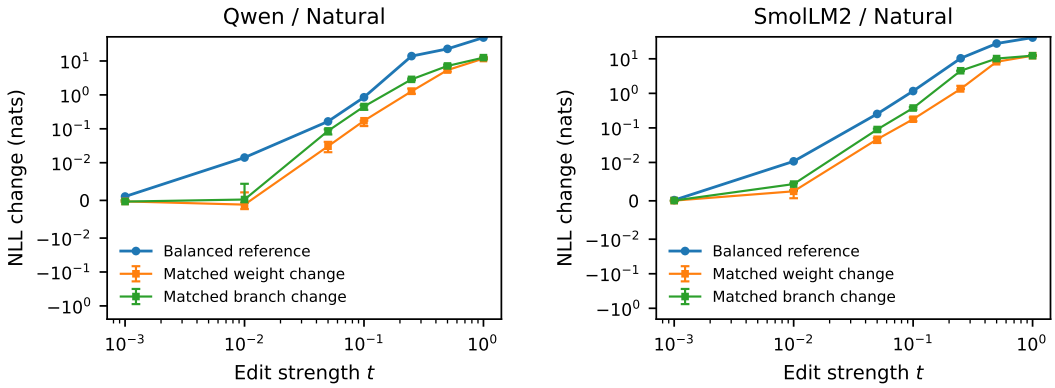


Figure 2: The complete positive-strength grid on the natural texts. Balanced-reference loss is target-weighted over all three passages. Each matched-control line is the median of three seed-level target-weighted means, with their full range, not an uncertainty interval. Zero strength is exactly the common baseline. All scalar-reference curves are tabulated in Appendix H; matching is reference-specific.

8 Testing suffix coefficients rather than fitting a decay law

Same-scale native-reference panel. Both paths receive exactly cH_o: direct final readout, or the original suffix followed by that readout. At cuts $\{7, 14, 27, 28\}$ for Qwen and $\{8, 16, 31, 32\}$ for SmolLM2, all five scales $\{1, 8, 64, 512, 4096\}$ are retained. Six clean full forwards, 90 nonempty float32 suffix evaluations and 30 empty controls form 120 conditions. Both final readouts are evaluated in float64 on those endpoints. All 18 nonempty passage/cut logit curves and 36 directed-KL curves decrease. Early Qwen cuts have descriptive tail exponents 0.626–0.705 for logits and 1.302–1.491 for KL, slower than the terminal/SmolLM2 curves. Appendix G retains the full panel and endpoint certificates.

A distinct coefficient experiment. Theorem 1 predicts v_i and $C_i = \mathrm{Var}_{p_{\infty},i}(v_i)/2$ by frozen branch evaluations, without using measured output differences. The new fixed panel reuses the same cuts and three synthetic token sequences at $c \in \{64, 512, 4096, 32768\}$. It contains 96 conditions: 72 nonempty reference suffix evaluations and 24 empty controls, plus six native clean full forwards, 18 $c = 1$ reference checks and 462 frozen branch evaluations. Every condition is included, whether or not its approximation is accurate.

This panel uses a *float64 reference of stored checkpoint computation*. Three native clean forwards per model capture states, masks and RoPE arrays. Stored weights are reloaded without changing their values; attention and RMS retain float64 instead of forcing intermediate float32. With $t = 1/c$, the exact rescaled recurrence is

$$Y_{j+1} = Y_j + tF_j(N_{j,\epsilon_j t^2}(Y_j)), \quad Y_0 = H_o. \tag{18}$$

Table 3: Frozen-ray coefficients at $c = 32768$, computed without fitting an output response. E_z is pooled relative logit-coefficient error; $R_{\rightarrow}$ and $R_{\leftarrow}$ divide the measured, scaled KL sums by the predicted coefficient sum. Rows weight all targets in the three passages within each model. Earlier scales and all individual passages remain in Appendix F.

Model	Cut	E_z	$R_{\rightarrow}$	$R_{\leftarrow}$
Qwen	7	0.012886	0.991338	0.989820
Qwen	14	0.004588	0.993299	0.992020
Qwen	27	0.000013	1.000001	0.999996
SmolLM2	8	0.000568	1.000137	1.000569
SmolLM2	16	0.000308	1.000221	1.000569
SmolLM2	31	0.000014	0.999974	0.999946

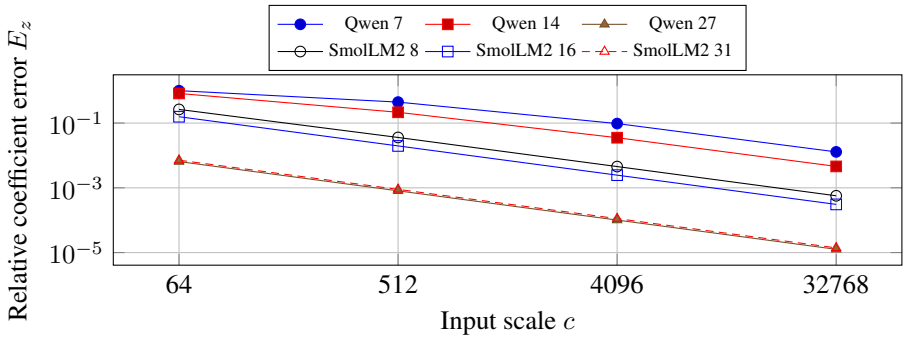


Figure 3: Convergence to the predicted logit coefficient, with all four fixed scales and all six nonempty cuts. Each curve aggregates every target within a model/cut; it is not a fitted decay law or a confidence interval. Early Qwen cuts remain far from the approximation at the smaller scales.

This rewrites the fixed-epsilon map at cH_o; it is not an epsilon intervention on that map. Native RoPE arrays are fixed, and Qwen's internal query/key normalization epsilons are not rescaled. The earlier float32 panel is not replaced by this reference calculation.

Coefficient convergence. For each cut define the pooled relative logit error $E_z(c) = (\sum_i \|c\Delta z_i - v_i\|_2^2 / \sum_i \|v_i\|_2^2)^{1/2}$, and $R_{\rightarrow}(c) = c^2 \sum_i \mathrm{KL}(p_i^{\mathrm{dir}} \| p_i^{\mathrm{suf}}) / \sum_i C_i$, with the reverse ratio analogously. At $c = 32768$, six model/cut forward ratios range from 0.99134 to 1.00022 and reverse ratios from 0.98982 to 1.00057. Qwen's early-cut logit errors are 1.289% and 0.459%; SmolLM2's are 0.0568% and 0.0308%. Last-block errors are about 0.0013% and 0.0014%. Table 3 and Figure 3 include earlier scales where the approximation is much less accurate. For Qwen cut 7 at $c = 64$, $E_z = 1.002$ and $R_{\rightarrow} = 0.0659$; this finite-scale transition is part of the result, not a removed outlier.

Scope of the numerical comparison. At $c = 1$, the largest logit differences between reference continuation and a reference head applied to the native endpoint are 2.75×10^{-5} (Qwen) and 1.32×10^{-4} (SmolLM2). These local agreement checks do not identify the two arithmetic policies at every large scale. Offline replay verifies all 96 conditions and 9180 nonempty target-position diagnostics, including exact constancy of each predicted coefficient across scales. It reaggregates retained diagnostic arrays; model-level regeneration uses the separately supplied runner and weights. The new panel tests E_0 and the leading coefficient; E_1 has not been measured on these pretrained models.

9 Related Work and Implications

Symmetry and local output expansions. Rescaling-invariant learning predates this construction (Neyshabur et al., 2015); our object is a channel-calibrated edit in a unique balanced SwiGLU representation. The categorical Fisher covariance and KL Taylor expansion are standard (Martens,

2020). We use these local readout identities to obtain the output coefficient of a specified finite residual suffix, not as independent discoveries.

Architecture and normalization. The exact transport/reaction and vorticity organization builds on Liang (2026). RMSNorm and gated FFNs are established components (Zhang and Sennrich, 2019; Shazeer, 2020). Normalization placement affects training and propagation (Xiong et al., 2020; Kim et al., 2025); Sun et al. (2025) studies depth-dependent layer ineffectiveness. Here weights and depth are fixed while the input radius varies. Our coefficients depend on frozen branch values and the output law, without a training-variance or isotropy assumption. Massive activations (Sun et al., 2024), fixed-head logit lenses (nostalgebraist, 2020) and learned lenses (Belrose et al., 2023) provide empirical context; direct readout here is the declared no-suffix comparator, not a claim to extract all information in a hidden state.

10 Conclusion

A structural edit must be specified consistently before its propagation can be interpreted. Balanced channel coordinates define one such tying operation. Gates and normalization determine the local derivative, and residual radius and direction have different roles in later computation and final readout. For a fixed suffix, these distinctions yield an explicit expansion with independently computable output coefficients. The two-model reference experiment supports coefficient convergence while retaining the model-dependent transition regimes.

The finite tying experiments and radius-scaling experiments test complementary interventions. The balanced-reference panel now tests the entire weak-to-strong path and its coordinate consistency with reference-specific controls. Removing an arbitrary channel-scale dependence defines a coherent intervention; it does not require that intervention to preserve the original language-model function. Historical checkpoint dependence, oracle-state access and the six-passage coverage are documented with the full results. The coefficient panel uses the three existing synthetic passages. No new training result or uniform finite-edit approximation is inferred.

REPRODUCIBILITY STATEMENT

The source package contains the complete manuscript, retained empirical records, fixed measurement protocols, model runners, offline replay code and synthetic checks for the results stated here. Offline replay checks all 2016 balanced/scalar-reference primary cells and 288 coordinate checks, all 120 same-scale conditions, and all 96 coefficient conditions, including 9180 nonempty coefficient target-position diagnostics. It reaggregates saved diagnostics and the balanced-panel per-target loss/divergence records, not the full unretained vocabulary logits. The uninterrupted balanced-panel design uses 2064 backbone evaluations. Timeout recovery adds documented checks: the saved records count at least 2160 calls, with possible unpersisted work at interruption. Head-only chunks and repeated zero-strength declarations are not counted as new backbone evaluations. Model-level regeneration uses the pinned checkpoints and the separately specified measurement environment. A clean extracted package builds without access to any other manuscript. Historical E4b checkpoints and parts of their original deployment provenance remain unavailable, as detailed in the empirical appendices.

AI USE STATEMENT

A generative AI assistant helped organize research records, retrieve related work, formulate and check algebraic claims, design the exploratory comparisons and sequential forward controls, implement and test measurement and analysis code, execute public-weight model forwards, interpret results, prepare figures and anonymous reproduction materials, generate the three synthetic evaluation passages, extract and tokenize the fixed WikiText sample, and draft and edit manuscript sections. Training-trajectory measurements are inherited from archived experiments; the public-weight intervention results were newly computed. The natural-text validation uses attributed WikiText excerpts; these texts were not AI-generated.

REFERENCES

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023. https://arxiv.org/abs/2303.08112.

nostalgebraist. Interpreting GPT: the logit lens. *LessWrong*, 31 August 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When Smol Goes Big—Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737*, 2025. https://arxiv.org/abs/2502.02737. Model: https://huggingface.co/HuggingFaceTB/SmolLM2-360M.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*, 2019. https://arxiv.org/abs/1902.09689.

Tamir David Hay and Lior Wolf. Dynamic Layer Tying for Parameter-Efficient Transformers. *arXiv preprint arXiv:2401.12819*, 2024. https://arxiv.org/abs/2401.12819.

Jeonghoon Kim, Byeongchan Lee, Cheonbok Park, Yeontaek Oh, Beomjun Kim, Taehwan Yoo, Seongjin Shin, Dongyoon Han, Jinwoo Shin, and Kang Min Yoo. Peri-LN: Revisiting normalization layer in the Transformer architecture. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:30400–30436, 2025. https://proceedings.mlr.press/v267/kim25u.html.

Zhihua Liang. The Geometry of Semantic Space: A Continuous Geometric Framework for the Transformer Architecture. *arXiv preprint arXiv:2607.17146*, 2026. https://arxiv.org/abs/2607.17146.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. *arXiv preprint arXiv:1609.07843*, 2016. https://arxiv.org/abs/1609.07843.

Qwen Team. Qwen3-0.6B model card and weights, 2025. Revision c1899de289a04d12100db370d81485cdf75e47ca. https://huggingface.co/Qwen/Qwen3-0.6B.

Noam Shazeer. GLU variants improve Transformer. *arXiv preprint arXiv:2002.05202*, 2020. https://arxiv.org/abs/2002.05202.

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

Jingcun Wang, Yu-Guang Chen, Ing-Chao Lin, Bing Li, and Grace Li Zhang. Basis Sharing: Cross-Layer Parameter Sharing for Large Language Model Compression. *arXiv preprint arXiv:2410.03765*, 2024. https://arxiv.org/abs/2410.03765.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the Transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:10524–10533, 2020. https://proceedings.mlr.press/v119/xiong20b.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 32, 2019. https://arxiv.org/abs/1910.07467.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. *arXiv preprint arXiv:1506.02617*, 2015. https://arxiv.org/abs/1506.02617.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The Curse of Depth in Large Language Models. *arXiv preprint arXiv:2502.05795*, 2025. https://arxiv.org/abs/2502.05795.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. https://www.jmlr.org/papers/v21/17-678.html.

A CANONICAL CHANNEL BALANCING

Proposition 2 (Canonical balancing under positive channel scaling)**.** *Assume* $\|u_j\| > 0$ *and* $\|d_j\| > 0$ *on every channel in this statement. Define*

$$C_* = \operatorname{diag}\left(\sqrt{\frac{\|d_j\|}{\|u_j\|}}\right), \qquad U_* = C_* U, \quad D_* = D C_*^{-1}.$$

This is the unique representative on the positive diagonal scaling orbit whose corresponding row and column norms agree. It satisfies

$$\|(U_*)_{j:}\| = \|(D_*)_{:j}\| = \sqrt{\|u_j\|\|d_j\|}.$$

Under any positive diagonal C*, balancing* (CU, DC^{-1}) *yields the same pair* (U_*, D_*)*. Applying ordinary transpose tying in these balanced coordinates produces precisely the weighted-adjoint branch*

$$f_*(x) = U^\top M z_U(x), \qquad M = C_*^2. \tag{19}$$

Proof. Balancing requires $c_j\|u_j\| = \|d_j\|/c_j$, whose unique positive solution is the displayed c_j. A preliminary scaling by C changes the balancing factor to $C_* C^{-1}$; substitution leaves both balanced matrices unchanged. Finally $U_*^\top z_{U_*} = U^\top C_*^2 z_U$. □

Positive diagonal scaling commutes with elementwise gating; arbitrary dense changes of hidden basis do not. Rescaling symmetry is classical (Neyshabur et al., 2015; Martens, 2020); the construction here identifies exactly which balanced-coordinate operation the proposed SwiGLU edit performs. Zero channels are treated in Appendix E rather than by dividing by a zero norm.

B PROOF OF THE FIXED-SUFFIX EXPANSION AND OUTPUT COEFFICIENTS

B.1 LOCAL CATEGORICAL-READOUT EXPANSION

For finite logits z, let $A(z) = \log \sum_y e^{z_y}$ and $p_z = \operatorname{softmax}(z)$. Differentiating the finite sum gives $DA(z) = p_z$ and $D^2 A(z) = \operatorname{diag}(p_z) - p_z p_z^\top$. Hence

$$\mathrm{KL}(p_z \| p_{z+\Delta}) = A(z+\Delta) - A(z) - p_z^\top \Delta = \tfrac{1}{2} \operatorname{Var}_{p_z}(\Delta) + O(\|\Delta\|^3). \tag{20}$$

Reverse KL has the same quadratic term. These are standard categorical information-geometry identities (Martens, 2020). Adding a common scalar to all logits leaves the softmax exactly unchanged. These local facts, together with the finite residual recursion, suffice for the following proofs.

Proposition 3 (Exact finite-chord curvature and a variance bound)**.** *For any finite logit increment* Δ*, let* $p_s = \operatorname{softmax}(z + s\Delta)$ *and* $w = \operatorname{osc}(\Delta) = \max_y \Delta_y - \min_y \Delta_y$*. Then*

$$\mathrm{KL}(p_0 \| p_1) = \int_0^1 (1-s) \operatorname{Var}_{p_s}(\Delta)\, ds, \tag{21}$$

$$\mathrm{KL}(p_1 \| p_0) = \int_0^1 s \operatorname{Var}_{p_s}(\Delta)\, ds. \tag{22}$$

For $w > 0$*, writing* $V_0 = \operatorname{Var}_{p_0}(\Delta)$ *gives*

$$\frac{e^{-w} - 1 + w}{w^2} V_0 \le \mathrm{KL}(p_0 \| p_1) \le \frac{e^{w} - 1 - w}{w^2} V_0. \tag{23}$$

At $w = 0$ *both KL and* V_0 *vanish; the displayed multipliers have continuous limit* $1/2$*.*

Proof. Set $a(s) = A(z + s\Delta)$. Then $a''(s) = \operatorname{Var}_{p_s}(\Delta)$; the two integral remainders give Eqs. (21)–(22). Also $e^{-sw} \le p_s(y)/p_0(y) \le e^{sw}$. Since variance is $\min_b \mathbb{E}(\Delta - b)^2$, these density bounds imply $e^{-sw} V_0 \le \operatorname{Var}_{p_s}(\Delta) \le e^{sw} V_0$. Integrating $(1-s)e^{\pm sw}$ proves Eq. (23). □

The variance bound retains how strongly each vocabulary direction is populated. It can be substantially smaller than replacing every variance by $w^2/4$. It is a finite-change statement about a known endpoint increment, not a way to obtain that increment without computing the network.

B.2 A SHARP LARGE-RADIUS THEOREM, INCLUDING ITS COEFFICIENT

Let the suffix have s sublayers

$$X_{j+1} = X_j + F_j(N_{j,\epsilon_j}(X_j)), \qquad X_0 = cH,$$

where $H_i \neq 0$ for every token and

$$N_{j,\epsilon}(X)_i = \frac{\sqrt{d}\,\Gamma_j X_i}{\sqrt{\|X_i\|^2 + d\epsilon}}.$$

All epsilons are fixed and nonnegative. Assume the F_j are C^3 on neighborhoods of the normalized limiting states. They may couple all permitted tokens. Define the zero-epsilon, degree-zero limiting vector fields and their frozen values by

$$q_j^0(X) = F_j(N_{j,0}(X)), \qquad a_j = q_j^0(H), \qquad L_j = Dq_j^0(H). \tag{24}$$

These frozen evaluations use the *same* direction state H for every sublayer, with that sublayer's own parameters and normalization gains. They are not the usual clean suffix trajectory.

Theorem 2 (Frozen-ray expansion and output coefficients)**.** *Put*

$$E_0 = \sum_j a_j, \qquad E_1 = \sum_{j<l} L_l a_j. \tag{25}$$

The complete suffix obeys

$$\mathcal{S}(cH) = cH + E_0 + c^{-1}E_1 + O(c^{-2}). \tag{26}$$

Let $A_f = W\Gamma_f$, and compare direct and suffix outputs using the same final RMS readout with fixed $\epsilon_f \geq 0$. For token i, write $H_i = r_i u_i$, $P_i = I - u_i u_i^\top$, and define

$$z_{\infty,i} = \sqrt{d}A_f u_i, \qquad v_i = \frac{\sqrt{d}}{r_i}A_f P_i E_{0,i}, \qquad p_{\infty,i} = \mathrm{softmax}(z_{\infty,i}). \tag{27}$$

Then

$$\Delta z_i(c) = c^{-1}v_i + O(c^{-2}), \tag{28}$$

$$\lim_{c\to\infty} c^2\,\mathrm{KL}(p_i^{\mathrm{dir}}(c)\|p_i^{\mathrm{suf}}(c)) = \frac{1}{2}\,\mathrm{Var}_{p_{\infty,i}}(v_i). \tag{29}$$

Reverse KL has the same limit. For any fixed target y,

$$\lim_{c\to\infty} c\,\Delta\mathrm{NLL}_{y,i}(c) = (p_{\infty,i} - e_y)^\top v_i. \tag{30}$$

In particular the KL is $\Theta(c^{-2})$ exactly when v_i is nonconstant across vocabulary coordinates. If v_i is constant, both KL directions are $O(c^{-4})$.

Proof. Set $t = 1/c$ and $Y_j(t) = tX_j(cH)$. Then $Y_0 = H$ and exactly

$$Y_{j+1} = Y_j + t\,q_j^t(Y_j), \qquad q_j^t(X) = F_j\left(\left(\frac{\sqrt{d}\,\Gamma_j X_i}{\sqrt{\|X_i\|^2 + d\epsilon_j t^2}}\right)_{i=1}^n\right).$$

On a neighborhood of H with every token radius bounded away from zero, $q_j^t = q_j^0 + O(t^2)$, with the corresponding local derivative control. Induction through the finite recursion gives

$$Y_s = H + t\sum_j a_j + t^2\sum_{j<l} L_l a_j + O(t^3).$$

Dividing by t proves Eq. (26). No statistical or isotropy assumption is used.

