AN EXACT ADVECTION–DIFFUSION–REACTION REPRESENTATION OF TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated Transformer architecture invites a dynamical-systems view, but leaves open how computation is organized and changes through depth. Building on recent work connecting self-attention to a Laplacian-like difference operator, we systematically analyze how attention, normalization, residual connections, and nonlinear feed-forward transformations compose. We obtain an exact advection–diffusion–reaction (ADR) representation of sequential pre-norm Transformer blocks, retaining the feed-forward response at the attention-updated state. This representation provides a common geometric framework for quantifying component actions, deriving constraints on causal modifications, and examining task behavior through controlled interventions. In the evaluated pretrained decoders, advection generally grows relative to the complete update while reaction's relative magnitude declines, despite absolute reaction growth in most checkpoints. On common inputs, neighboring complete updates are more aligned than a fixed distant-block control, while their transport and reaction actions remain directionally distinct. Separating the identity baseline reveals that apparent diffusion persistence can coexist with changing symmetric transport. An attention-approximation intervention further shows that an opposing feed-forward response can amplify the block-output error. By preserving the complete nonlinear computation, exact ADR provides a quantitative basis for relating component geometry to depthwise behavior and the effects of modifying a block. This connection supplies constraints for architecture design and evaluation criteria for inference approximation.

1 INTRODUCTION

Large language models demonstrate broad capabilities in language understanding, reasoning, and code generation (Team, 2024). Yet their empirical success is not matched by a comparably complete account of the internal mechanisms that produce it (Geshkovski et al., 2023). For Transformers, an important part of that account is how information exchange between tokens combines with local feature transformations inside a block. Repeated architecture also raises a question about depth: how does the computation change as information passes through successive blocks?

Recent work represents self-attention as an exact connection walk, in which routing weights select source tokens and feature maps transform the information they carry (Lin et al., 2026). Its associated Laplacian-like difference operator provides a geometric starting point for analyzing a complete Transformer block. Such a block combines attention with normalization, residual addition, and a nonlinear feed-forward transformation of the attention-updated state (Xiong et al., 2020). Building on the earlier block-level ADR interpretation (Lin et al., 2026, Sec. 6.4), we analyze how these operations compose and how the resulting geometry describes computation through depth.

We derive an exact advection–diffusion–reaction (ADR) representation of the complete sequential pre-norm Transformer block. For each input, the attention connection splits into an identity term, skew-adjoint advection, and self-adjoint diffusion, defined using a fixed positive weighting of tokens. Reaction includes attention bias and full nonlinear feed-forward network (FFN), evaluated after attention update. The representation preserves computation exactly (Theorem 2). Applying it across blocks gives a discrete dynamical system whose component effects can be measured through depth.

This geometry guides both measurement and modification. We derive a coupling constraint on scalar advection and diffusion gains that preserves causal attention, so each token still depends

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.

only on earlier tokens and itself. We also show how a normalization rescaling with compensating weight changes can alter identity and diffusion magnitudes while preserving the block function. The symmetric transport obtained by separating the identity baseline, together with advection and reaction, remains unchanged under this rescaling (Section 4).

We use the decomposition to study two complementary aspects of depth. First, *how do component magnitudes change along the model's normal computation?* Across seven pretrained decoders, advection generally grows relative to the complete block update, while reaction's relative magnitude declines. Reaction nevertheless grows in absolute magnitude in most checkpoints. Depth therefore changes the balance of component effects without uniformly weakening reaction. Second, *how similarly do neighboring blocks act when given the same input?* Their complete updates are more aligned than those of a fixed distant-block comparison, while their advection, symmetric transport, and reaction components continue to change direction. In the five larger checkpoints, separating the identity baseline reveals that high diffusion alignment coexists with changing symmetric transport. Together, these comparisons reveal local organization across depth alongside continued changes within each update (Section 5).

Preserving the attention–FFN sequence also matters when modifying a model. Changing attention changes the FFN's input. A separate intervention that approximates the attention output projection shows that the FFN response can oppose the attention error yet amplify it at the block output. Local approximation quality must therefore be assessed after the FFN responds, with prediction quality evaluated separately. An exploratory task study in one checkpoint illustrates how ADR can also guide controlled component interventions (Appendix D.1).

Exact ADR connects these analyses to the computation the model performs. It makes geometric components measurable within the complete nonlinear update, relates their changes to depthwise behavior, and constrains how they can be modified. This gives architecture design and inference approximation a common basis for evaluating both component changes and their consequences for the complete block (Appendix E).