Write R_t for final readout in these rescaled coordinates. It has the expansion $R_t = R_0 + t^2 C_f + O(t^4)$ locally. In the difference $R_t(Y_s) - R_t(H)$ the common order-t^2 epsilon term cancels, giving

$$\Delta z = tDR_0(H)E_0 + t^2\left(DR_0(H)E_1 + \tfrac{1}{2}D^2R_0(H)[E_0, E_0]\right) + O(t^3). \tag{31}$$

Since $DR_{0,i}(H_i)e = \sqrt{d}A_f P_i e/r_i$, this proves Eq. (28). The base logits converge to z_∞ with error $O(t^2)$. Applying the local KL expansion and differentiating NLL proves Eqs. (29)–(30). Finite logits imply full support, so variance is positive exactly for nonconstant v_i. If v_i is constant, remove its common logit shift without changing softmax; the remaining increment is $O(t^2)$ and its KL is $O(t^4)$. □

The second coefficient is explicit. Let $\gamma_i = u_i^\top E_{0,i}$ and $v_{H,i} = P_i E_{0,i}$. The coefficient of c^{-2} in Eq. (31) is

$$b_i = \sqrt{d} A_f \left[\frac{P_i E_{1,i}}{r_i} - \frac{\gamma_i v_{H,i}}{r_i^2} - \frac{\|v_{H,i}\|^2 u_i}{2 r_i^2} \right]. \tag{32}$$

When v_i is a common shift, $\lim c^4 \,\mathrm{KL} = \mathrm{Var}_{p_{\infty,i}}(b_i)/2$ in both directions. This gives the next nonzero output coefficient when the first term is a common shift.

Coefficient evaluation. Section 8 tests the leading coefficient on a fixed two-model reference panel. The explicit second coefficient is a derived result and has not been tested by that pretrained panel.

C Local rotation and residual-state transport

Gates and normalization. Differentiating both paths of the SwiGLU branch gives

$$K(x) = D\big[\mathrm{diag}(\phi(Gx))U + \mathrm{diag}(Ux)\,\mathrm{diag}(\phi'(Gx))G\big]. \tag{33}$$

The first tied summand is symmetric, while the independent gate contributes the second term. Use $N(h) = \Gamma \rho_\delta(h)$ with $\rho_\delta(h) = \sqrt{d}\, h/\sqrt{\|h\|^2 + \delta}$ and $\delta = d\epsilon$ for standard RMSNorm. At $h = ru$, $\|u\| = 1$, write

$$J_\rho = aI - buu^\top, \qquad a = \frac{\sqrt{d}}{\sqrt{r^2+\delta}}, \quad b = \frac{\sqrt{d}\, r^2}{(r^2+\delta)^{3/2}}. \tag{34}$$

Absorb the actual learned gains into $A = K(N(h))\Gamma = S + B$, where $S^\top = S$ and $B^\top = -B$. The residual Jacobian is $I + AJ_\rho$. The inherited vorticity identity (Liang, 2026) then reads

$$\Omega = (AJ_\rho)^\top - AJ_\rho = -[S, J_\rho] - \{B, J_\rho\}. \tag{35}$$

Proposition 4 (A rank-two normalization component)**.** *Set $v_S = (I - uu^\top)Su$. For every effective derivative $A = S + B$, its symmetric-normalization contribution is*

$$\Omega_S = -[S, J_\rho] = b(v_S u^\top - u v_S^\top), \quad \mathrm{rank}\,\Omega_S \le 2, \quad \|\Omega_S\|_{\mathrm{op}} = b\|v_S\|_2. \tag{36}$$

It acts only on $\mathrm{span}\{u, v_S\}$ and vanishes exactly when Su is parallel to u. If $B = 0$, this is the full vorticity.

The statement identifies the contribution of symmetric deformation even in the general gated case. The remaining term in Eq. (35) captures the skew derivative. This is local Jacobian geometry; finite residual transport is described separately below. Proofs and the discrete Euler norm identity are in Appendix E.

Exact radial–directional transport. For a token update $h^+ = h + q$, decompose $h = ru$ and $q = \gamma u + v$, where $\gamma = u^\top q$ and $v \perp u$. When $h^+ \neq 0$,

$$r^+ = \sqrt{(r+\gamma)^2 + \|v\|^2}, \qquad u^+ = \frac{(r+\gamma)u + v}{r^+}, \qquad u^+ - u = \frac{v}{r} + O\!\left(\frac{\|q\|^2}{r^2}\right). \tag{37}$$

The expansion is for $\|q\|/r \to 0$; the first two equalities are exact. Attention may couple the whole sequence in q, without changing this tokenwise decomposition. Radius sets the relative angular step of a bounded increment. In contrast, the final normalized readout is

$$\mathcal{R}_\epsilon(ru) = \frac{\sqrt{d}\, W \Gamma_f u}{\sqrt{1 + d\epsilon/r^2}}. \tag{38}$$

It depends only on the positive-ray direction at $\epsilon = 0$, with an explicit finite-ϵ radial correction. Thus radius controls subsequent transport while direction governs normalized output.

D A FINITE-RADIUS BOUND ON SUFFIX INFLUENCE

Consider a fixed-length sequence $H = (h_i)$ followed by a fixed finite suffix of pre-normalized residual sublayers, $H_{j+1} = H_j + F_j(N_j(H_j))$, with output $\mathcal{S}(H)$. Bounded normalization ranges and continuity of each branch on their closures give finite tokenwise bounds $\|F_j(N_j(Y))_i\| \le B_j$ for all Y. Set $B_{\mathrm{suf}} = \sum_j B_j$. This premise includes fixed-weight SwiGLU and fixed-mask attention on RMS-normalized inputs; explicit bounds are given in Appendix E.

Theorem 3 (Observable influence of a fixed suffix)**.** *Let $r_{\min} = \min_i \|h_i\| > B_{\mathrm{suf}}$ and use the same readout for H and $\mathcal{S}(H)$. Define*

$$\eta = \sqrt{d}\,\|W\Gamma_f\|_{2\to\infty}\frac{B_{\mathrm{suf}}}{r_{\min} - B_{\mathrm{suf}}}. \tag{39}$$

For each token, the logit difference and any target NLL difference satisfy

$$\|\Delta\mathrm{logits}\|_\infty \le \eta, \qquad |\Delta\mathrm{NLL}| \le 2\eta, \qquad \mathrm{KL}(p\|q),\ \mathrm{KL}(q\|p) \le \eta^2/2, \tag{40}$$

where p, q are the two softmax distributions.

Proof sketch. Telescoping gives $\mathcal{S}(H)_i = h_i + e_i$ with $\|e_i\| \le B_{\mathrm{suf}}$. The segment joining these states has radius at least $r_{\min} - B_{\mathrm{suf}}$. Integrating the RMSNorm derivative, whose norm is at most $\sqrt{d}/\|h\|$, gives the logit bound. The log-sum-exp increment lies between the smallest and largest logit increments. Its second directional derivative is their variance, at most $\|\Delta\mathrm{logits}\|_\infty^2$, giving the NLL and both KL bounds. Appendix E supplies the full argument. □

For a fixed nonzero state per token and a scale-independent suffix budget, comparison at the identical scaled input cH yields

$$\Delta\mathrm{logits} = O(c^{-1}), \qquad \Delta\mathrm{NLL} = O(c^{-1}), \qquad \mathrm{KL} = O(c^{-2}). \tag{41}$$

These are upper rates, not required empirical slopes. The increments may remain nonzero while their output influence attenuates. At finite radius, a verified net-displacement bound or measured sum of increment norms gives a path-specific version of Eq. (39).

A direction-resolved certificate. The worst-case budget can be loose. Appendix E.5 derives an exact endpoint decomposition of the logit change into radial and tangential terms, and bounds NLL and KL using its vocabulary range. This certificate remains applicable beyond $r_{\min} > B_{\mathrm{suf}}$ and gives zero influence for any positive purely radial displacement at $\epsilon = 0$. It uses the actual suffix endpoint; it is not a predictor that avoids running the suffix.

E PROOFS FOR THE SCALE AND RESIDUAL-STATE RESULTS

E.1 FEATURE METRIC AND CALIBRATED OBSERVABLES

All feature transformations here act on a bias-free branch, keep G fixed, and use positive invertible diagonal C. Since C commutes with $\mathrm{diag}(\phi(Gx))$, $z_{CU} = Cz_U$, proving original function preservation for every input, not only calibration tokens. In the new coordinates, unit transpose replacement is $(CU)^\top Cz_U = U^\top C^2 z_U$. In contrast,

$$(CU)^\top (C^{-\top} M C^{-1})(Cz_U) = U^\top M z_U.$$

For $M_{jj} = \|d_j\|/\|u_j\|$, the norm transformations give $M_{C,jj} = M_{jj}/c_j^2$ and $(T_M)_{:j} = M_{jj}u_j$, of norm $\|d_j\|$. Summing squared column norms proves $\|T_M\|_F = \|D\|_F$. When $D = U^\top$, $M = I$. Among replacements whose jth column is a positive multiple of u_j and has norm $\|d_j\|$, this column is uniquely determined.

The norm-ratio construction is stated on nonzero channels. If $u_j = 0$, the original gated feature is identically zero and its effective branch contribution can be omitted. If $d_j = 0$ with $u_j \neq 0$, setting $M_{jj} = 0$ keeps the column zero but produces a positive-semidefinite, not positive-definite, metric. The proposition uses the nondegenerate active-channel domain. It does not extend the feature symmetry to arbitrary dense basis changes, which generally fail to commute with the gate.

For $C = cI$, $\alpha_c = \|D/c\|_F/\|cU\|_F = \alpha/c^2$. Thus $\alpha_c(cU)^\top(cz_U) = \alpha U^\top z_U$. Original activations are unchanged and the singular directions can be chosen consistently; the squared singular value, and hence g_ℓ, acquires c^2. Therefore αg_ℓ is invariant even though raw g_ℓ ranges over all positive numbers if its starting value is positive. A repeated leading singular value requires a direction convention; no claim of directional identifiability is obtained from scalar invariance. Scalar calibration alone is generally not invariant under nonuniform channel scaling.

E.2 Gated derivative and the normalization plane

Differentiating each factor of $\phi(Gx) \odot Ux$ gives Eq. (8). Under unit tying its first term is $U^\top \operatorname{diag}(\phi(Gx))U$, symmetric but potentially indefinite for SiLU. The second term contains the independent gate derivative. For standard RMSNorm $N(h) = \Gamma h/\sqrt{\operatorname{mean}(h^2) + \epsilon}$, set $\delta = d\epsilon$ and $A = K(N(h))\Gamma$. Differentiation gives Eq. (34). With $A = S + B$,

$$(AJ_\rho)^\top - AJ_\rho = J_\rho(S - B) - (S + B)J_\rho = -[S, J_\rho] - \{B, J_\rho\}.$$

For its symmetric contribution,

$$-[S, J_\rho] = b(Suu^\top - uu^\top S).$$

Write $Su = (u^\top Su)u + v_S$ with $v_S \perp u$. The radial terms cancel, leaving Eq. (36). This matrix annihilates the orthogonal complement of $\operatorname{span}\{u, v_S\}$. If $v_S \neq 0$, its matrix on the orthonormal basis $u, v_S/\|v_S\|$ has off-diagonal entries $\pm b\|v_S\|$, proving its rank and norm. Since $b > 0$ for $r > 0$, vanishing is equivalent to $v_S = 0$. The result applies to this component for arbitrary B, and to the full vorticity only when $B = 0$ or its remaining contribution vanishes.

Our vorticity convention is $\Omega = J^\top - J$; the usual skew part is $-\Omega/2$. The identity in the full residual Jacobian contributes no skew part. Local rotation is distinct from norm preservation of a finite Euler step: for real skew B,

$$\|(I + tB)u\|^2 = \|u\|^2 + t^2\|Bu\|^2. \tag{42}$$

This follows by expanding the square and using $u^\top Bu = 0$.

E.3 Polar transport and final readout

Orthogonality gives $\|h + q\|^2 = (r + \gamma)^2 + \|v\|^2$ and the exact formulas in Eq. (37). The derivative of $h \mapsto h/\|h\|$ at ru is $(I - uu^\top)/r$. Taylor expansion for $\|q\|/r \to 0$ gives the displayed angular expansion. For the continuous flow $\dot{h} = q(h)$,

$$\dot{r} = u^\top q(h), \qquad \dot{u} = (I - uu^\top)q(h)/r.$$

Equation (38) follows by substituting $h = ru$ in final RMSNorm and the fixed head. The fixed-direction finite-ϵ correction is $O(\epsilon/r^2)$ for fixed d and head. Positive rays do not identify u and $-u$.

E.4 Bounded updates for a finite pre-normalized suffix

Fix weights, sequence length, masks, and a finite number of sublayers. A bounded normalization image has a compact closure in the finite-dimensional sequence space. A branch continuous on that closure attains a finite maximum token norm, yielding B_j. Standard RMSNorm has token norm at most $\sqrt{d}\|\Gamma\|_{\mathrm{op}}$ for nonnegative epsilon on its domain. At epsilon zero exclude undefined zero-state evaluations; positive epsilon gives a globally defined normalization.

For a SwiGLU branch on $\|x\| \le M$, $|\mathrm{SiLU}(t)| \le |t|$ and $\|a \odot b\|_2 \le \|a\|_2\|b\|_2$ imply

$$\|D[\phi(Gx) \odot Ux]\|_2 \le \|D\|_{\mathrm{op}}\|G\|_{\mathrm{op}}\|U\|_{\mathrm{op}}M^2.$$

A fixed-mask attention head with at least one allowed key per query is a convex combination of value vectors. Their norms are bounded on normalized inputs; finite concatenation of heads and a fixed output map remain bounded. Fixed biases, if present in attention, can be included in the bounds. The feature-symmetry result separately uses bias-free SwiGLU.

Telescoping along the actual suffix path gives $e_i = \mathcal{S}(H)_i - h_i$ with $\|e_i\| \le \sum_j B_j = B_{\mathrm{suf}}$. For every $t \in [0,1]$, $\|h_i + te_i\| \ge \|h_i\| - \|e_i\| \ge r_{\min} - B_{\mathrm{suf}}$. The bare final radial derivative has eigenvalues

$$\lambda_{\tan} = \frac{\sqrt{d}}{\sqrt{\|h\|^2 + d\epsilon}}, \qquad \lambda_{\mathrm{rad}} = \frac{\sqrt{d}\, d\epsilon}{(\|h\|^2 + d\epsilon)^{3/2}}.$$

For $\epsilon \ge 0$ both are bounded by $\sqrt{d}/\|h\|$ on that segment. Integrating and multiplying by the head gives the first bound in Eq. (40). Here $\|A\|_{2\to\infty} = \max_v \|A_{v:}\|_2$ is the maximum row norm.

For completeness, let $z' = z + \Delta$, $\Lambda(z) = \log \sum_v e^{z_v}$ and $w = \max_v \Delta_v - \min_v \Delta_v$. The log-sum-exp increment lies in the interval between the smallest and largest coordinate increments; subtracting Δ_y gives $|\Delta \mathrm{NLL}_y| \le w$. With $a(t) = \Lambda(z + t\Delta)$,

$$a''(t) = \mathrm{Var}_{\mathrm{softmax}(z+t\Delta)}(\Delta_v) \le w^2/4.$$

Its integral remainder $\int_0^1 (1-t)a''(t)\, dt$ equals $\mathrm{KL}(\mathrm{softmax}(z) \| \mathrm{softmax}(z'))$, and is at most $w^2/8$. Interchanging endpoints gives the reverse direction. Since $w \le 2\|\Delta\|_\infty \le 2\eta$, Eq. (40) follows. The same bounds hold for fixed-weight averages over targets/tokens.

For cH the uniform budget is unchanged and $r_{\min}(cH) = c r_{\min}(H)$. Substitution proves Eq. (41). Positive tokenwise rescalings whose minimum tends to infinity have the analogous bound through their minimum input radius. These are sufficient bounds: faster convergence, exact equality, or cancellation is compatible with them. The theorem does not assert that every large pooled-standard-deviation excess $\mathcal{E}_L$ has a small budget-to-radius ratio.

E.5 An endpoint certificate resolving radial and tangential effects

This refinement uses the actual suffix endpoints rather than a uniform worst-case budget. For a token, let $h = ru$, $e = \gamma u + v$ with $v \perp u$, and $h' = h + e \ne 0$. Put

$$s = \sqrt{(r+\gamma)^2 + \|v\|^2 + d\epsilon}, \qquad a = \frac{r+\gamma}{s} - \frac{r}{\sqrt{r^2 + d\epsilon}}, \qquad A_f = W\Gamma_f.$$

The exact logit difference is

$$\Delta z = \sqrt{d}\left[aA_f u + \frac{A_f v}{s} \right]. \tag{43}$$

Define $\mathrm{osc}(x) = \max_v x_v - \min_v x_v$ across vocabulary coordinates, and

$$\xi = \sqrt{d}\left[|a| \,\mathrm{osc}(A_f u) + \frac{\mathrm{osc}(A_f v)}{s} \right]. \tag{44}$$

Then $\mathrm{osc}(\Delta z) \le \xi$ and the preceding log-sum-exp argument proves

$$|\Delta \mathrm{NLL}_y| \le \xi, \qquad \mathrm{KL}(p\|q),\ \mathrm{KL}(q\|p) \le \xi^2/8. \tag{45}$$

To prove Eq. (43), substitute $h' = (r+\gamma)u + v$ in the exact readout and subtract the readout of ru. The oscillation inequality follows from subadditivity and positive homogeneity of oscillation, using $\mathrm{osc}(ax) = |a| \,\mathrm{osc}(x)$.

No condition $r > B_{\mathrm{suf}}$ is needed here. In particular, if $\epsilon = 0$, $v = 0$ and $r + \gamma > 0$, then $a = 0$ and $\Delta z = \xi = 0$, for arbitrarily large positive radial displacement. For finite epsilon the residual radial contribution is explicit. This certificate also removes vocabulary-common logit shifts from the distributional bound. It is an a posteriori output certificate, not a method for predicting the suffix endpoint or skipping its computation. It refines the finite-radius theorem's observable consequences without changing any existing experimental measurements.

F Frozen-ray coefficient experiment: protocol and complete outcomes

The measurement protocol fixed the two existing public model/tokenizer revisions, all three existing synthetic token sequences, architecture-relative cuts $\{L/4, L/2, L-1, L\}$ and scales

$\{64, 512, 4096, 32768\}$ before pretrained execution. No model outcome selects a passage, cut, scale or token. The coefficient at a cut is evaluated from all remaining branches at the same clean cut state; it is not fitted to any suffix output, logit change or loss. For Qwen the model revision is `c1899de289a04d12100db370d81485cdf75e47ca`; for SmolLM2 it is `f8027fd0eaeea54caa13c31d31b9fdc459c38b49`. Model-file hashes and exact input IDs match the preceding study.

Arithmetic and branch definition. Native float32 clean forwards capture all needed cut states, attention masks and RoPE arrays; these states are cast to double. The checkpoint is reloaded in double without modifying any stored weight value. The new reference explicitly retains double in RMSNorm and eager attention softmax. Original fixed RoPE arrays are cast to double, with no recomputation across scales. Query/key normalization internal to Qwen attention retains its saved epsilon. Only the outer residual RMS epsilon becomes ϵt^2 in the rescaled recurrence $Y = X/c$, as required by the exact coordinate identity. The final epsilon scales by the same rule in both direct and suffix readouts. Therefore the new panel tests the mathematical stored-weight reference at fixed original epsilon, separately from the native-float32 measurements.

Frozen quantities. For each branch $q_j^0(H)$, the outer RMS epsilon is zero, while attention's internal normalization parameters and all other operations stay fixed. Compute $E_0 = \sum_j q_j^0(H)$, then each v_i, $p_{\infty,i}$ and $C_i = \mathrm{Var}_{p_{\infty,i}}(v_i)/2$. The implementation verifies that these coefficient diagnostics are byte-identical across all four scales. None uses the subsequently computed suffix response. This panel measures the leading coefficient; the second coefficient E_1 is not evaluated.

KL evaluation. Let $\Delta = z' - z$ and $\widetilde{\Delta} = \Delta - \mathbb{E}_{p_z}\Delta$. Evaluate

$$\mathrm{KL}(p_z \| p_{z'}) = \log\left(1 + \mathbb{E}_{p_z}[\exp(\widetilde{\Delta}) - 1 - \widetilde{\Delta}]\right).$$

The code uses `expm1` and a local series for $\exp(v) - 1 - v$ at $|v| < 10^{-3}$. This avoids subtracting nearly equal log-partition values at large c. The reverse direction is evaluated independently. Six tiny-model tests, including Qwen3 and Llama graphs, verify algebraic rescaling, the coefficient and KL formulas, common shifts, exact empty suffixes and agreement with the decoder on controlled inputs before pretrained execution.

Metrics and crosschecks. The main ratios are ratios of target sums. The Fisher-weighted coefficient error is

$$E_{\mathrm{F}}(c) = \sqrt{\frac{\sum_i \mathrm{Var}_{p_{\infty,i}}(c\Delta z_i - v_i)}{\sum_i \mathrm{Var}_{p_{\infty,i}}(v_i)}}.$$

Signed loss coefficients $(p_{\infty,i} - e_{y_i})^\top v_i$ and their errors, and $\|c(Y_s - H) - E_0\|$, are also retained. At $c = 1$, each nonempty cut is continued in the reference and compared with a reference final head on the native endpoint. Maximal logit discrepancies are 2.74875×10^{-5} for Qwen and 1.31526×10^{-4} for SmolLM2; state discrepancies are 0.004224 and 0.006782. These are measured agreement checks, not an assumed identity of floating-point policies at arbitrary scale.

Inventory and reproducibility. The panel contains 96 conditions, with 9180 nonempty prediction-target conditions. Separately: six native full forwards, 72 nonempty scaled reference suffix evaluations, 24 empty controls, 18 additional $c = 1$ reference continuations and 462 frozen branch evaluations (216 Qwen, 246 SmolLM2). The cumulative native full-forward count is $2064 + 6 = 2070$. Model computations use Python 3.12.3, PyTorch 2.8.0+cpu, Transformers 4.56.2, NumPy 2.2.6 and two CPU threads. The per-target NPZ contains scalar diagnostic arrays, not full logit vectors or E_0 vectors. Offline replay checks its exact inventory, shape, finiteness, hashes, scale-invariant coefficient fields and all recorded reductions. Recomputing those scalar measurements themselves requires the published measurement code and checkpoints. Scientific closeness of any coefficient ratio to one is never an acceptance condition for storing a run.

Table 4: All coefficient-panel cut/scale aggregates. Empty-cut ratios are undefined (zero predictions and zero measured changes), not excluded.

Model	Cut	c	E_z	$R_{\rightarrow}$	$R_{\leftarrow}$	E_{F}
Qwen	7	64	1.00156	0.0658833	0.0621806	0.969818
Qwen	7	512	0.444572	0.571036	0.595143	0.359595
Qwen	7	4096	0.0964871	0.935713	0.935021	0.0531586
Qwen	7	32768	0.0128859	0.991338	0.98982	0.00671795
Qwen	14	64	0.821504	0.0929075	0.0891945	0.892374
Qwen	14	512	0.215352	0.663961	0.654988	0.219767
Qwen	14	4096	0.0352998	0.947832	0.939131	0.0296175
Qwen	14	32768	0.00458812	0.993299	0.99202	0.00372864
Qwen	27	64	0.00646993	1.00053	0.998155	0.0056803
Qwen	27	512	0.000807741	1.00006	0.999758	0.000708629
Qwen	27	4096	0.000100952	1.00001	0.99997	8.85566e-05
Qwen	27	32768	1.26187e-05	1	0.999996	1.10692e-05
Qwen	28	64	–	–	–	–
Qwen	28	512	–	–	–	–
Qwen	28	4096	–	–	–	–
Qwen	28	32768	–	–	–	–
SmolLM2	8	64	0.264017	1.08829	1.37956	0.140519
SmolLM2	8	512	0.035968	1.00928	1.03867	0.0197209
SmolLM2	8	4096	0.00453859	1.00111	1.00458	0.00250099
SmolLM2	8	32768	0.000567976	1.00014	1.00057	0.000313185
SmolLM2	16	64	0.157154	1.08236	1.28438	0.0884983
SmolLM2	16	512	0.0196794	1.01382	1.03683	0.0109729
SmolLM2	16	4096	0.00246286	1.00177	1.00456	0.00137149
SmolLM2	16	32768	0.000307911	1.00022	1.00057	0.000171438
SmolLM2	31	64	0.00714331	0.986522	0.972651	0.00631605
SmolLM2	31	512	0.000895095	0.998309	0.996525	0.000782547
SmolLM2	31	4096	0.000111919	0.999789	0.999565	9.77111e-05
SmolLM2	31	32768	1.39904e-05	0.999974	0.999946	1.22122e-05
SmolLM2	32	64	–	–	–	–
SmolLM2	32	512	–	–	–	–
SmolLM2	32	4096	–	–	–	–
SmolLM2	32	32768	–	–	–	–

G SAME-SCALE SUFFIX EXPERIMENT: PROTOCOL, NUMERICS AND COMPLETE CURVES

Fixed inputs and executed panel. The protocol was committed after revision-4 theory and the previous edit/state results, before this panel's pretrained-model forwards. Model and tokenizer revisions remain `c1899de289a04d12100db370d81485cdf75e47ca` (Qwen3-0.6B) and `f8027fd0eaeea54caa13c31d31b9fdc459c38b49` (SmolLM2-360M). Every original model/config/tokenizer-file hash is checked against the existing input lock. All three synthetic passages and their exact stored token IDs are reused without truncation, new text selection, chat templates or padding. Cuts are the quarter, half, last-block entrance, and final-normalization entrance; all five scales are fixed before execution. No model parameter is edited or trained.

Continuation and state measurements. A clean forward captures every block-entry state and the actual mask, position, and rotary-embedding arguments. We call the original layers from the selected cut with the same arguments and input cH_o, not a separately calibrated approximation. Attention and FFN outputs and intervening residual states are captured. The path budget sums norms of *realized* residual displacements; raw branch-output norms and floating-point addition discrepancies are saved separately. Per-sublayer radial/tangential summaries and per-token initial/final radii, net displacements, signed radial components, tangential norms and direction cosines are exported. Parameter version counters remain unchanged, and hooks are removed on success and exceptions.

Readout and numerical meaning. The suffix environment is Python 3.12, PyTorch 2.8.0+cpu, Transformers 4.56.2, NumPy 2.2.6, huggingface-hub 0.36.2, safetensors 0.8.0 and tokenizers 0.22.2, with two PyTorch CPU threads, eager attention, evaluation mode and no cache or generation. The primary readout evaluates the stated mathematical RMSNorm and the unchanged head in float64 on the float32 endpoints. This avoids subtractive softmax error obscuring small KL values; it is not silently equated with native float32 readout. Logits at positions $0, \dots, N-2$ predict targets $1, \dots, N-1$. Every summary uses precisely these target positions, with equal token weights within a model; no cross-tokenizer NLL average is formed.

Certificates and domains. We project the two endpoints independently and compare their logit difference with Eq. (43). All per-token values of ξ, logit oscillation, target-NLL change and both KL directions are retained. For a measured path budget B_i, the finite-radius certificate is evaluated only where $r_i > B_i$; a separate net-displacement budget is also checked. Such budgets certify the executed path and are not estimates of a scale-uniform global worst-case increment bound. Invalid-radius cases are recorded, not discarded from the experiment. The endpoint certificate does not require that radius condition. Inequality comparisons allow 10^{-9} absolute rounding tolerance, the endpoint identity 10^{-8}, and KL values no smaller than -10^{-10}. Raw values, including any tiny negative roundoff, are preserved. Tail fits use exactly scales $64, 512, 4096$ and all nonempty passage/cut curves; the theorem does not require them to saturate its upper rate.

Checks and inventory. Before either pretrained run, seven tests cover 30 random endpoint examples, a large positive radial displacement at zero epsilon, vocabulary-common shifts, invalid-radius cases, identity readout, every-cut continuation in tiny Qwen3 and Llama models, target indexing, and exception cleanup. Both model jobs pass all seven. For the pretrained models, every $c = 1$ continuation reproduces the clean final residual within 10^{-5}, and every no-suffix control is exactly identical. Six clean full-model forwards, 90 nonempty suffix evaluations and 30 empty-suffix readout controls are separate from the earlier 2058 full forwards. Thus the cumulative complete-forward count is 2064; suffix-only evaluations are not counted as complete model forwards. No training or new natural-text evaluation is added here. Full JSON/NPZ records retain execution times, source/protocol/library/model hashes and the complete inventory. The maximum native-clean versus float64-readout logit discrepancy is 1.30×10^{-5} for Qwen and 2.15×10^{-5} for SmolLM2.

H Balanced-reference intervention protocol and complete strength curves

Fixed operators and input inventory. The new panel uses the same two immutable checkpoints, three synthetic passages and three 1024-token natural-text excerpts as the earlier public-weight studies. The original calibration token IDs, model/tokenizer files and their hashes are unchanged. Each model/text pair is a fixed scientific shard, with all strengths and all controls, rather than a text chosen after measuring the edit. Nine of the twelve completed in one execution; three natural-text Qwen shards required the missing conditions to be recovered in smaller execution partitions after a time limit. The first shard was SmolLM2's library passage. All use identical measurement functions and locked inputs. No model is trained, screened or selected by its response.

For each layer, compare T_M from Proposition 1, $T_F = \alpha U^\top$ and $T_B = \beta U^\top$. All paths are $D(t) = D + t(T - D)$ at

$$t \in \{0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 1\}.$$

Targets use original weights and original calibration features only and remain fixed across strengths within each execution shard. Calibration is computed on the same locked calibration sequence in each original shard, not fitted to the evaluation text or loss. Recovery reuses the exact parent scalar coefficients and Gaussian coefficients and verifies every original and rescaled target-matrix hash, rather than recalibrating missing conditions. Cross-shard target hashes and coefficient spreads are reported, rather than assuming cross-machine byte identity. Gate, attention, normalization, embeddings and output head remain unchanged. A zero up row has an identically zero gated feature, so its balanced target column is set to zero and counted. Column-norm preservation is checked on active channels; no division by zero or outcome-dependent clipping occurs.

Table 5: Complete nonempty same-scale curves for Qwen. All entries are target-weighted means; $\Delta\mathcal{L}$ is suffix minus direct NLL, and coverage is the fraction satisfying the measured path-budget radius condition. All five final-normalization controls per passage have exactly zero differences and coverage one.

Cut	c	$\|\Delta z\|_\infty$	$\mathrm{KL}(p\|q)$	$\mathrm{KL}(q\|p)$	$\Delta\mathcal{L}$	Coverage
7	1	23.006	1.08×10^{1}	8.26×10^{0}	-7.620	0.008
7	8	20.685	7.77×10^{0}	6.44×10^{0}	-7.066	0.008
7	64	9.323	2.57×10^{0}	2.42×10^{0}	-3.416	0.008
7	512	3.615	3.48×10^{-1}	3.63×10^{-1}	-0.054	0.971
7	4096	0.575	8.91×10^{-3}	8.90×10^{-3}	0.019	1.000
14	1	22.402	9.56×10^{0}	6.39×10^{0}	-6.001	0.008
14	8	16.644	6.16×10^{0}	4.20×10^{0}	-4.957	0.008
14	64	6.805	1.07×10^{0}	1.03×10^{0}	-1.049	0.639
14	512	2.638	1.20×10^{-1}	1.18×10^{-1}	0.118	1.000
14	4096	0.380	2.67×10^{-3}	2.65×10^{-3}	0.021	1.000
27	1	10.006	2.13×10^{-1}	2.48×10^{-1}	-0.381	1.000
27	8	1.018	2.54×10^{-3}	2.50×10^{-3}	-0.048	1.000
27	64	0.123	3.94×10^{-5}	3.93×10^{-5}	-5.97×10^{-3}	1.000
27	512	0.015	6.15×10^{-7}	6.15×10^{-7}	-7.46×10^{-4}	1.000
27	4096	1.92×10^{-3}	9.61×10^{-9}	9.61×10^{-9}	-9.33×10^{-5}	1.000

Table 6: Complete nonempty same-scale curves for SmolLM2. All entries are target-weighted means; $\Delta\mathcal{L}$ is suffix minus direct NLL, and coverage is the fraction satisfying the measured path-budget radius condition. All five final-normalization controls per passage have exactly zero differences and coverage one.

Cut	c	$\|\Delta z\|_\infty$	$\mathrm{KL}(p\|q)$	$\mathrm{KL}(q\|p)$	$\Delta\mathcal{L}$	Coverage
8	1	26.619	1.62×10^{1}	1.30×10^{1}	-13.081	0.000
8	8	19.970	1.06×10^{1}	1.02×10^{1}	-10.560	0.010
8	64	4.767	5.31×10^{-1}	6.73×10^{-1}	-4.334	1.000
8	512	0.645	7.69×10^{-3}	7.92×10^{-3}	-0.623	1.000
8	4096	0.081	1.19×10^{-4}	1.20×10^{-4}	-0.079	1.000
16	1	25.457	1.45×10^{1}	1.07×10^{1}	-10.673	0.000
16	8	16.059	8.70×10^{0}	7.82×10^{0}	-8.418	0.008
16	64	3.729	3.79×10^{-1}	4.50×10^{-1}	-3.361	1.000
16	512	0.485	5.55×10^{-3}	5.68×10^{-3}	-0.463	1.000
16	4096	0.061	8.57×10^{-5}	8.60×10^{-5}	-0.058	1.000
31	1	17.194	4.93×10^{-1}	5.94×10^{-1}	-0.775	0.721
31	8	2.181	1.21×10^{-2}	1.09×10^{-2}	-0.090	1.000
31	64	0.265	2.08×10^{-4}	2.05×10^{-4}	-0.011	1.000
31	512	0.033	3.29×10^{-6}	3.29×10^{-6}	-1.42×10^{-3}	1.000
31	4096	4.12×10^{-3}	5.16×10^{-8}	5.15×10^{-8}	-1.77×10^{-4}	1.000

Separately matched ordinary edits. For every reference T there are six Gaussian controls: seeds 20260916, 20260917, 20260918, separately calibrated to the weight-displacement norm and the original-input branch-displacement RMS. A layer's generator uses the seed plus 1009ℓ; its Gaussian matrix is shared across reference families and matching rules. Coefficients are frozen across the strength grid:

$$D_{R,F}(t) = D + t\frac{\|T - D\|_F}{\|R\|_F}R, \qquad D_{R,B}(t) = D + t\frac{\mathrm{RMS}((T - D)z)}{\mathrm{RMS}(Rz)}R.$$

The two rules match different quantities and are not asserted to match both simultaneously. Realized native-arithmetic errors are recorded at each nonzero strength. Equal t between T_M, T_F, T_B is not by itself an equal-magnitude comparison; each reference has its own matched controls.

A coordinate-consistency experiment. Independently of loss, repeat each reference path after the fixed diagonal rescaling with entries $(1/2, 1, 2, 4)$ in channel order. Recompute the target from CU, DC^{-1} and Cz; the gate remains fixed. Compare every block state and the final output against the corresponding unscaled-reference path, not against the original unedited model. Positive powers of two reduce input-rounding confounding. General positive diagonal scales are covered by the algebraic proof and float64 synthetic tests, not by claiming that one native-arithmetic experiment covers every possible floating-point condition. Scalar references need not be invariant, so their rescaling differences are retained as observations, not classified as software failures.

Measurement arithmetic and storage. Stored weights are evaluated with native float32, eager attention, dropout disabled, no cache, and a recorded CPU-thread count capped at four in the pinned implementation. The actual thread count is retained per shard, rather than inferred from the configured cap. The final normalization and head are also native float32; the head is evaluated in chunks of eight prediction targets. NLL, both KL directions and their reductions use float64. This arithmetic policy differs from the mathematical float64 readout in the suffix-coefficient experiment and is labeled separately.

Hooks capture every residual-block entry and the actual last-block output before final normalization. Saved profiles contain per-depth population standard deviation and radius, radius-ratio and direction-chord summaries, not every hidden-state vector. Per-target NLL, clean NLL, both KL directions and logit L_2/L_∞ changes are saved. Offline reconstruction checks these scalars and summaries; full hidden states and vocabulary logits can be regenerated with the supplied runner and fixed model files. NLL and KL aggregate by prediction-target count within one model and text panel. Control ranges span three perturbation seeds and do not represent independent model replications.

Complete inventory and recovery checks. Three references with six controls each give $21 \times 8 \times 12 = 2016$ declared primary cells, plus $3 \times 8 \times 12 = 288$ coordinate checks. Zero-strength rows reuse measured baselines. The uninterrupted design has one calibration, one clean reference, 147 positive-strength primary evaluations, one rescaled clean reference, 21 positive-strength coordinate checks, and one final clean recheck: 172 per shard, or 2064 overall. Repeated recovery baselines and checks increase the recorded count to at least 2160; unsaved work at the three interruptions may add calls beyond that lower bound. Scientific condition counts remain 2016 and 288, not inflated by these checks. Head-only chunks are not counted as backbone forwards.

Every shard runs the ten algebra/restoration tests and the three native-architecture tests before loading pretrained measurements. Checkpoint files, input IDs, measurement sources and dependency versions are checked against the executed manifest. Successful complete executions and recovery partitions hash all checkpoint parameters before and after, and their final clean evaluations exactly recover their own initial states and readouts. Edit setup and exceptions restore modified matrices. The three interrupted original sessions did not reach their final clean check; their previously saved cells are retained, not assigned a nonexistent final check. The original-clean loss must also agree with the preceding archived value within 2×10^{-5} nats. Fixed tolerances are 5×10^{-4} for the specified local matching rule, 2×10^{-5} for original-function coordinate checks, 2×10^{-4} for balanced-path state checks and -10^{-10} as the lower rounding tolerance for tokenwise KL. These numerical tolerances do not require a favorable scientific result.

Timeout recovery and session baselines. The original Qwen natural-text jobs reached a 90-minute execution limit after saving part of the fixed grid. Every saved condition, including its

original numerical values, is retained. Six predetermined arm groups per text partition the missing positive-strength keys without overlap; zero-strength keys reuse the original measured clean baseline. Each partition verifies the frozen parent targets and uses its own freshly checked clean baseline for NLL damage, KL and geometric differences. No outcome is shifted to force cross-session equality. The maximum clean-NLL difference is 2.781×10^{-7} nats. A repeated $t = 0.001$ balanced-path sentinel checks agreement with a saved condition but is excluded from the scientific panel. All raw parent and recovery observations accompany the complete record.

Three recovery preflights stopped before new scientific cells because their aggregate vocabulary-L_2 sentinel tolerance was not normalized for vocabulary size. The correction applies the same absolute tolerance in per-coordinate RMS units for that one metric; NLL, KL, L_∞, frozen matrix hashes, local matching and all measurement formulas remain unchanged. Original failures, corrected thresholds, actual differences and their additional evaluation counts are retained. This correction concerns cross-machine numerical validation, not a criterion requiring smaller prediction damage. The fixed grid is completed without discarding unfavorable results or treating execution partitions as independent model replications.

Display and completeness. Table 2 displays $t = 0.05$ from the full grid. This compact display choice was made after the first shard, before the remaining panel was analyzed; it does not change the fixed strength inventory. Figure 2 shows every positive strength on the natural texts. The following tables and figures retain both text panels, all references and both matching rules; no strength is selected as a universal functional-safety threshold.

Table 7: Complete balanced/scalar-reference strength panel. Every row is a target-weighted model/text aggregate. Matched random-control loss ranges use all three seeds and are calibrated separately to the reference in each row. Zero-strength rows share the clean baseline. Two-line control entries show the lower endpoint above the upper endpoint.

Model/text	Reference	t	ΔL	KL	Random ΔL Weight match	Random ΔL Branch match
Qwen/Natural	Balanced	0	0	0	0–0	0–0
Qwen/Natural	Balanced	0.001	0.0010	3.956×10^{-5}	-3.487×10^{-4} 7.052×10^{-5}	-3.337×10^{-4} 1.262×10^{-4}
Qwen/Natural	Balanced	0.01	0.0142	0.0038	-0.0023–0.0021	-5.754×10^{-4} 0.0044
Qwen/Natural	Balanced	0.05	0.1650	0.1211	0.0204–0.0413	0.0680–0.0962
Qwen/Natural	Balanced	0.1	0.8515	0.7680	0.1204–0.1732	0.3625–0.4913
Qwen/Natural	Balanced	0.25	13.9224	13.8644	1.0773–1.5353	2.7064–3.0383
Qwen/Natural	Balanced	0.5	22.7509	23.5094	4.8586–5.8368	6.8949–7.8524
Qwen/Natural	Balanced	1	49.0459	51.0319	11.1856–12.4674	11.3179–12.8516
Qwen/Natural	Frobenius	0	0	0	0–0	0–0
Qwen/Natural	Frobenius	0.001	0.0015	2.288×10^{-4}	-3.481×10^{-4} 6.963×10^{-5}	-4.221×10^{-4} 5.663×10^{-4}
Qwen/Natural	Frobenius	0.01	0.0363	0.0247	-0.0023–0.0021	0.0145–0.0240
Qwen/Natural	Frobenius	0.05	1.3013	1.2548	0.0205–0.0413	0.6833–0.8410
Qwen/Natural	Frobenius	0.1	6.0747	5.9930	0.1206–0.1735	2.5096–2.9089
Qwen/Natural	Frobenius	0.25	15.4632	15.4029	1.0799–1.5392	4.3172–5.8471
Qwen/Natural	Frobenius	0.5	19.3322	19.9853	4.8689–5.8519	9.7317–11.2120
Qwen/Natural	Frobenius	1	33.1587	34.4548	11.1815–12.4932	11.6764–12.4199
Qwen/Natural	Branch RMS	0	0	0	0–0	0–0
Qwen/Natural	Branch RMS	0.001	0.0012	4.536×10^{-5}	-3.100×10^{-4} 3.800×10^{-5}	-3.456×10^{-4} 1.625×10^{-4}
Qwen/Natural	Branch RMS	0.01	0.0161	0.0045	-0.0023–0.0014	-4.500×10^{-4} 0.0050
Qwen/Natural	Branch RMS	0.05	0.1949	0.1439	0.0116–0.0287	0.0753–0.1068
Qwen/Natural	Branch RMS	0.1	1.0468	0.9467	0.0803–0.1186	0.4054–0.5451
Qwen/Natural	Branch RMS	0.25	13.9918	13.9046	0.6952–1.0236	2.7949–3.1680
Qwen/Natural	Branch RMS	0.5	24.9558	25.7187	3.6665–4.4742	6.6465–7.6439
Qwen/Natural	Branch RMS	1	33.4626	34.7825	10.7149–11.9412	11.1502–13.0800
Qwen/Short	Balanced	0	0	0	0–0	0–0

Model/text	Reference	t	ΔL	KL	Random ΔL Weight match	Random ΔL Branch match
Qwen/Short	Balanced	0.001	0.0035	5.801×10^{-5}	-4.475×10^{-4} -2.948×10^{-6}	-3.677×10^{-4} 2.608×10^{-6}
Qwen/Short	Balanced	0.01	0.0388	0.0060	-0.0040–0.0011	-0.0029 9.855×10^{-4}
Qwen/Short	Balanced	0.05	0.2744	0.1795	-0.0065–0.0369	0.0227–0.0882
Qwen/Short	Balanced	0.1	1.0016	0.9587	0.0516–0.2020	0.2851–0.5243
Qwen/Short	Balanced	0.25	11.2215	11.3813	1.2834–1.7586	3.6727–4.0698
Qwen/Short	Balanced	0.5	17.6216	19.3262	5.9614–6.3936	6.6137–8.2710
Qwen/Short	Balanced	1	47.3731	51.9701	10.1125–11.1689	10.3752–11.3550
Qwen/Short	Frobenius	0	0	0	0–0	0–0
Qwen/Short	Frobenius	0.001	0.0030	4.498×10^{-4}	-4.449×10^{-4} -3.129×10^{-6}	-7.592×10^{-4} 5.868×10^{-4}
Qwen/Short	Frobenius	0.01	0.0544	0.0497	-0.0040–0.0010	0.0018–0.0104
Qwen/Short	Frobenius	0.05	1.6060	1.6950	-0.0065–0.0369	0.7128–1.0406
Qwen/Short	Frobenius	0.1	6.6271	6.7832	0.0519–0.2023	3.3517–3.8514
Qwen/Short	Frobenius	0.25	9.9277	10.7586	1.2886–1.7637	5.1462–6.5378
Qwen/Short	Frobenius	0.5	30.5375	33.6558	5.9683–6.4115	8.7862–10.5390
Qwen/Short	Frobenius	1	31.2814	34.4544	10.1037–11.1953	10.7677–12.0019
Qwen/Short	Branch RMS	0	0	0	0–0	0–0
Qwen/Short	Branch RMS	0.001	0.0033	6.808×10^{-5}	-3.399×10^{-4} 3.317×10^{-5}	-4.050×10^{-4} 5.378×10^{-5}
Qwen/Short	Branch RMS	0.01	0.0375	0.0073	-0.0031–0.0011	-0.0031–0.0017
Qwen/Short	Branch RMS	0.05	0.2790	0.2135	-0.0069–0.0259	0.0264–0.0954
Qwen/Short	Branch RMS	0.1	1.1415	1.1298	0.0286–0.1338	0.3411–0.5993
Qwen/Short	Branch RMS	0.25	13.0368	12.9954	0.6458–1.0769	3.6629–4.1403
Qwen/Short	Branch RMS	0.5	17.8961	19.8184	4.7513–4.9901	6.5814–8.0995
Qwen/Short	Branch RMS	1	31.5626	34.7697	9.4901–10.6564	10.1942–11.8631
SmolLM2/Natural	Balanced	0	0	0	0–0	0–0
SmolLM2/Natural	Balanced	0.001	1.805×10^{-4}	1.041×10^{-4}	-1.036×10^{-4} 8.106×10^{-5} 6.759×10^{-4}	-6.347×10^{-5} 7.755×10^{-5}
SmolLM2/Natural	Balanced	0.01	0.0108	0.0101	0.0025	0.0028–0.0046
SmolLM2/Natural	Balanced	0.05	0.2536	0.2494	0.0365–0.0504	0.0776–0.0949
SmolLM2/Natural	Balanced	0.1	1.1645	1.1514	0.1560–0.2010	0.3392–0.4023
SmolLM2/Natural	Balanced	0.25	10.3469	10.2584	1.3403–1.6536	4.4116–4.5841
SmolLM2/Natural	Balanced	0.5	27.7243	28.0303	8.0261–8.4983	9.8024–10.1901
SmolLM2/Natural	Balanced	1	41.1800	41.4010	12.0743–12.3611	12.2932–12.6544
SmolLM2/Natural	Frobenius	0	0	0	0–0	0–0
SmolLM2/Natural	Frobenius	0.001	1.641×10^{-4}	3.614×10^{-4}	-1.039×10^{-4} 8.123×10^{-5} 6.748×10^{-4}	2.148×10^{-4} 2.373×10^{-4}
SmolLM2/Natural	Frobenius	0.01	0.0260	0.0276	0.0025	0.0191–0.0211
SmolLM2/Natural	Frobenius	0.05	1.3224	1.3316	0.0366–0.0505	0.5193–0.5557
SmolLM2/Natural	Frobenius	0.1	7.4728	7.3627	0.1561–0.2012	3.5828–4.0602
SmolLM2/Natural	Frobenius	0.25	20.0021	20.1291	1.3416–1.6551	8.5351–9.3995
SmolLM2/Natural	Frobenius	0.5	26.5225	26.6446	8.0253–8.4953	10.0965–10.9510
SmolLM2/Natural	Frobenius	1	39.1745	39.4373	12.0719–12.3598	12.2541–13.4221
SmolLM2/Natural	Branch RMS	0	0	0	0–0	0–0
SmolLM2/Natural	Branch RMS	0.001	2.568×10^{-4}	1.211×10^{-4}	-1.037×10^{-4} 6.871×10^{-5} 3.778×10^{-4}	-6.783×10^{-5} 8.540×10^{-5}
SmolLM2/Natural	Branch RMS	0.01	0.0130	0.0117	0.0021	0.0029–0.0048
SmolLM2/Natural	Branch RMS	0.05	0.2944	0.2866	0.0302–0.0431	0.0809–0.0998
SmolLM2/Natural	Branch RMS	0.1	1.3342	1.3136	0.1306–0.1681	0.3551–0.4232
SmolLM2/Natural	Branch RMS	0.25	10.5917	10.4852	1.0386–1.2632	4.6147–4.7855
SmolLM2/Natural	Branch RMS	0.5	33.7123	34.0143	7.2795–7.5009	9.8541–10.5460
SmolLM2/Natural	Branch RMS	1	42.4830	42.7686	12.0851–12.2617	12.2136–12.6660
SmolLM2/Short	Balanced	0	0	0	0–0	0–0
SmolLM2/Short	Balanced	0.001	0.0044	2.030×10^{-4}	-2.423×10^{-4} 7.251×10^{-4}	-2.798×10^{-4} 8.949×10^{-4}

Table 8: Same-scale suffix influence. D_c is the target-position mean of $\|\Delta z\|_\infty$; $K_c = \mathbb{E}_i \mathrm{KL}(p_i \| q_i)$, with direct p_i and suffix q_i. Each entry includes all three passages (382 Qwen or 383 SmolLM2 targets). Both readouts receive the same cH_o; reverse KL and every scale appear in Appendix G.

Model	Cut	D_{64}	D_{4096}	K_{64}	K_{4096}
Qwen	7	9.323	0.575	2.57×10^{0}	8.91×10^{-3}
Qwen	14	6.805	0.380	1.07×10^{0}	2.67×10^{-3}
Qwen	27	0.123	1.92×10^{-3}	3.94×10^{-5}	9.61×10^{-9}
SmolLM2	8	4.767	0.081	5.31×10^{-1}	1.19×10^{-4}
SmolLM2	16	3.729	0.061	3.79×10^{-1}	8.57×10^{-5}
SmolLM2	31	0.265	4.12×10^{-3}	2.08×10^{-4}	5.16×10^{-8}

Model/text	Reference	t	ΔL	KL	Random ΔL Weight match	Random ΔL Branch match
SmolLM2/Short	Balanced	0.01	0.0586	0.0169	0.0024–0.0133	0.0087–0.0215
SmolLM2/Short	Balanced	0.05	0.4726	0.2777	0.0763–0.1167	0.1391–0.1717
SmolLM2/Short	Balanced	0.1	1.5092	1.1313	0.2211–0.2953	0.4264–0.4825
SmolLM2/Short	Balanced	0.25	8.8290	8.2133	1.3843–1.6998	4.3363–4.6881
SmolLM2/Short	Balanced	0.5	28.9848	29.7654	6.9805–7.2097	8.3101–9.4979
SmolLM2/Short	Balanced	1	37.6173	38.7105	10.9166–11.6476	11.0525–12.1819
SmolLM2/Short	Frobenius	0	0	0	0–0	0–0
SmolLM2/Short	Frobenius	0.001	0.0102	0.0015	-2.431×10^{-4} 7.263×10^{-4}	-8.604×10^{-5} 0.0017
SmolLM2/Short	Frobenius	0.01	0.1269	0.0527	0.0024–0.0133	0.0545–0.0610
SmolLM2/Short	Frobenius	0.05	1.9394	1.5697	0.0764–0.1168	0.6017–0.6971
SmolLM2/Short	Frobenius	0.1	6.7102	6.3698	0.2213–0.2957	3.0739–4.2813
SmolLM2/Short	Frobenius	0.25	18.7583	19.1770	1.3856–1.7017	7.1250–7.8586
SmolLM2/Short	Frobenius	0.5	28.0109	28.9967	6.9791–7.2090	8.7762–9.8083
SmolLM2/Short	Frobenius	1	39.6699	40.8151	10.9102–11.6490	11.1928–12.4227
SmolLM2/Short	Branch RMS	0	0	0	0–0	0–0
SmolLM2/Short	Branch RMS	0.001	0.0048	2.445×10^{-4}	-2.052×10^{-4} 7.189×10^{-4}	-2.848×10^{-4} 9.619×10^{-4}
SmolLM2/Short	Branch RMS	0.01	0.0659	0.0202	0.0017–0.0120	0.0089–0.0225
SmolLM2/Short	Branch RMS	0.05	0.5313	0.3188	0.0615–0.1057	0.1415–0.1786
SmolLM2/Short	Branch RMS	0.1	1.6993	1.2932	0.1848–0.2646	0.4388–0.5052
SmolLM2/Short	Branch RMS	0.25	8.6510	8.1455	1.0620–1.3822	4.4962–4.8804
SmolLM2/Short	Branch RMS	0.5	30.2154	30.8511	6.4650–7.0213	8.4547–9.8308
SmolLM2/Short	Branch RMS	1	40.4291	41.8876	11.0538–11.5267	11.0831–12.3622

I Additional native suffix results and controls

I.1 Complete same-scale contrasts

We now test the suffix prediction at fixed depth by changing only the input radius. At each selected clean cut state, both the direct head and the original suffix receive exactly cH_o. The fixed panel uses the three existing synthetic passages, cuts $\{7, 14, 27, 28\}$ for Qwen and $\{8, 16, 31, 32\}$ for SmolLM2, and all scales $c \in \{1, 8, 64, 512, 4096\}$. The last cut has no remaining block. Six clean full-model forwards supply the states; 90 nonempty suffix evaluations and 30 no-suffix controls form the 120 paired conditions. Suffixes execute in float32; both mathematical final readouts are evaluated in float64 on these endpoints (Appendix G).

Radius changes the observable contribution at the same depth. Table 8 and Figure 6 quantify distributional attenuation, rather than inferring it from similar average NLL. All 18 passage/cut logit curves and all 36 directed-KL curves decrease across the complete grid. On the fixed tail $c = 64, 512, 4096$, Qwen's terminal block and all three SmolLM2 cuts give log–log decay exponents 0.976–1.001 for mean logit difference and 1.993–2.088 across both KL directions, close to inverse-radius and inverse-square behavior. Earlier Qwen cuts instead give 0.626–0.705 and 1.302–1.491 on

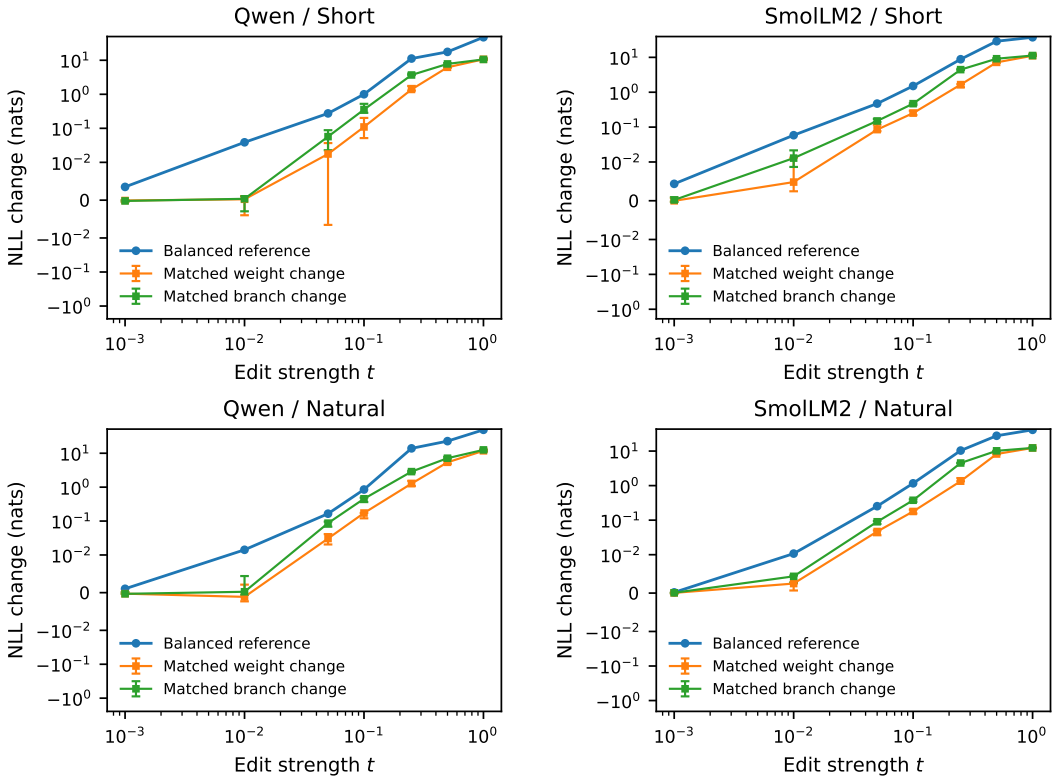


Figure 4: All positive strengths of the balanced reference and its two separately matched random families. NLL differences are target-weighted within each model/text panel. Each random line is the median of three seed-level means and its bars span the three means, not confidence intervals. The symmetric-log vertical scale retains signed losses near zero. Zero strength is exactly the common clean baseline and is listed in the tables.

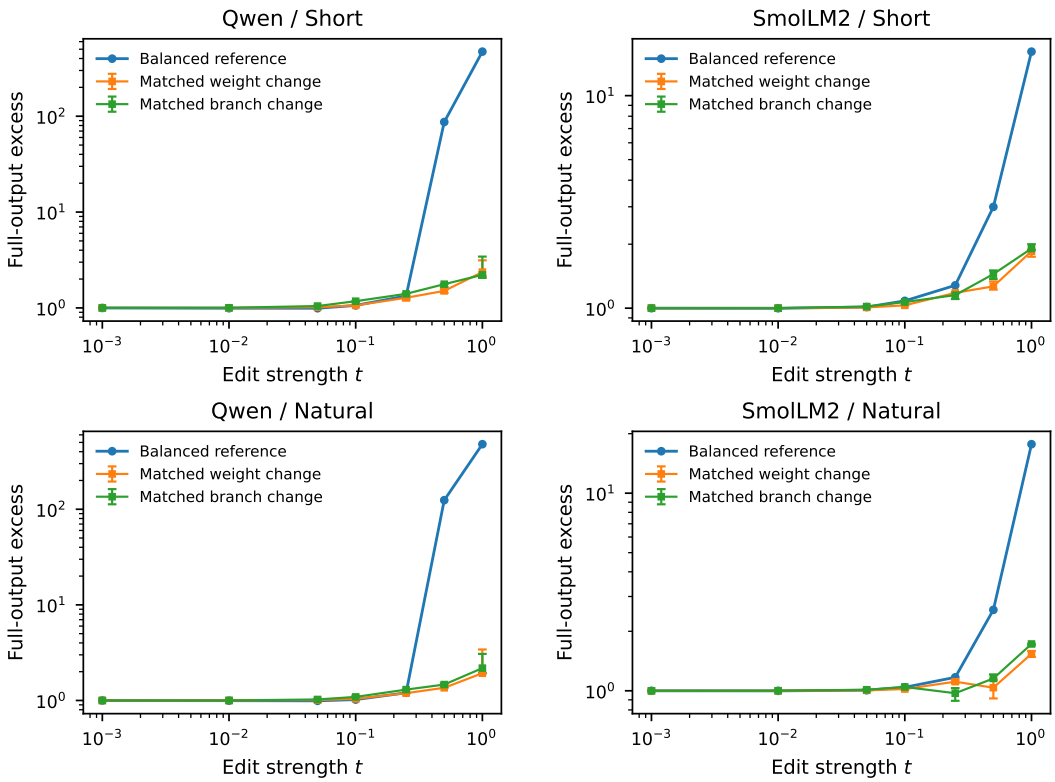


Figure 5: Complete-output amplification along the same balanced-reference paths. For display, the reference is the geometric mean of its three text-level excess values; each random family shows the median and full range of three seed-level text geometric means. All individual text outcomes and their extrema remain in the raw records and the generated result tables. These display summaries do not turn texts or perturbation seeds into independent model samples.

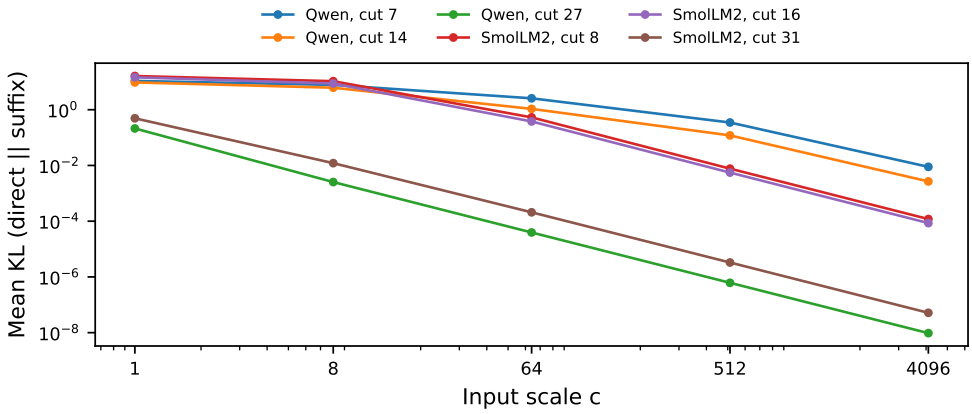


Figure 6: Output-distribution change at every fixed scale and nonempty cut. Each curve is a within-model target-weighted mean over all three passages; lines join measured conditions. Zero no-suffix controls are omitted from the logarithmic axis. This measures suffix influence, not a comparison of model quality across tokenizers.

this finite range. Their measured update budgets and directions still evolve with scale; these curves show attenuation but do not yet exhibit the same tail rate. The theorem is asymptotic and does not assign one finite-window exponent to every cut. Both earlier and terminal nonempty suffixes attenuate: the effect is not restricted to a late-layer cut. At $c = 1$, every continuation reproduces its clean final state; all 30 no-suffix conditions have zero paired logit, NLL, and KL difference. Distributional attenuation also differs from functional improvement: at Qwen cut 14, suffix-minus-direct NLL changes from -1.049 at $c = 64$ to $+0.118$ at $c = 512$, even as forward KL falls from 1.074 to 0.120.

Direction-resolved certificates quantify the output change. The exact radial–tangential endpoint identity has maximum numerical discrepancy 2.03×10^{-13} across 11,475 nonempty target-position conditions. Every endpoint NLL and bidirectional-KL bound passes. The measured path-budget radius condition holds for 0.8–100.0% of targets at $c = 64$ across cuts and for all targets at $c = 4096$; invalid-radius cases are retained. At $c = 4096$, the six cut-level median ratios $\|\Delta z\|_\infty / \eta_{\text{path}}$ are 0.030–0.088: the norm-only upper bound is conservative. The direction-resolved endpoint certificate is much closer to the logit oscillation, with median $\operatorname{osc}(\Delta z)/\xi$ of 0.98990–0.99996. Its subsequent KL bound remains looser: median $\max\{\mathrm{KL}(p\|q), \mathrm{KL}(q\|p)\}/(\xi^2/8)$ is 0.010–0.060. Thus measuring direction removes substantial geometric slack, while the general distributional inequality still leaves slack. These are pathwise certificates, not endpoint-free predictions.

J Additional native edit and state results

J.1 Scale controls and single-layer response

Severe damage survives both scale controls. For Qwen, the calibration branch RMS matches to relative error below 10^{-7}. On the new texts, frozen-input branch ratios range from 0.90 to 1.91, so this is calibration transfer rather than exact per-text matching. Unit tying produces $4{,}987$–$5{,}213$-fold full-output excess; Frobenius matching leaves $4{,}185$–$4{,}363$, and branch matching leaves $2{,}459$–$2{,}580$. Token-weighted loss rises from 4.211 to 35.382, 35.492 and 35.773 nats, respectively (Table 9). Thus neither tested scale control restores propagation or function. Matching a branch on original inputs does not match its action along the edited network's changed trajectory.

Layerwise localization and functional damage. Every-layer interventions allow direct comparison of the original g_ℓ with the response of editing that layer. Across 28 layers, its signed correlation with loss damage is 0.271–0.373 for unit edits and 0.261–0.340 for Frobenius-matched edits. A

Table 9: All-layer edits on the same three texts, evaluated separately with each tokenizer. Original losses are 4.211 over 382 targets (Qwen) and 3.513 over 383 targets (SmolLM2). Ranges span passages; $\mathcal{E}_L$ uses the complete output of each model. Absolute NLL is not compared across tokenizers.

Model	Edit	$\Delta\mathcal{L}$ (nats)	$\mathcal{E}_L$ range
Qwen3-0.6B	Unit transpose	31.171	4986.51–5212.67
Qwen3-0.6B	Frobenius matched	31.281	4184.66–4362.94
Qwen3-0.6B	Branch matched	31.563	2459.01–2579.78
SmolLM2-360M	Unit transpose	40.149	19.36–22.71
SmolLM2-360M	Frobenius matched	39.670	17.81–20.77
SmolLM2-360M	Branch matched	40.429	16.81–19.80

predeclared original-input branch-RMS ratio reaches 0.633–0.709 and 0.683–0.729, respectively. This advantage is intervention-dependent: after branch matching its correlation is 0.104–0.305, versus 0.343–0.353 for g_ℓ. All six fixed predictors, including edit-distance baselines, appear in Appendix T. These are descriptive comparisons within one model, not independent-layer significance tests.

Editing layer 27 alone leaves its entry response $\mathcal{E}_{27} = 1$ exactly, yet the unit edit increases loss by 19.54–22.01 nats. The complete-output excess is only 2.44–2.60 in these cases. On the repair passage, editing layer 26 instead gives much larger excess 19.94 but smaller loss damage 0.81 nats. This demonstrates why response coverage and language-model loss must both be measured: a quiet entry endpoint can coexist with severe damage, and amplification magnitude alone does not order that damage. Spectral-gap, multi-initialization and float64 SVD checks support the numerical directions used here; their scope is this public model, not the archived checkpoints.

J.2 Residual-state controls and fixed-head readout

At a fixed cut k, edit only blocks before k and retain the original suffix. For each token, compare the edited state h_e with $(\|h_o\|/\|h_e\|)h_e$ and $(\|h_e\|/\|h_o\|)h_o$, where h_o is its clean state. These oracle interventions restore original length or direction, respectively. Across all 45 pre-final passage/arm/cut comparisons in each model, direction restoration gives lower NLL than length restoration. The corresponding contrast also holds for the ordinary edits (Section 6), identifying a shared residual-state phenomenon. Full results and patch checks appear in Appendices U and W.

At the Qwen last-block entrance, restoring length still produces 459–486-fold full-output excess and NLL 14.944–14.968. Restoring direction instead gives NLL about 4.591 (original 4.211), despite large residual scale. The equivalent SmolLM2 cut gives direction-restored NLL 4.202–4.208 versus 15.639–15.759 after length restoration (original 3.513). This recurrence concerns two recipes, not independent training seeds. Diagnostic ranking also varies: SmolLM2 single-layer loss correlations favor spectrum (0.474–0.777) over g_ℓ (0.119–0.315).

What does running the suffix add to fixed-head readout? On the synthetic probes, a separate control reads each clean cut state through only final RMSNorm and the original language-model head, omitting the suffix. For the three transpose arms, direction restoration followed by the original suffix has a substantially lower NLL at earlier cuts: Qwen cut 7 gives 5.118–5.214, versus 11.831 for direct readout; SmolLM2 cut 8 gives 5.671–6.073, versus 16.594. At late Qwen cut 27 the additional NLL advantage is only 0.000511–0.001024; SmolLM2 cut 31 retains 0.0807–0.0863 (Appendix V).

The depth-dependent NLL comparison motivates the suffix-influence theorem. Its baseline uses the unscaled clean state, whereas direction restoration uses a positively scaled clean state. Theorem 3 instead compares identical input scales and additionally controls distributional differences. The existing panel covers synthetic probes and three transpose arms; final-output cuts serve as readout checks. Section 8 completes the same-scale comparison with logits, NLL and both KL directions.

K REDUCED OBSERVABLES AND RETAINED ARCHIVED COMPARISONS

K.1 ARCHIVED SETTING AND EXTRACTION

The archived trajectory uses 24 pre-normalized SwiGLU blocks of hidden width 1536 (legacy "600M" label; full configuration in Appendix Y). Its 438-token probe excludes padding; the scanner captures post-norm FFN inputs and residual-block entries. The public-weight experiments use separate 28- and 32-layer models and new probes (Section 6).

The short and constrained runs have 11 checkpoints on 25K-step cosine horizons; the longer run has 41 checkpoints through step 99,999 with a 100K horizon. Their schedules differ. The longer run processes approximately 13.1B tokens by revisiting a 3.93B-token corpus, so later checkpoints do not supply independent training text.

For the archive's $L = 24$, H_{23} depends on FFN edits in layers 0–22, while $\partial H_{23}/\partial W_{\text{down}}^{(23)} = 0$. The original diagnostic nevertheless aggregates layers 0–23. We preserve that historical definition and separately evaluate endpoint-supported aggregates, including $g_{\max}^{<23} = \max_{0\le\ell<23} g_\ell$. Section K.3 applies this restriction to every layerwise baseline.

Diagnostic extraction. For the matrix X_ℓ of FFN-input vectors, r_ℓ is the leading eigenvector of $X_\ell^\top X_\ell$. The centered variant subtracts the token mean before forming this second moment. The scanner also records a residual-entry alignment, but the field used in g_ℓ is `align_ffn_raw`. Power iteration estimates the top singular pair of the up-projection. The code includes a full-SVD check for the first scanned layer at the first checkpoint; this check is not a certification of every singular pair in the dataset.

Archived comparisons use full-precision JSON; historical plots retain rounded summaries. New forwards restore immutable public weights. Appendix Y separates these sources and the unavailable original E4b checkpoints.

K.2 LEADING-MODE MOTIVATION

For a simplified positive constant gate $C_g = cI$ and a unit signal direction r, the Gram component satisfies

$$r^\top W_{\text{up}}^\top (cI) W_{\text{up}} r = c \sum_i \sigma_i(W_{\text{up}})^2 |v_i^\top r|^2. \tag{46}$$

We summarize the leading input-side singular direction $v_1^{(\ell)}$ and the leading uncentered FFN-input direction r_ℓ by

$$g_\ell = \sigma_1(W_{\text{up}}^{(\ell)})^2 \left| \left\langle v_1^{(\ell)}, r_\ell \right\rangle \right|^2, \qquad g_{\max} = \max_\ell g_\ell, \tag{47}$$

The top-singular contribution is cg_ℓ. This supplies a concrete motivation for multiplying spectral gain by alignment: neither a large singular value in an unoccupied direction nor alignment to a weak direction has the same interpretation as their combination. It does not make g_ℓ an exact gain for a gated, normalized Transformer. Gate statistics, the remaining singular subspace, attention, and the changed activation trajectory after surgery are omitted.

We consider two aggregates,

$$R = \sum_\ell \log(1 + g_\ell), \qquad g_{\max} = \max_\ell g_\ell. \tag{48}$$

The first tests an accumulated-resonance picture; the second tests a dominant-layer picture. Neither is fitted numerically to the response, but choosing between them after inspection is still model selection. In particular, e^R is not a certified bound on $\mathcal{E}$. The constrained-training observations below provide an empirical reason not to infer explosion from a large R alone.

K.3 SPECTRAL ALIGNMENT ACROSS TRAINING CONFIGURATIONS

The archived trajectory tests whether the leading-mode diagnostic provides information beyond simpler measurements. The earlier short-run comparison selected the maximum aggregate after observing failure of the sum: their rank correlations were 0.842 and 0.079, respectively. The later project

Table 10: Reviewer-triggered endpoint-support analysis. The response ends at block entry 23; the new aggregates use layers 0–22. Original all-layer results are retained. Every checkpoint is included in its original window.

	Full (41)		Late (31)	
Predictor	0–23	0–22	0–23	0–22
$g_{\max}$	0.250	0.211	0.785	0.788
$\max_\ell \sigma^2_{1,\ell}$	0.838	0.825	0.696	0.693
$\max_\ell a^2_\ell$	-0.184	-0.255	-0.063	-0.140
$\sum_\ell \log(1+g_\ell)$	-0.352	-0.369	-0.063	-0.154

ledger specified a late-window criterion before the long-run result; the 31 checkpoints at or after 25K give $\rho_{\mathrm{S}}(g_{\max}, \mathcal{E}) = 0.785$. This preserves an internally documented PASS, not an independently certified public preregistration. Across all 41 checkpoints, the same correlation falls to 0.250, versus 0.838 for spectrum alone. Late spectrum gives 0.696 and original amplification 0.743 (Table 15). The latter is also the denominator of $\mathcal{E}$, so its association is mathematically coupled to the response.

The association does not supply amplitude calibration or a local change law. Late Pearson correlation is 0.089; first-difference rank correlations after 25K, 30K and 50K are -0.169, 0.022 and 0.114. The diagnostic peaks at 25K and collapses at 27.5K while excess increases to its peak. Figure 8 and the complete comparisons remain in Appendix P.

Fixed trend and coverage controls. New comparisons use identical checkpoints, windows and response definitions, specified before the new calculations but after the original results were known. Partial rank correlations controlling linear and quadratic time are 0.625 and 0.407 late; adding spectrum and original amplification gives 0.366. These particular adjustments retain some association. However, deleting contiguous blocks of five late checkpoints can reverse the diagnostic's advantage over spectrum or original amplification. All block deletions and ordered models are retained in Appendix R; no independent-checkpoint significance is inferred.

Matching response support to FFNs 0–22 changes the gain correlations to 0.211150 overall and 0.787903 late, compared with 0.824913 and 0.693145 for spectrum (Table 10). The original layer-23 maxima at 25K and 30K cannot localize a cause of the entry-23 response. The revised aggregates have no inherited prospective PASS. Spectral gaps and numerical direction checks for those archived checkpoints remain unavailable; the new public-model checks do not certify them.

Forecasting against response persistence. A last-known-response predictor holds the final training-prefix $\log \mathcal{E}$ fixed across the next test block, reading neither the excluded gap label nor labels within that block. Its RMSE is 0.227383 and 0.126982 on the two fixed split schemes, lower than every one of the 11 original OLS models and their endpoint-supported counterparts. Their first test points are 42.5K and 57.5K; neither tests the early 20K–30K transition. Like OLS calibration, persistence uses historical surgical labels and addresses forecasting rather than label-free diagnosis. Appendix S retains all predictions and split boundaries. Thus the late rank association survives several checks, while broad predictive superiority does not.

K.4 Stable training under the tying constraint

The companion archived run trains stably with $W_{\mathrm{down}} = W_{\mathrm{up}}^\top$ from initialization and an independent gate. Its reported loss summary is 0.7243, versus 0.7045 for same-width untied and 0.7099 for a narrower parameter-matched model. These are means of logged training-step losses over the final 1,000 steps, not held-out validation losses. The available implementation truly shares the parameter across both projection uses, so differentiation includes both paths. Appendix Y records the architecture, optimizer, data provenance and unresolved deployed-runtime details.

Stable tying is compatible with a transient $g_{\max}$ peak of 10.80 and sum aggregate R between approximately 7 and 18 after initialization. Neither large diagnostic alone is sufficient for explosion. The different cosine horizons prevent a same-schedule causal training comparison (Appendix Q). This trajectory establishes a trainable tied configuration; it does not identify which adaptations pro-

duce stability. Historical randomized rescue, whose operator and seeds remain unresolved, is kept separately in Appendix M.

L ALGEBRA AND MEASUREMENT CONVENTIONS

Coordinate convention. The scanner stores an up-projection kernel as a $d \times m$ array acting on row vectors. Its power iteration uses $KK^\top$ to recover the input-side direction. Equations in the main text use column vectors and $W_{\text{up}} = K^\top$. These are the same linear map under transposition, not different singular-vector choices. For both conventions, the diagnostic is the squared product of the top singular value and its absolute alignment to the leading FFN-input direction.

The input direction matters. The code forms the raw and centered principal directions from the tensor named `ffn_input`. It separately forms a principal direction from `residual_input`, stored as `align_res_raw`. The published diagnostic field `g_raw` uses `align_ffn_raw`, not the residual alternative. Earlier prose that describes r_ℓ simply as a residual-stream principal component therefore needs this precision. No change to any source measurement is made here.

The vorticity identity. With $K = S + B$ and symmetric J_ρ, transposition gives $(KJ_\rho)^\top = J_\rho(S - B)$. Subtracting $(S + B)J_\rho$ gives $-[S, J_\rho] - \{B, J_\rho\}$. Both summands are skew-symmetric. Here $\delta = d\epsilon$ for the standard RMSNorm epsilon. For a learned normalization $N(h) = \Gamma\rho_\delta(h)$ with diagonal Γ, $DN = \Gamma J_\rho$ need not be symmetric. One must use $(KDN)^\top - KDN$ or explicitly absorb Γ into the base derivative before applying the identity.

A scalar counterexample to automatic positivity. Let $d = m = 1, W_{\text{up}} = W_{\text{down}} = 1, W_{\text{gate}} = -1$, and $x = 1$. The tied branch is $f(x) = x\phi(-x)$ and $f'(1) = \phi(-1) - \phi'(-1) < 0$. Transpose tying consequently does not guarantee a positive derivative even in one dimension. In more than one dimension, the independent gate creates an additional nonsymmetric term. This counterexample is algebraic clarification, not an empirical finding about a production checkpoint.

Norm containment. For bare RMSNorm, $\|\rho_\delta(h)\| < \sqrt{d}$. With fixed coordinate gains, the bound becomes $\|\Gamma\rho_\delta(h)\| \le \|\Gamma\|_{\text{op}}\sqrt{d}$. Applying such normalization to a whole residual output bounds that output. Applying it only to the branch increment bounds the increment instead, not the accumulated residual state. These two post-normalization placements must not share a name without a computational definition.

M HISTORICAL FROZEN-MODEL OBSERVATIONS

Table 11 retains the cross-model survey as reported in the earlier framework. The edit binds both the gate and up-projection paths as described in Section 2; the table is not a multi-model replication of the primary gate-preserving checkpoint diagnostic. Ratios in the source are rounded independently, so the displayed instability is not recomputed by dividing already rounded growth values.

Table 11: Historical whole-tie survey. Growth is the final/initial residual norm ratio; the last column is the source-reported excess relative to the original configuration.

Model	Original growth	Tied growth	Reported excess
Qwen3-0.6B	93	142955	1541
Mistral-7B-v0.3	155	58189	375
Qwen3-4B	124	27811	224
LLaMA-3.1-8B	89	6929	78
Gemma-3-1B	76	71	≈ 1

The Gemma exception is retained. The available observation does not identify the component responsible for it. In particular, a soft-cap applied to final output logits should not be presented as a demonstrated explanation of containment earlier in the residual stack.

The historical Qwen3 replacement series reports final norms 815 (original), $9{,}880{,}199$ (extensive tie), 7680 (reported antisymmetric rescue), and $23{,}269$ (reported symmetric control). These values describe one particular series, not the same sample average as Table 11. The implied reductions are approximately 1287 and 425. Our source audit did not recover the corresponding implementation or seed-level records. Consequently these numbers are unaudited historical context, with no role in the main diagnostic conclusions. A causal skew-parity claim would require the rectangular/square-operator distinction, the actual norm-matching procedure, and paired-seed records. Matching only the operator norm would not match the entire singular spectrum.

The associated projected spectral table reports a layer-26 radius increase from approximately 6.2 to $20{,}077$, with projected eigenvalues becoming real to reported precision. It is a result for a 128-dimensional projection. Distance preservation of a projection must not be substituted for a full nonnormal eigenspectrum certificate.

N Recorded test chronology and complete audit statistics

The archived ledger records the definition of the diagnostic, static anchors, failure of the sum aggregate, and adoption of the maximum aggregate on 8 July 2026. It then records a criterion of Spearman correlation at least 0.6 on the 25K–100K segment. The completion and automatic verdict are dated 9 July. A later re-audit dated 24 July explicitly retains the PASS while narrowing its interpretation to a recipe-specific regime correlate.

This is an internal chronological record. This preparation has not independently certified when each checkpoint first became accessible relative to every analysis decision. The immutable criterion, checkpoint-generation logs, and script versions should accompany any external claim of prospective testing. The threshold applies to the recorded segment, including step 25,000; silently dropping that boundary or moving it would change the tested statement.

Table 12: Full-precision re-audit values transcribed from the archived report. For difference rows, the stated window is applied before differencing.

Window	Predictor	n	Spearman	Pearson
25K run, step > 0	R	10	0.078788	0.010680
25K run, step > 0	$g_{\max}$	10	0.842424	0.765683
100K, all	R	41	-0.351742	-0.142244
100K, all	$g_{\max}$	41	0.250348	-0.063092
100K, $\geq$ 25K	R	31	-0.063306	0.065114
100K, $\geq$ 25K	$g_{\max}$	31	0.785484	0.088535
100K, $\geq$ 25K	ΔR	30	-0.554171	-0.671870
100K, $\geq$ 25K	$\Delta g_{\max}$	30	-0.169299	-0.774937
100K, $\geq$ 30K	ΔR	28	-0.455391	-0.388356
100K, $\geq$ 30K	$\Delta g_{\max}$	28	0.021894	-0.138131
100K, $\geq$ 50K	ΔR	20	-0.103759	0.087870
100K, $\geq$ 50K	$\Delta g_{\max}$	20	0.114286	-0.018461

O Checkpoint summaries

P Extended archived-trajectory analysis

P.1 Static anchors and competing resonance aggregates

The archived static comparison gives $R = 8.94$ for the mature Qwen3-0.6B configuration and $R = 0.11$ for its random-initialization comparator. The mature model has an alignment band at layers 17–21, with absolute alignment approximately 0.31–0.36, compared with the reported isotropic chance reference 0.0249 at $d = 1024$. The corresponding centered alignments are much smaller. These observations motivate an association with an uncentered activation direction; they do not establish that a single activation coordinate is the causal mediator.

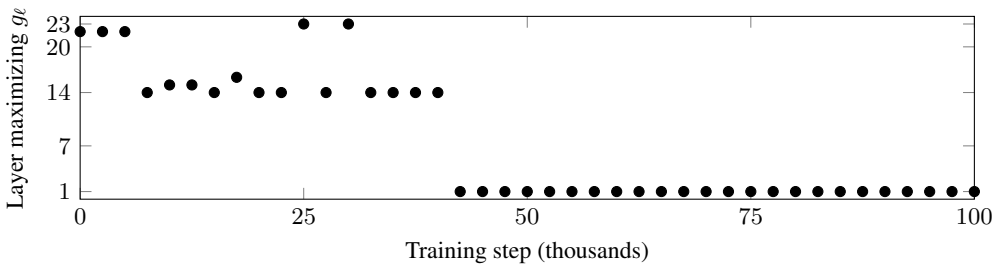


Figure 7: Historical all-layer diagnostic argmax. The layer-23 peaks at 25K and 30K lie outside the response's FFN support. These indices describe the measured statistic; they do not establish the location of a causal amplification source or its numerical robustness.

Table 13: All 41 checkpoints of the 100K-horizon run. Values are rounded source summaries.

Step	$g_{\max}$	Layer	$\mathcal{E}$	R	Step	$g_{\max}$	Layer	$\mathcal{E}$	R
0	0.15	22	3.71	0.46	52500	4.41	1	28.37	8.45
2500	3.11	22	0.45	12.46	55000	4.51	1	28.35	8.04
5000	3.32	22	0.93	12.52	57500	3.45	1	25.86	8.33
7500	4.03	14	1.06	11.42	60000	2.98	1	24.51	7.79
10000	22.29	15	0.87	10.33	62500	3.57	1	26.24	7.83
12500	65.55	15	1.04	16.22	65000	3.86	1	26.31	8.39
15000	9.14	14	1.36	9.34	67500	3.53	1	25.56	8.56
17500	23.31	16	1.27	11.55	70000	3.51	1	20.90	8.79
20000	6.67	14	9.56	10.58	72500	3.33	1	21.37	8.84
22500	7.06	14	24.03	10.07	75000	3.64	1	21.61	9.11
25000	123.27	23	36.31	14.92	77500	3.44	1	21.61	9.37
27500	5.91	14	70.19	10.08	80000	3.69	1	25.51	9.59
30000	13.36	23	52.89	15.98	82500	3.86	1	25.07	9.70
32500	4.72	14	54.00	9.77	85000	3.52	1	26.18	9.63
35000	6.47	14	56.44	7.17	87500	3.71	1	26.07	9.91
37500	5.84	14	48.75	6.63	90000	3.67	1	26.90	9.95
40000	4.93	14	42.25	7.65	92500	3.57	1	27.06	9.99
42500	3.82	1	56.42	7.28	95000	3.78	1	26.60	10.13
45000	4.64	1	40.59	9.81	97500	3.63	1	27.42	10.03
47500	3.84	1	42.13	7.31	99999	3.64	1	25.97	10.14
50000	3.61	1	34.83	8.57					

Static separation does not imply temporal prediction. On the short-horizon trajectory, excluding the ill-conditioned step-zero comparison, the originally recorded sum aggregate has rank correlation 0.078788 with surgical excess across ten checkpoint levels. The maximum aggregate has 0.842424. Choosing the maximum was a *post-hoc revision* on these data, not the successful outcome of the original hypothesis. Alignment alone was reported to perform less well; a definitive multi-feature comparison would require held-out recipes and uncertainty estimates.

P.2 Recorded prospective segment and full-trajectory behavior

After the short-run analysis, the project ledger recorded a criterion of $\rho_{\mathrm{S}}(g_{\max}, \mathcal{E}) \geq 0.6$ on steps at or after 25K of the longer run. The later 31-checkpoint segment yields 0.785484, preserving the recorded PASS. This is an internally documented prospective test. The chronology is supported here by the project records, not by an independently certified public preregistration.

Figure 8 shows why the boundary matters. Across all 41 checkpoints, the rank correlation is only 0.250348. Within the recorded segment the Pearson correlation is 0.088535, so even there the strong rank relation is not a stable linear calibration. At 25K, $g_{\max}$ peaks at 123.27; at 27.5K it falls to 5.91 while excess rises from 36.31 to its peak 70.19. This reversal is visible and is not removed by

Table 14: Constrained-training scan (25K horizon). No additional surgery is applied.

Step	$g_{\max}$	Layer	R
0	0.01	22	0.09
2500	10.80	20	17.85
5000	8.82	23	15.76
7500	4.20	22	14.23
10000	3.67	1	10.73
12500	4.31	1	6.97
15000	4.13	1	9.15
17500	3.93	1	7.31
20000	4.21	1	8.72
22500	4.09	1	9.12
24999	4.06	1	9.20

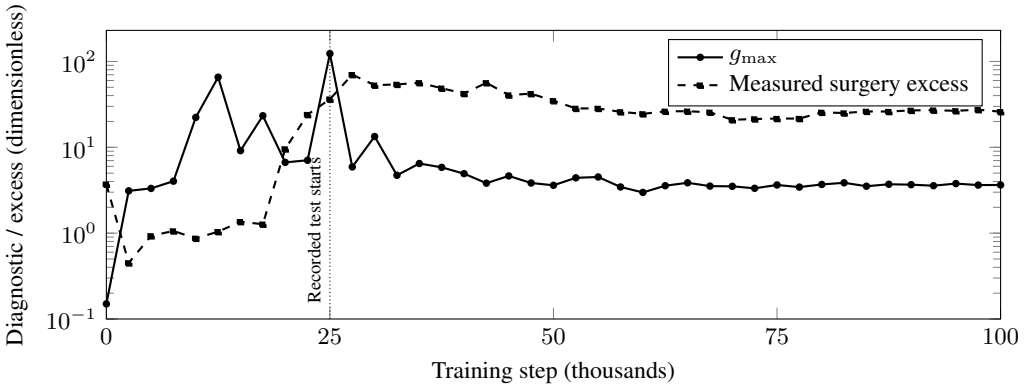


Figure 8: Full 41-checkpoint trajectory. The solid curve is the resonance diagnostic and the dashed curve is measured surgical excess; equal vertical values do not imply calibration between them. The dotted line marks the recorded late-segment test boundary. All checkpoints are shown, including early high-g observations without comparable excess and the mismatch between the two peaks. Plot values are the archived rounded summaries.

selecting the late window. Peak locations separated by one checkpoint do not establish a time-local mechanism.

The subsequent dynamical re-audit. The re-audit retains the prospective PASS and separately tests consecutive first differences. It first restricts the checkpoints to each declared window and then differences them, avoiding a boundary-crossing difference. The rank relation between $\Delta g_{\max}$ and $\Delta\mathcal{E}$ is -0.169299 after 25K, 0.021894 after 30K, and 0.114286 after 50K (Table 12). These are descriptive diagnostics on a dependent trajectory, not independent-sample significance tests. They do not identify a stable sign or magnitude for a change law.

The recorded association does not identify causal mediation by $g_{\max}$. Training phase can influence both the predictor and the response. We therefore directly compare the diagnostic with simpler measurements before interpreting its additional information.

P.3 Simple baselines and training trends

The new comparisons are exploratory: the trajectory and its earlier correlations were already known. Before computing them, we fixed the full and late windows, the response in Eq. (3), and six predictors: $g_{\max}$, maximum squared singular value, maximum squared absolute alignment, training step, original amplification, and R. Spearman correlations use average ranks. Squaring the nonnegative spectral and alignment baselines does not change their ranks. The training-step coefficient is signed; its absolute magnitude measures monotone association in either direction.

For the original all-24-layer definition, Table 15 changes the interpretation of the late result. Across all 41 checkpoints, spectrum alone has correlation 0.838, versus 0.250 for $g_{\max}$. On the late segment, $g_{\max}$ reaches 0.785, compared with 0.696 for spectrum, -0.684 for training step, and 0.743 for original amplification. Alignment alone has -0.063. Original amplification is also the denominator of $\mathcal{E}$, so its association is mathematically coupled to the response and is not independent mechanistic evidence.

To assess specified training trends, we regress predictor and response ranks on linear or quadratic standardized step ranks and correlate the residuals. The late $g_{\max}$ coefficients are 0.625 and 0.407, respectively; controlling additionally for spectrum and original amplification gives 0.366. These results reject neither arbitrary training confounding nor probe dependence. They show that the measured association is not exhausted by these particular time adjustments.

The apparent advantage over simple baselines is fragile. Removing every contiguous block of five late checkpoints makes the difference in absolute rank correlation between $g_{\max}$ and spectrum range from -0.064 to 0.304; against original amplification it ranges from -0.156 to 0.223. These are sensitivity ranges, not confidence intervals. In expanding-prefix, one-checkpoint-gap prediction of $\log \mathcal{E}$, the late-window RMSE is 0.176 for spectrum alone and 0.429 for $g_{\max}$ alone. Adding gain to a quadratic-time model improves RMSE from 0.529 to 0.254, but adding it to linear time worsens 0.327 to 0.411. Appendix R reports every fixed model, including improvements in joint spectral/activation models. Thus some additional information survives, but robust superiority over simple spectral predictors remains unestablished.

P.4 Matching response coverage and forecasting information

Matching the recorded endpoint uses layers 0–22 for every layer aggregate, keeping the response and windows fixed. The gain correlation becomes 0.211150 over all 41 checkpoints and 0.787903 late; spectrum gives 0.824913 and 0.693145 (Table 10). The late association persists. This corrected definition has no inherited prospective PASS; the archived trajectory's missing final-output response still requires its original checkpoints.

The same addendum adds a last-known-response predictor to the original ordered splits: each test block receives the final training-prefix $\log \mathcal{E}$, without using the gap label or updating within the block. Its full-precision RMSE is 0.227383 and 0.126982, lower than all 11 original OLS models and their endpoint-supported counterparts. The two split schemes first evaluate steps 42.5K and 57.5K, respectively; neither tests the 20K–30K transition. Persistence requires historical surgical labels, as does OLS calibration, and addresses that forecasting task rather than label-free diagnosis. Appendix S gives both comparisons and exact split coverage.

P.5 Depth structure and the diagnostic maximum

The original all-layer maximum shifts from late to middle layers and settles at layer 1 from approximately 42.5K onward. Layer 23 attains the maximum at 25K and 30K, but its FFN cannot cause amplification at the recorded endpoint. These points can encode training-state information, not localize the source of this response. Late in training, $g_{\max}$ is approximately 3–4 while excess remains approximately 21–28. Interpreting the argmax also requires spectral gaps and direction-stability checks, absent from the archive (Appendix S).

We retain the fixed top-one statistic and failed sum result; further aggregation choices require new evidence.

Q Extended constrained-training observations

A stable tied trajectory. The companion constrained run imposes $W_{\text{down}} = W_{\text{up}}^{\top}$ from initialization, retains an independent gate, and trains stably. Its reported training-loss summary is 0.7243, versus 0.7045 for the same-width untied run and 0.7099 for a narrower untied run at matched parameter count. The run driver defines terminal summaries as means of logged training-step losses over the last 1,000 steps; these are not independently evaluated validation losses. Complete original loss logs and the deployed token-reduction details remain unavailable. The available implementation uses the same up-projection parameter in both directions of the computation, so differentiation

Table 15: Original all-24-layer exploratory comparisons on identical E4b checkpoints and excess-amplification responses. Signed Spearman correlations; partial correlations remove a linear or quadratic trend from ranks. Checkpoints are temporally dependent. Original amplification also occurs in the response denominator.

Predictor	Full (41)	Late (31)	Late partial t	Late partial t, t^2
$g_{\max}$	+0.250	+0.785	+0.625	+0.407
$\max_\ell \sigma^2_{1,\ell}$	+0.838	+0.696	+0.180	+0.525
$\max_\ell a^2_\ell$	-0.184	-0.063	+0.111	+0.201
Training step	+0.252	-0.684	–	–
Original amplification	-0.131	+0.743	+0.413	+0.240
$\sum_\ell \log(1 + g_\ell)$	-0.352	-0.063	+0.233	-0.405

includes both uses; it does not copy a separate down parameter after each step. Appendix Y records configuration and provenance.

Within the constrained checkpoint scan, $g_{\max}$ rises from approximately 0.01 to 10.80 by 2.5K steps, then decreases to 4.20 at 7.5K and 4.06 at the endpoint. The peak-layer alignment at 2.5K is reported as 0.904. The optimizer therefore does not simply prevent the diagnostic from becoming large; a transient develops and later subsides. No surgery is needed to measure this trajectory, because the constraint already belongs to the trained forward map.

Appendix Y, Figure 15, preserves the cross-horizon diagnostic comparison.

The constrained trajectory's sum aggregate spans approximately 7–18 after initialization while the run remains stable. This is evidence against treating a high value of R as a sufficient explosion condition. Independent-gate terms, normalization, and adaptation of the full configuration are possible compensating mechanisms; the experiment does not apportion their contributions.

The earlier randomized replacement series, whose operator construction and seed-level provenance remain unresolved, is documented separately in Appendix M.

R Exploratory baseline protocol and ordered prediction

The analysis specification was written before the new comparisons, after the historical trajectory and its previously reported correlations were known. All new analyses are exploratory. The anonymous package supplies the specification, full layer records, exact computed values, and code.

Matched measurements and trend controls. Every primary comparison uses the same full 41-checkpoint E4b series or the same 31-checkpoint window starting at step 25000, with $\mathcal{E} = A_{\text{surgery}}/A_{\text{original}}$. Ranks use average ties. Linear and quadratic partial associations regress both ranked variables on an intercept, standardized step ranks, and optionally their square. The joint control adds spectral and original-amplification ranks. The late joint residual correlation has 25 residual degrees of freedom; no independent-sample test is attached. Training step is the stage baseline: the realized checkpoint-level learning-rate series was not recovered. The configured cosine schedule is not substituted for a measured series.

Temporal dependence and deletion. For each simple baseline we compare absolute Spearman correlation with that of $g_{\max}$, so an inverse monotone baseline is not discarded. All contiguous blocks of lengths 3, 5, and 8 are deleted in turn, recomputing both correlations on the same retained points. These are sensitivity ranges, not confidence intervals. First differences are formed after restricting each window, including the historical 30K and 50K sensitivity windows. Every result is retained in the accompanying JSON.

Ordered prediction. An expanding training prefix is followed by one excluded checkpoint and a contiguous test block. The full series starts with 16 training checkpoints and test blocks of 5; the late series starts with 12 and uses blocks of 4. After each split the prefix grows by the test-block size. The final block may be shorter. This yields 24 and 18 held-out predictions, respectively. Later splits may

train on earlier test checkpoints once those observations precede the split; no future observations enter a training prefix. OLS includes an intercept and training-only feature scaling. Features are $t = \text{step}/100000$, t^2, $\log(1 + g_{\max})$, $\log \max_\ell \sigma^2_{1,\ell}$, $\max_\ell a^2_\ell$, and $\log A_{\text{original}}$. The response is $\log \mathcal{E}$. No parameter or split is selected using held-out errors. These short-horizon extrapolations on one observed trajectory are not prospective or independent-seed validation.

Table 16: Original all-layer OLS models and the reviewer-triggered persistence baseline. RMSE of held-out log excess; lower is better. Test coverage is 42.5K–99,999 (24 points) for the full-prefix split and 57.5K–99,999 (18 points) for the late-prefix split. Neither tests the early transition.

Model	Test from 42.5K (24)	Test from 57.5K (18)
Intercept only	1.041	0.443
Time	2.086	0.327
Quadratic time	2.190	0.529
Spectrum	0.478	0.176
Alignment	0.880	0.457
Original amplification	0.502	0.223
Gain	0.996	0.429
Time + gain	2.209	0.411
Quadratic time + gain	2.116	0.254
Quadratic time + spectrum + amplification	0.533	0.559
Quadratic time + spectrum + amplification + gain	0.329	0.435
Last known response (added)	0.227	0.127

The added persistence baseline holds the final training-prefix log response fixed across the entire test block. It reads neither the gap response nor responses within that block. It outperforms all original OLS models. Like OLS calibration, it requires historical surgical labels; this forecasting task differs from diagnosis at a checkpoint without such labels. The original OLS improvements between nested models therefore do not demonstrate forecasting value beyond response persistence.

S Endpoint support and persistence addendum

These reviewer-triggered comparisons were specified after the earlier results and the review were available. They use the same checkpoints, response and prediction splits. The original protocol and results are retained separately in the package.

The recorded response ends at H_{23}; only FFNs in layers 0–22 can affect that endpoint. We recompute the gain, spectrum, alignment and sum aggregates on those layers at every checkpoint. Layer 23 attains the original gain maximum at steps 25K and 30K. The resulting statistic has no inherited prospective PASS. A response covering the final block output requires new forwards.

Table 17: Fixed models using endpoint-supported layer aggregates. All splits match Table 16; the historical-response baseline is unchanged.

Model	Test from 42.5K (24)	Test from 57.5K (18)
Intercept only	1.041	0.443
Time	2.086	0.327
Quadratic time	2.190	0.529
Spectrum	0.482	0.144
Alignment	0.944	0.422
Original amplification	0.502	0.223
Gain	0.843	0.281
Time + gain	2.255	0.332
Quadratic time + gain	2.018	0.553
Quadratic time + spectrum + amplification	0.817	0.553
Quadratic time + spectrum + amplification + gain	0.585	0.560
Last known response	0.227	0.127

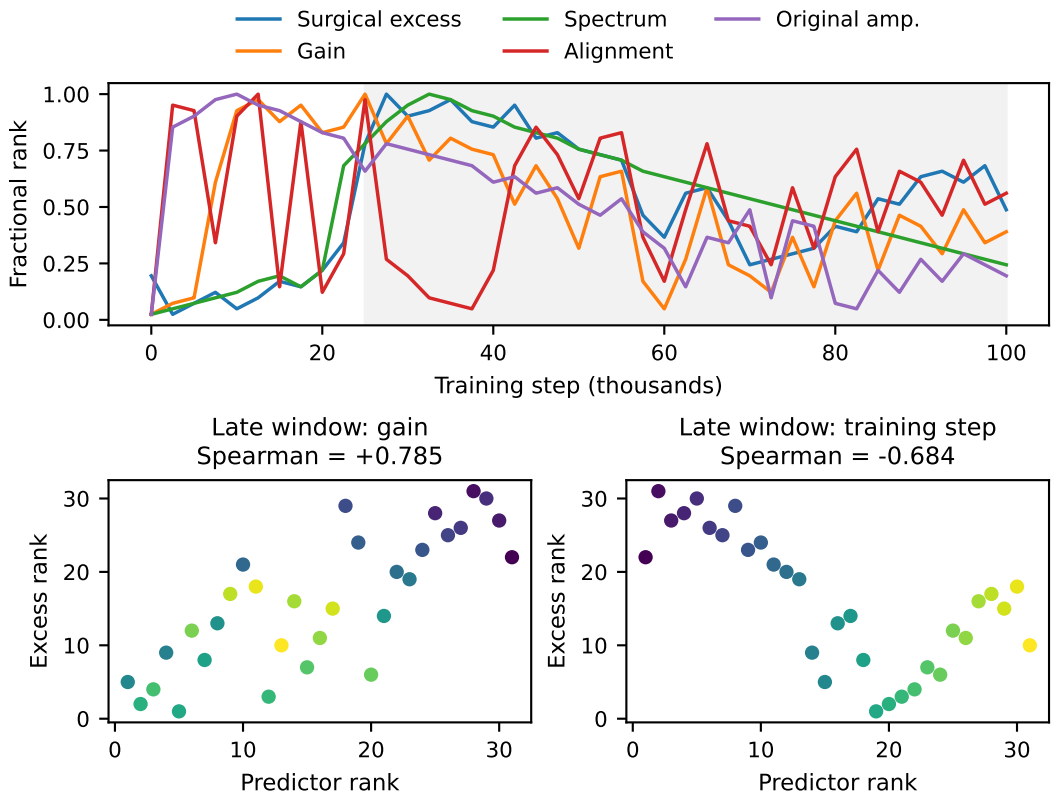


Figure 9: Full-precision rank trajectories and late-segment comparisons. Colors in the two scatterplots encode increasing training step. The shaded region in the trajectory is the recorded late segment. No fitted calibration or independent-checkpoint uncertainty is implied.

Persistence also outperforms every fixed model after matching the layer support. Full-precision persistence RMSE is 0.227382804/0.126982448, compared with 0.227270316/0.126915539 when independently reconstructing the two-decimal checkpoint table. The package records each training prefix, excluded gap, evaluated step and prediction horizon.

Spectral identifiability. For $U = I_2$ and $r = e_1$, either e_1 or e_2 is a valid leading right singular vector, giving gain 1 or 0 for the same Gram operator. Leading activation directions have the analogous issue at repeated top eigenvalues. The archived JSON records a leading singular value and activation top-energy fraction, but not both top spectral gaps or iteration residuals. The scanner uses one fixed power-iteration initialization for weights and an eigendecomposition for the activation direction. Its initial-layer SVD smoke check cannot establish directional stability at the peak checkpoints. The follow-up protocol specifies gap, residual and multi-initialization checks before interpreting a maximizing layer as mechanistically informative.

T PUBLIC-WEIGHT FORWARD CONTROLS AND COMPLETE COMPARISONS

Executed model and runtime. The fixed public Qwen3-0.6B revision is `c1899de289a04d12100db370d81485cdf75e47ca` for both weights and tokenizer. The loaded model has 596,049,920 parameters, 28 layers, hidden width 1024 and intermediate width 3072. Stored weights are evaluated in float32, with eager attention, dropout disabled, no cache, no generation and 16 CPU threads. Runtime versions are Python 3.10.12, PyTorch 2.8.0+cpu, Transformers 4.56.2, NumPy 2.2.6, huggingface-hub 0.36.2, safetensors 0.8.0 and tokenizers 0.22.2. Exact model files and measurement code are hashed. This recovers a public model; it does not restore the historical E4b training checkpoints.

Tokens and outcome conventions. The library, river and repair passages contain 125, 125 and 135 tokens with no special tokens or chat template. They are processed separately without padding or truncation. Logits at positions $0, \ldots, N-2$ predict tokens $1, \ldots, N-1$; NLL sums are accumulated in float64 over 124, 124 and 134 targets. The reported combined loss weights passages by those counts. All 29 residual depths are measured before final normalization, using population standard deviation over real tokens and features. The entry response uses layers 0–26 for diagnostic aggregation; the complete output uses layers 0–27. In all nine last-layer-only edits across arms and probes, the captured last-block entry is unchanged exactly.

Fixed and sequential protocols. Before any new public-model forward, the first protocol fixed three probes, two edit arms, all-layer edits and every single-layer edit: 177 evaluations including originals. After seeing those results, a separately locked addendum added branch-RMS matching. Its calibration concatenates token IDs from the ten historical passages with no inserted separators, totaling 438 tokens. It uses only original-model branch inputs and outputs. Coefficients are fixed across the three new texts; no edited response or loss enters calibration. This adds one calibration forward, three original rechecks and 87 edited forwards. The original rechecks reproduce the first-stage losses exactly. All-layer and single-layer records, full activation profiles, every tested predictor, protocol hashes and chronology accompany the supplement.

Numerical checks. Float64 eigendecomposition of $U^\top U$ supplies the weight reference; thin SVD of the actual FFN-input matrix supplies the activation direction. Relative top singular gaps range from 0.0152 to 0.4688, and activation-moment gaps from 0.6988 to 0.9529. Maximum reference eigenpair residuals are 3.4×10^{-15} and 1.2×10^{-14}. Across all layers and four initializations, 200-iteration power directions have minimum absolute cosine 0.9999805 with the reference and maximum residual 1.89×10^{-4}; at 2000 iterations these become $1 - 2.7 \times 10^{-15}$ and 2.7×10^{-15}. The largest 200-iteration absolute gain error is 3.15×10^{-5} (relative error reaches 0.109 where gain is small). Direct SVD at fixed layers 1, 14, 26 and 27 agrees in direction to absolute cosine at least $1 - 10^{-15}$. These checks establish numerical accuracy for the fresh public-model measurements only.

What scale matching fixes. Frobenius coefficients range from 0.8251 to 1.0149; calibration branch coefficients range from 0.2927 to 0.9604. Local branch matching on the calibration tokens has maximum relative RMS error 8.3×10^{-8}. On evaluation inputs the corresponding branch-RMS ratios span 0.904–1.908. The controls therefore compare precisely defined amplitude conventions; they do not hold the edited network's own inputs or its function fixed.

U SEQUENTIAL CONTROLS OF RESIDUAL LENGTH AND DIRECTION

Motivation and chronology. Exploratory reanalysis of all 29 recorded residual depths found that blocks 19, 26 and 27 have the three largest relative standard-deviation steps in every passage and edit arm. A relative step divides the edited between-depth gain by the original gain; it is not a Jacobian estimate. This observation motivated a separately locked state-intervention protocol after the 268-forward study. Cuts 7, 14 and 21 follow architecture quarters; 26, 27 and 28 examine the terminal pair. All six cuts, all three passages and all three edit arms were fixed before the new model outcomes. No weight, branch or diagnostic coefficient was refit.

State intervention and original suffix. At cut k, down projections in blocks $0, \ldots, k-1$ receive the specified transpose edit. The remaining blocks retain their original weights. For every token, let h_o and h_e be the original and edited residual states at H_k, with lengths $a = \|h_o\|$ and $b = \|h_e\|$. Compare h_e, $(a/b)h_e$, and $(b/a)h_o$ before running the original suffix. The second state restores original length while retaining edited direction; the third restores original direction while retaining edited length. These controls use clean states from the same text and are therefore oracle state interventions, not a deployable repair or a weight-spectrum-preserving alignment edit. A residual direction contains the entire preceding computation, so these results do not isolate the first singular direction or a particular gate term.

Execution and boundary checks. The same immutable weights, tokenizer, three passages, coefficients, float32 forward environment and next-token convention are used. Three original rechecks

Table 18: All fixed single-layer comparisons. Each interval is the minimum and maximum signed Spearman correlation across the three passages, with 28 layers per passage. These are descriptive ranges, not confidence intervals. The branch-matched arm is the sequential addendum.

Edit	Predictor	$\rho(\cdot, \mathcal{E}_{28})$	$\rho(\cdot, \Delta\mathcal{L})$
Unit	g_ℓ	$[0.194, 0.217]$	$[0.271, 0.373]$
Unit	σ_1^2	$[0.141, 0.229]$	$[-0.009, 0.075]$
Unit	Alignment^2	$[0.114, 0.157]$	$[0.218, 0.322]$
Unit	Relative branch perturbation	$[-0.107, 0.013]$	$[-0.120, 0.022]$
Unit	Branch RMS ratio	$[0.126, 0.239]$	$[0.633, 0.709]$
Unit	Relative Frobenius edit	$[-0.076, 0.052]$	$[-0.413, -0.271]$
Frobenius	g_ℓ	$[0.138, 0.190]$	$[0.261, 0.340]$
Frobenius	σ_1^2	$[0.074, 0.221]$	$[0.001, 0.056]$
Frobenius	Alignment^2	$[0.062, 0.140]$	$[0.205, 0.283]$
Frobenius	Relative branch perturbation	$[-0.055, 0.059]$	$[-0.157, -0.032]$
Frobenius	Branch RMS ratio	$[0.175, 0.374]$	$[0.683, 0.729]$
Frobenius	Relative Frobenius edit	$[-0.089, 0.099]$	$[-0.461, -0.323]$
Branch	g_ℓ	$[0.144, 0.149]$	$[0.343, 0.353]$
Branch	σ_1^2	$[0.033, 0.082]$	$[0.078, 0.134]$
Branch	Alignment^2	$[0.055, 0.094]$	$[0.275, 0.285]$
Branch	Relative branch perturbation	$[-0.398, -0.333]$	$[-0.493, -0.264]$
Branch	Branch RMS ratio	$[0.278, 0.429]$	$[0.104, 0.305]$
Branch	Relative Frobenius edit	$[-0.576, -0.501]$	$[-0.581, -0.423]$

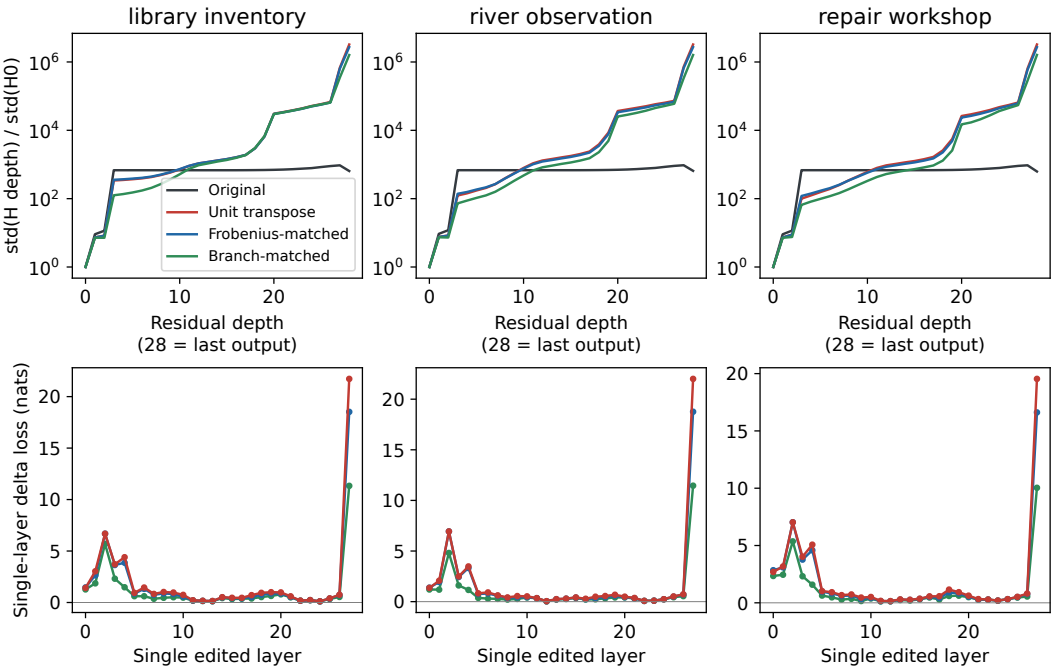


Figure 10: Fresh public-model measurements. Top: all-layer edits, with each depth normalized by the original input standard deviation; depth 28 is the actual last-block output. Bottom: next-token loss damage from editing one layer at a time. Branch matching uses the separate calibration text. Every new passage and edited layer is shown.

plus $3 \times 3 \times 6 \times 3 = 162$ interventions give 165 new forwards, or 433 across all stages. All original and cut-28 unpatched-prefix results reproduce their preceding records. The 108 actual patches have maximum relative length error 4.93×10^{-8} and unit-direction error 4.18×10^{-8}; pre-patch states reproduce exactly. Tiny-model tests verify that a complete clean-state reset reproduces the

Table 19: Sequential state controls with an original suffix. Losses average 382 next-token targets; original loss is 4.211. At cut k, only blocks $0, \ldots, k-1$ are edited. The two patches restore either the original per-token length or direction. Cut 28 precedes final RMSNorm and has no Transformer suffix; its scaling invariance is a readout control.

Edit	Cut k	Edited prefix	Original length	Original direction
Unit	7	13.923	15.712	5.195
Unit	14	16.123	15.520	6.550
Unit	21	17.040	23.095	7.087
Unit	26	15.871	22.852	5.091
Unit	27	30.673	14.964	4.591
Unit	28	35.382	35.382	4.211
Frobenius	7	13.669	15.580	5.214
Frobenius	14	15.565	15.600	6.490
Frobenius	21	17.227	23.080	7.069
Frobenius	26	15.754	22.846	5.090
Frobenius	27	30.349	14.968	4.591
Frobenius	28	35.492	35.492	4.211
Branch	7	14.535	15.152	5.118
Branch	14	14.703	15.754	6.070
Branch	21	18.785	23.150	7.012
Branch	26	15.819	22.822	5.089
Branch	27	27.235	14.944	4.591
Branch	28	35.773	35.773	4.211

original suffix and loss exactly, check non-collinear tokenwise swaps, and check exception cleanup. Together with the four previous measurement tests, seven forward-runtime tests pass. The offline package verifies records and calculations without rerunning those model tests or forwards.

Complete outcomes and endpoint interpretation. Restoring the original direction gives lower loss than restoring original length in all 45 passage/arm/cut comparisons before the final-output cut. These dependent comparisons are reported individually, without an IID significance calculation. Neither patch is assumed to preserve the function, and neither fixes every outcome. At cut 27, restoring original length still gives full-output excess between 459.25 and 485.70 through the unmodified final block. Restoring original direction reduces token-weighted loss to approximately 4.591 across arms, while excess ranges from 455.42 to 1106.34. At cut 26, length restoration instead increases token-weighted loss from 15.75–15.87 to 22.82–22.85; restoring direction gives 5.089–5.091. Every other fixed cut remains in the table and figure.

At cut 28, the only remaining computation is final RMSNorm and the output head. Positive per-token rescaling is approximately removed by RMSNorm (its epsilon and floating-point arithmetic are retained). Accordingly, restoring length leaves loss near 35 while restoring direction returns approximately the original 4.211. This endpoint is an expected readout control and is not evidence of a newly identified multi-block amplification mechanism.

V CLEAN-STATE DIRECT-READOUT CONTROL

For the three fixed synthetic probes, we compare the existing direction-restored states from the three transpose arms (unit, Frobenius matched and branch matched) with a fixed-head direct readout. One clean original forward per model/probe supplies every previously selected cut state $h_o^{(k)}$. We omit all remaining Transformer blocks and compute

$$\mathcal{L}_{\mathrm{direct}}(k) = \mathrm{CE}\Big(W_{\mathrm{head}}\ \mathrm{RMSNorm}_{\mathrm{final}}(h_o^{(k)}), y\Big)\,.$$

This is a logit-lens-style use of the existing final normalization and output head (nostalgebraist, 2020; Belrose et al., 2023), not a learned probe. The protocol fixes the existing cuts and token IDs without fitting a coefficient or choosing a cut from the new outcomes. The six new clean forwards reproduce the previous original losses within 10^{-6} before their readouts are accepted. Model-file

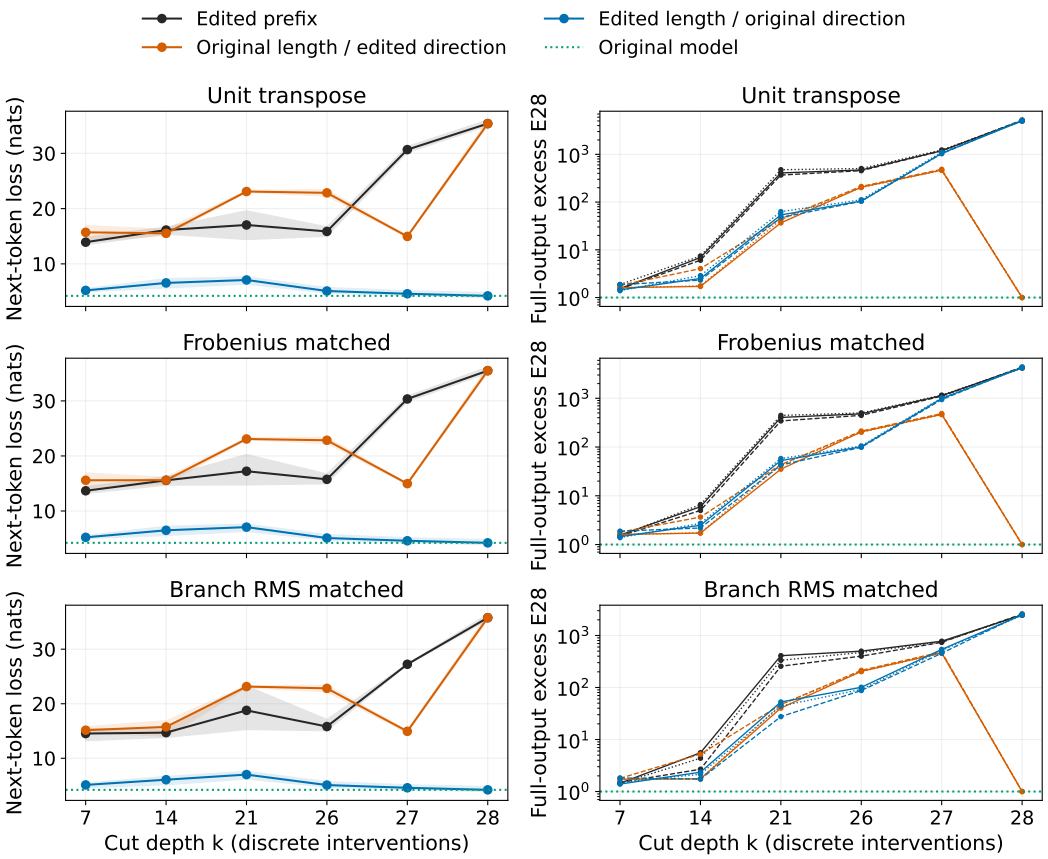


Figure 11: All fixed state-intervention cuts, spaced equally as discrete conditions. Left: token-weighted loss; shaded ranges span the three passages and are not confidence intervals. Right: full-output excess for each passage (solid, dashed and dotted curves), using the same mode colors. Edited prefixes are followed by original suffixes. Cut 28 has no remaining Transformer block.

Table 20: Fixed-head direct readout and direction-restored original suffixes on the three synthetic probes. NLL is target-weighted separately within each model. Intervals span the three transpose arms, not uncertainty. The final column is direct-readout NLL minus direction-plus-suffix NLL; it is a descriptive comparison, not an exactly scale-matched causal effect.

Model	Cut	Direct	Direction + suffix	NLL difference
Qwen3-0.6B	7	11.8305	[5.1181, 5.2138]	[6.6167, 6.7124]
	14	10.2118	[6.0703, 6.5497]	[3.6622, 4.1415]
	21	7.3022	[7.0117, 7.0867]	[0.2154, 0.2905]
	26	5.1060	[5.0893, 5.0912]	[0.0148, 0.0167]
	27	4.5916	[4.5906, 4.5911]	[0.000511, 0.001024]
	28	4.2107	[4.2107, 4.2107]	$[-1.72, 6.30] \times 10^{-8}$
SmolLM2-360M	8	16.5938	[5.6711, 6.0728]	[10.5210, 10.9227]
	16	14.1863	[6.0199, 6.5642]	[7.6220, 8.1664]
	24	9.3526	[5.6555, 5.8635]	[3.4890, 3.6971]
	30	4.8760	[4.5688, 4.5935]	[0.2824, 0.3072]
	31	4.2884	[4.2021, 4.2078]	[0.0807, 0.0863]
	32	3.5130	[3.5130, 3.5130]	$[-0.39, 1.01] \times 10^{-7}$

hashes are checked against the earlier locks. The execution note records Python 3.12 on the new CPU runner, rather than the earlier Python 3.10.12 runtime. Head-only evaluations are not counted as full-model forwards.

The direction intervention uses $c_i h_{o,i}^{(k)}$ with $c_i = \|h_{e,i}^{(k)}\|/\|h_{o,i}^{(k)}\| > 0$ for each token, while the direct control uses the unscaled clean state. For a final RMSNorm with diagonal gain Γ,

$$\mathrm{RMSNorm}_\epsilon(ch) = \frac{\Gamma h}{\sqrt{\mathrm{mean}(h^2) + \epsilon/c^2}}.$$

Thus positive scaling is only approximately removed at finite ϵ. The NLL difference is a comparison with this fixed-head baseline, not an exactly matched estimate of the suffix's isolated causal contribution.

At earlier cuts, the direction-restored original suffix has substantially lower NLL than direct readout: the weighted difference is 6.6167–6.7124 nats at Qwen cut 7 and 10.5210–10.9227 at SmolLM2 cut 8. This contrast persists through Qwen cut 14 and SmolLM2 cut 24. At Qwen cut 27 the difference is only 0.000511–0.001024 nats; SmolLM2 cut 31 retains 0.0807–0.0863. Hence the extra NLL advantage depends strongly on depth, with little remaining advantage at the late Qwen cuts.

Comparable average NLL does not establish identical logits, distributions, or negligible residual updates. These observables were not measured by this audit, so it does not demonstrate that the suffix stops computing. Likewise, poor fixed-head readout at an earlier layer does not show that its representation lacks task information; a learned readout can behave differently (Belrose et al., 2023). The table covers the synthetic probes and the three transpose arms. The ordinary-edit and natural-text direction-versus-length results are separate evidence; the new direct-readout comparison is not claimed to validate their entire panel or to identify a tying-specific mechanism.

W FIXED TRANSFER STUDY ON A SECOND PUBLIC MODEL

Selection, configuration and chronology. After the 433 Qwen forwards, we selected one different public model before measuring any of its pretrained forwards: HuggingFaceTB/SmolLM2-360M (Ben Allal et al., 2025), revision `f8027fd0eaeea54caa13c31d31b9fdc459c38b49`. Selection required public access, a different provider/recipe, fewer than 1B parameters and compatible bias-free SwiGLU/pre-RMSNorm blocks; no models were screened by intervention outcomes. The loaded model has 361,821,120 parameters, 32 blocks, hidden width 960, FFN width 2560, 15 query heads, 5 KV heads, vocabulary 49152, shared input/output embeddings and RMSNorm epsilon 10^{-5}. Model/tokenizer files, protocol, token IDs and exact code were hashed before execution. This is a different-recipe comparison, not a controlled training-seed replication.

Fixed measurements. The same three texts yield 125, 125 and 136 tokens, or 383 next-token targets, under this tokenizer. The separate calibration sequence yields 515 tokens. Special tokens, padding, truncation and chat templates are omitted as before. Within-model NLL is target-weighted; losses are neither pooled nor used to compare model quality across tokenizers. The unchanged float32/eager/16-thread measurement and float64-reference implementations repeat all three weight edits, all 32 single-layer edits and the six architecture-relative cuts $\{8, 16, 24, 30, 31, 32\}$. Each cut repeats the edited-prefix, original-length and original-direction states before an original suffix. The final cut is a readout control. One calibration, three originals, 288 single-layer edits, 162 prefix/state conditions and three original rechecks give 457 forwards. All outcomes and all six original-input predictors are retained.

Execution checks. Seven tiny-model measurement tests passed on the Llama implementation, covering the same hook, next-token, causal-support, patch and restoration checks used for Qwen. All original rechecks reproduce their losses and profiles exactly, every last-layer-only edit preserves H_{31} exactly, and original weights are restored. All 108 state patches reproduce their pre-patch state exactly; length and direction errors are below 4.8×10^{-8}. Statistical replay checks the complete inventories, loss counts, executed-source hashes and all fixed comparisons, without rerunning model forwards.

Scale and numerical checks. Frobenius coefficients range from 0.8819 to 1.0186; calibration branch coefficients range from 0.2605 to 1.3868. Calibration RMS relative error is below 5.3×10^{-8}; original-evaluation-input branch ratios range from 0.375 to 1.289. Relative weight gaps span 0.0226–0.5176, and input-moment gaps 0.6202–0.9899. Maximum reference residuals

Table 21: Every fixed SmolLM2 single-layer predictor. Intervals are descriptive minima and maxima across three texts, with 32 dependent layers per text, not confidence intervals.

Edit	Predictor	$\rho(\cdot, \mathcal{E}_{32})$	$\rho(\cdot, \Delta\mathcal{L})$
Unit	g_ℓ	$[-0.123, -0.092]$	$[0.120, 0.315]$
Unit	σ_1^2	$[0.146, 0.185]$	$[0.597, 0.777]$
Unit	Alignment^2	$[-0.175, -0.127]$	$[0.071, 0.192]$
Unit	Relative branch perturbation	$[0.161, 0.238]$	$[0.609, 0.755]$
Unit	Branch RMS ratio	$[-0.214, -0.110]$	$[-0.271, -0.207]$
Unit	Relative Frobenius edit	$[0.162, 0.256]$	$[0.637, 0.639]$
Frobenius	g_ℓ	$[-0.118, -0.091]$	$[0.119, 0.306]$
Frobenius	σ_1^2	$[0.144, 0.167]$	$[0.583, 0.775]$
Frobenius	Alignment^2	$[-0.170, -0.100]$	$[0.074, 0.184]$
Frobenius	Relative branch perturbation	$[0.209, 0.248]$	$[0.618, 0.769]$
Frobenius	Branch RMS ratio	$[-0.096, -0.066]$	$[-0.376, -0.317]$
Frobenius	Relative Frobenius edit	$[0.224, 0.291]$	$[0.657, 0.690]$
Branch	g_ℓ	$[-0.097, -0.071]$	$[0.134, 0.307]$
Branch	σ_1^2	$[0.130, 0.187]$	$[0.474, 0.754]$
Branch	Alignment^2	$[-0.144, -0.107]$	$[0.113, 0.184]$
Branch	Relative branch perturbation	$[0.162, 0.166]$	$[0.659, 0.714]$
Branch	Branch RMS ratio	$[-0.485, -0.303]$	$[-0.469, -0.440]$
Branch	Relative Frobenius edit	$[0.278, 0.348]$	$[0.572, 0.709]$

Table 22: All SmolLM2 state-control cuts, original suffix retained. Losses are weighted over 383 next-token targets; original loss is 3.513. Cut 32 has only final RMSNorm and readout remaining.

Edit	Cut k	Edited prefix	Original length	Original direction
Unit	8	14.293	14.307	6.073
Unit	16	17.963	12.373	6.564
Unit	24	20.506	14.490	5.864
Unit	30	20.138	20.038	4.594
Unit	31	18.686	15.639	4.203
Unit	32	43.662	43.662	3.513
Frobenius	8	14.287	14.345	6.037
Frobenius	16	17.874	12.499	6.525
Frobenius	24	20.459	14.471	5.814
Frobenius	30	20.044	20.061	4.587
Frobenius	31	18.409	15.701	4.202
Frobenius	32	43.183	43.183	3.513
Branch	8	14.561	14.796	5.671
Branch	16	17.335	13.123	6.020
Branch	24	21.412	14.171	5.656
Branch	30	20.482	20.725	4.569
Branch	31	17.558	15.759	4.208
Branch	32	43.942	43.942	3.513

are 5.4×10^{-15} and 1.2×10^{-14}. At 200 power iterations, minimum reference cosine is 0.9999989, maximum residual 6.87×10^{-5}, and maximum absolute/relative g errors are $0.00240/0.00430$. At 2000 iterations the minimum cosine is $1 - 1.3 \times 10^{-15}$ and maximum residual 2.3×10^{-15}. Direct SVD at fixed indices 1, 14, 26 and 27 agrees to cosine at least $1 - 10^{-15}$. Primary diagnostics use the float64 reference directions.

What transfers and what varies. Direction restoration gives lower loss than length restoration in all 45 pre-final comparisons in this model, matching the sign in all 45 Qwen comparisons. These are two models, not 90 independent replications. At $k = 31$, direction restoration gives weighted loss 4.202–4.208, versus 15.639–15.759 for length restoration and original 3.513. However, the latter has only 4.62–5.69-fold final-output excess, versus 459–486 in Qwen. All-layer SmolLM2

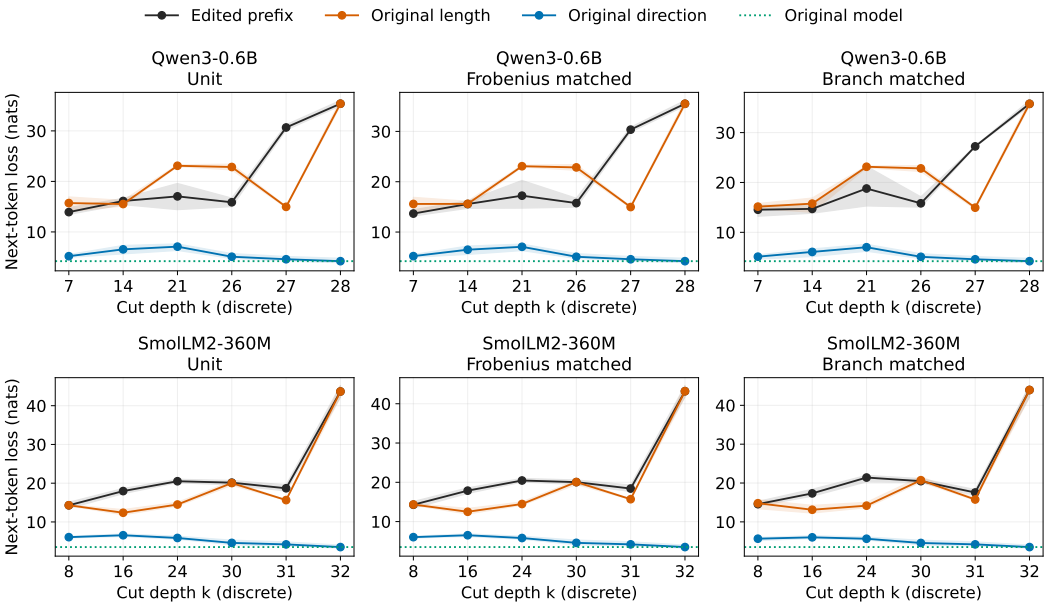


Figure 12: Fixed state controls in both public models. Cuts are spaced equally as discrete conditions; each panel uses its own tokenizer and original-loss reference. Lines are within-model target-weighted losses; shading spans the three passages, not uncertainty over independent runs. No model-quality comparison is inferred from absolute loss across tokenizers. The last cut in each row is the final-normalization readout control.

excess is 16.81–22.71 across the three edits, while its loss still rises sharply. Likewise, spectrum has stronger positive single-layer loss association than g, and the branch-RMS association is negative. The directional contrast recurs; amplification magnitude and simple diagnostic ranking vary across these recipes.

X Ordinary edits and longer natural-text validation

The ordinary-edit protocol was fixed after the two-model results and before its 770 new forwards. The natural-text protocol was fixed before its 392 forwards. The anonymous package includes both exact protocols, input locks, executed runners and raw records. Each model keeps its original revision, runtime and tokenizer. The reference is the previously calibrated branch-matched transpose, rather than a newly optimized edit.

Magnitude controls. Independently per layer, draw float32 standard-normal R_ℓ with a local CPU generator seeded by $s+1009\ell$, for $s \in \{20260916, 20260917, 20260918\}$. Apply $D_\ell + \lambda_\ell R_\ell$, where the two predeclared coefficients are

$$\lambda_{F,\ell} = \frac{\|T_\ell - D_\ell\|_F}{\|R_\ell\|_F}, \qquad \lambda_{B,\ell} = \frac{\mathrm{RMS}(T_\ell z_\ell - D_\ell z_\ell)}{\mathrm{RMS}(R_\ell z_\ell)}.$$

The two rules separately match realized weight change and original-calibration-input branch perturbation. They do not jointly match replacement spectrum, output scale, or final functional damage. All layerwise coefficients and realized matches are exported, including their transfer to original evaluation inputs. Perturbation seeds are not training seeds. Original suffixes, clean-state oracle patches and final-normalization controls retain their earlier meaning.

Text and inventory. The short-text study evaluates seven arms, six cuts, three states and three texts (378 intervention forwards), plus one calibration, three originals and three clean rechecks per model. The natural study evaluates seven arms, three cuts, three states and three texts (189), plus the

Table 23: All pre-final direction-minus-length loss contrasts. Counts and ranges are dependent within-model comparisons, not independent-run uncertainty.

Text	Model	Family	Direction lower	Contrast range
Short	qwen	Transpose	15/15	$[-18.08, -8.88]$
Short	qwen	Random: weight match	45/45	$[-10.88, -5.56]$
Short	qwen	Random: branch match	45/45	$[-11.63, -7.67]$
Short	smollm2	Transpose	15/15	$[-16.37, -5.87]$
Short	smollm2	Random: weight match	45/45	$[-10.57, -5.43]$
Short	smollm2	Random: branch match	45/45	$[-12.20, -7.14]$
Natural	qwen	Transpose	6/6	$[-13.34, -8.19]$
Natural	qwen	Random: weight match	18/18	$[-12.03, -7.55]$
Natural	qwen	Random: branch match	18/18	$[-13.41, -7.95]$
Natural	smollm2	Transpose	6/6	$[-15.96, -5.26]$
Natural	smollm2	Random: weight match	18/18	$[-11.14, -8.48]$
Natural	smollm2	Random: branch match	18/18	$[-11.80, -9.20]$

same seven original/calibration forwards per model. Thus these stages add 770 and 392 forwards; the shared inventory reports 2058 full-model forwards across all completed studies, including the six clean forwards from the direct-readout audit. All-layer outputs are reused from the final unpatched cut and counted once.

Natural text comes from `Salesforce/wikitext`, revision `b08601e04326c79dfdd32d625aee71d232d685c3`, `wikitext-2-raw-v1/test`. Article bodies are formed by concatenating source rows until the next single-equals heading. Take the first 8192 Unicode characters and select the first three articles having at least 1024 tokens under both pinned tokenizers; retain the first 1024 IDs for each tokenizer. They are Robert Boulter, Du Fu, and Kiss You (One Direction song). Source row spans, hashes, excerpts, extraction code and Wikipedia/dataset attribution are supplied. The final token may end at a different character under each tokenizer. No model outcome participates in selection; source presence in public-model training data is not excluded.

Execution checks. Three new runtime tests verify realized matching, clean-state recovery of an original suffix, and restoration after exceptions or partial edit setup. They pass on both tiny Qwen3 and Llama implementations (six executions). Shared hook/NLL/state tests previously passed on both implementations (14 executions). All original rechecks and pre-patch states are exact; original weights are restored after each cut. Natural-study calibration metrics and every replacement-matrix hash exactly match the preceding study.

Short qwen: original loss 4.210725 over 382 targets; maximum local matching relative error 5.71×10^{-8}, patch-length error 4.67×10^{-8}, and patch-direction error 4.07×10^{-8}.

Short smollm2: original loss 3.513031 over 383 targets; maximum local matching relative error 5.40×10^{-8}, patch-length error 4.91×10^{-8}, and patch-direction error 4.25×10^{-8}.

Natural qwen: original loss 2.872617 over 3069 targets; maximum local matching relative error 5.71×10^{-8}, patch-length error 5.15×10^{-8}, and patch-direction error 4.29×10^{-8}.

Natural smollm2: original loss 2.377545 over 3069 targets; maximum local matching relative error 5.40×10^{-8}, patch-length error 4.26×10^{-8}, and patch-direction error 4.42×10^{-8}.

The shared direction-restoration contrast constrains its interpretation: clean-state direction contains task-relevant information, so recovery alone does not distinguish transpose tying from ordinary edits. The controlled amplitude comparison shows that equal local perturbation magnitude does not imply equal downstream propagation. Spectral structure, gates, normalization and changed trajectories remain candidate contributors; these tests do not isolate a unique mediator or match final functional damage.

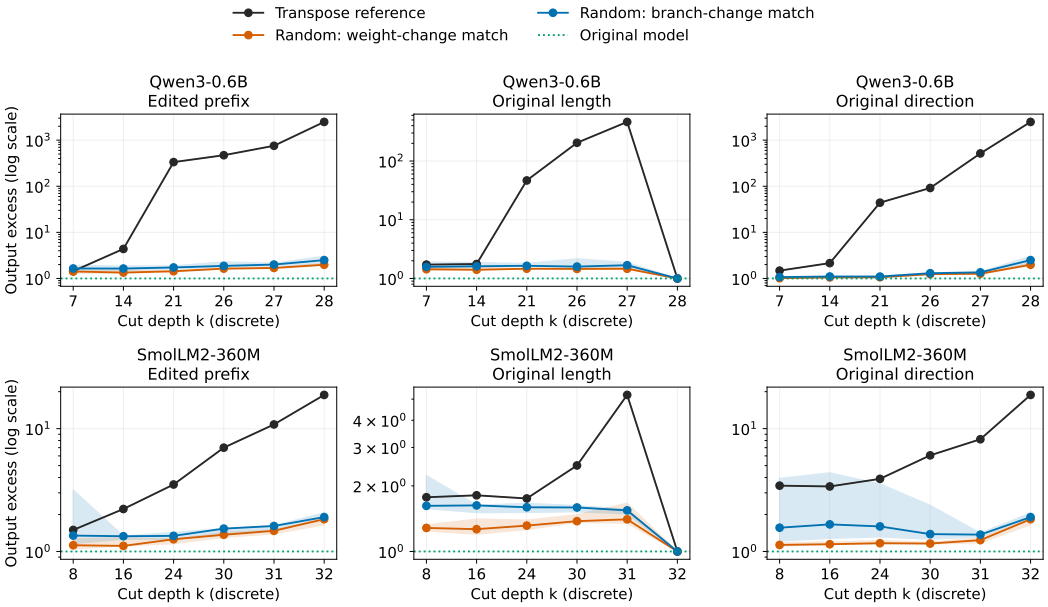


Figure 13: Short-text propagation under the fixed edits and state controls. Each seed first takes the median excess over the three passages; random-family lines and shading show the median and range over three perturbation seeds. The reference is its three-passage median. Cuts are discrete, equally spaced conditions, and the final cut is the readout control. Ranges are not confidence intervals. All loss curves and full seed/text profiles are supplied in the reproduction package.

Y Reproduction and source separation

The manuscript retains rounded 41-checkpoint and 11-checkpoint summaries for its historical figures and tables. The anonymous reproduction package additionally contains full-precision archived layer records for the short, long, and constrained trajectories and the static anchors. Only private checkpoint-location metadata is removed; scientific numerical fields are preserved. The baseline analysis validates layer inventories, singular-value/alignment products, stored maxima, finite values, exact checkpoint grids, and the identity of measured surgical excess. The short-run response itself was rounded in the original record and is explicitly treated as such.

The package regenerates the statistics, tables and figures from the archived and fresh forward records without a model or private repository. It includes exact analysis-environment versions, executable tests, hash manifests and measurement sources. The new public-model experiments supply immutable weights/tokenizer identities, executed source and a separate pinned forward environment. Historical E4b model-level replication still requires readable checkpoints and verified historical runtime versions. Available historical source snapshots are not a certified copy of that deployed environment.

The original source-location mapping and internal editorial records are excluded from the anonymous package. The package's file hashes establish integrity of the exported records, not an independently timestamped preregistration. Reproduction of an archived correlation does not independently certify when the corresponding prediction was first specified.

Training implementation and reported loss. The archived recipe specifies 24 blocks, hidden width 1536, FFN width 4096 (2731 in the narrower comparison), head dimension 128, 12 query heads, 4 key/value heads, vocabulary 151936, and RMSNorm epsilon 10^{-6}. Input/output embeddings are shared, logits are float32, and the recipe disables BOS insertion. The archive's "600M" designation is a run-family label. Its ledger reports total parameter counts near 0.988B for the untied and 0.837B for the tied/narrower arms, but those totals are not consistent with a direct count from the available configuration alone. Under the listed shared-embedding GQA/SwiGLU dimensions, the

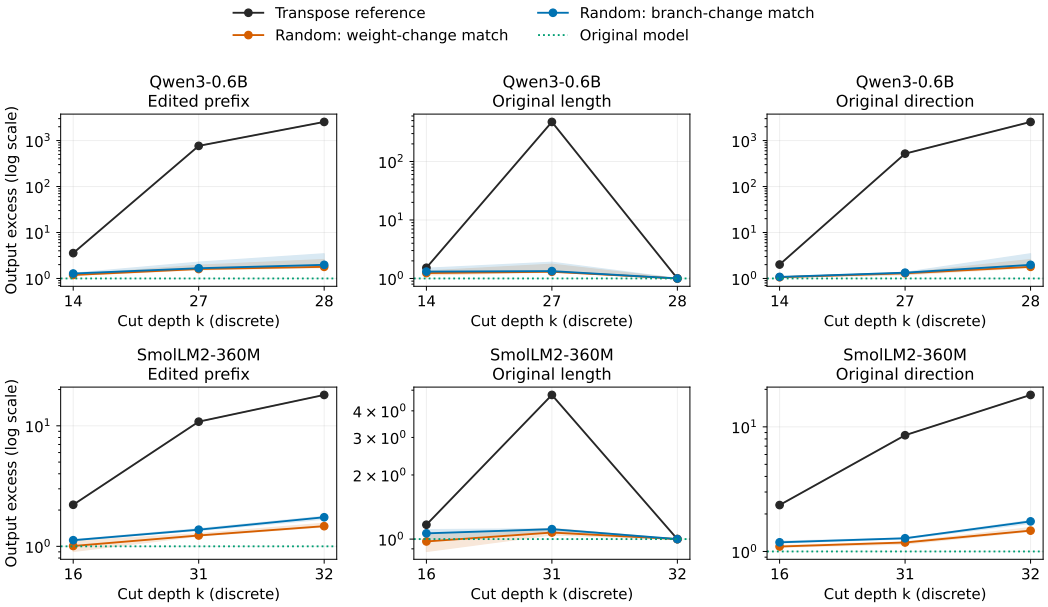


Figure 14: Natural-text propagation under the fixed edits and state controls. Each seed first takes the median excess over the three passages; random-family lines and shading show the median and range over three perturbation seeds. The reference is its three-passage median. Cuts are discrete, equally spaced conditions, and the final cut is the readout control. Ranges are not confidence intervals. All loss curves and full seed/text profiles are supplied in the reproduction package.

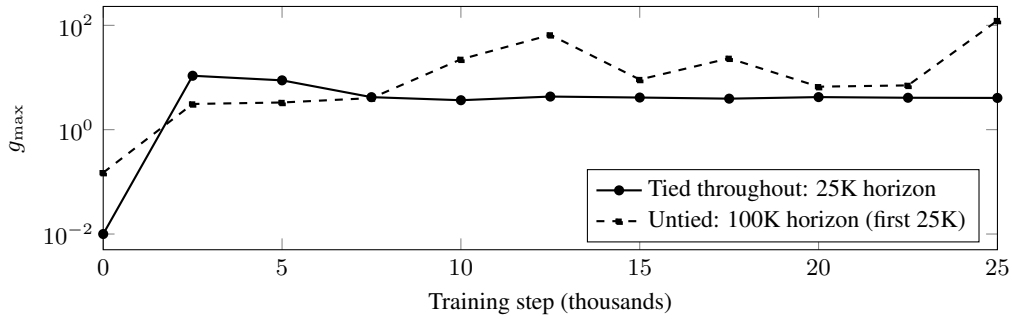


Figure 15: Constrained-training diagnostic and the first 25K steps of the longer untied run. The schedules have different cosine horizons. Their peak ratio, approximately 0.09, is therefore a descriptive cross-horizon contrast, not a clean estimate of the causal effect of tying under an identical learning-rate trajectory.

corresponding direct counts are approximately 0.8374B (untied, FFN width 4096), 0.6864B (tied, FFN width 4096), and 0.6865B (untied, FFN width 2731), up to small normalization/bias terms. We therefore treat the ledger totals as unresolved historical provenance rather than configuration-derived counts. The provided YAML and model sources specify the available configuration, but are not a recovered immutable deployment manifest; a missing historical module or a ledger counting discrepancy cannot be distinguished from the surviving records.

AdamW uses weight decay 0.1, gradient clipping, peak learning rate 3×10^{-4}, 1% warmup, and cosine decay to one tenth of the peak. Each step processes $64 \times 2048 = 131{,}072$ tokens. The 25K and 100K cosine horizons therefore do not provide an identical-schedule comparison.

The rank-zero tied FFN creates one up-projection parameter and uses it in both Ux and $U^\top h$; no independent down-projection or corrective parameters exist in that arm. The two uses both contribute to the shared-parameter derivative. The recipe selects Nemotron v1.1 training and validation

directories through Grain, with packed sequences of length 2048; the original file-level split manifest and deployed token-averaging implementation were not recovered. The training driver averages the available logged step losses in the last 1,000-step interval. The reported $0.7243/0.7045/0.7099$ values are training-log summaries from the experiment ledger, not a new held-out evaluation or token-weighted validation estimate. The exact logged samples underlying those summaries remain to be recovered.

Z OPEN EXPERIMENTS, NOT COMPLETED RESULTS

The original E4b follow-up protocol uses nine fixed checkpoints and the same three new text probes, measuring masked surgical excess and original/edited next-token loss. That checkpoint panel remains unexecuted because its stored weights are inaccessible. The public-model forwards in Appendix T have evaluated all three texts under their separately stated configuration and protocols; Appendix X adds matched ordinary edits and longer natural texts. Same-schedule independent-seed validation and a matched-singular-value alignment intervention remain later experiments. For that alignment intervention, matching weight singular values alone would not hold gate outputs or language-model damage fixed; it would also require perturbation and functional-damage controls.

The E4b control specification adds Frobenius-matched transpose edits, branch-perturbation baselines, both entry-23 and complete-block-output endpoints, activation profiles, targeted single-layer edits and spectral checks. Its status remains NOT_EXECUTED. The analogous public-model measurements are completed; they do not certify the archived trajectory's numerical directions or peak layers.