2 Native Block and Attention Connection

We begin with the block's computation and the attention connection of Lin et al. (2026), which we then extend to the complete block. Appendix A.1 gives the routing and transport derivations. Token states form the rows of $X_\ell \in \mathbb{R}^{n\times d}$, with destination token i reading source token j. At evaluation time, a sequential pre-norm block forms $Z_\ell = N_{1,\ell}(X_\ell)$ and computes

$$X_\ell^a = X_\ell + \Delta X_\ell^{\mathrm{attn}}, \qquad X_{\ell+1} = X_\ell^a + \mathrm{FFN}_\ell(N_{2,\ell}(X_\ell^a)). \tag{1}$$

Both normalizations and the FFN act on each token separately. Attention reads the normalized input Z_ℓ, the residual connection carries X_ℓ, and the FFN reads the updated state X_ℓ^a.

Realized connection. Fix a block input and let $a_h(i,j)$ be its attention weights for each of the H query heads. With the head's feature map $M_h = W_{V,h}W_{O,h}$ acting on row vectors, define

$$P_{ij} = \frac{1}{H}\sum_h a_h(i,j), \qquad \alpha_h(i,j) = \frac{a_h(i,j)}{\sum_r a_r(i,j)}, \qquad O_{ij} = \sum_h \alpha_h(i,j) M_h^\top. \tag{2}$$

On a zero-mass edge, any nonnegative completion of α_h summing to one gives the same action. Routing is P; calligraphic $\mathcal{A}$ below denotes advection. The column-field connection has kernel $T_{ij} = P_{ij}O_{ij}$ and acts as $(Tv)_i = \sum_j P_{ij}O_{ij}v_j$. For a row-stacked field V, its transposed action is

$$\mathcal{T}(V)_i = \sum_j P_{ij}V_j O_{ij}^\top = \frac{1}{H}\sum_{j,h} a_h(i,j)V_j M_h. \tag{3}$$

For a column-field operator C, $C(V)$ denotes its corresponding transposed action on row-stacked V, so $T(V) = \mathcal{T}(V)$. Thus T acts linearly on a field once its coefficients are fixed, while the coefficients themselves depend on the block input. We write $T_\ell[X]$ to make this dependence explicit when needed. The full attention Jacobian would also differentiate those coefficients.

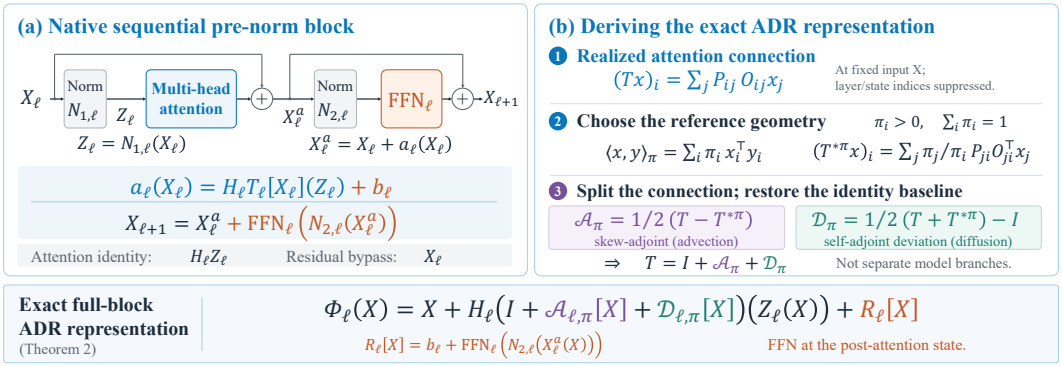


Figure 1: ADR preserves the native attention–FFN sequence. (a) Attention reads normalized Z_ℓ; the FFN reads the updated state X_ℓ^a. The identity action $H_\ell Z_\ell$ differs from the residual bypass X_ℓ. (b) A fixed positive token reference defines the connection's unique skew- and self-adjoint split. Reaction includes bias and the post-attention FFN. This is a construction schematic.

Native scale and bias. Row-stochasticity makes the value/output bias $b = \sum_h b_{V,h} W_{O,h} + b_O$ the same at every token. Restoring the native head sum yields

$$\Delta X_\ell^{\mathrm{attn}} = H_\ell \mathcal{T}_\ell[X_\ell](Z_\ell) + b_\ell. \tag{4}$$

The split acts on the normalized field and retains the head-count factor outside the connection.

3 Exact ADR Decomposition of Transformer Blocks

3.1 Geometric split of a realized connection

For a fixed block input and a column token field $x = (x_0, \ldots, x_{n-1})$, row stochasticity gives

$$((T - I)x)_i = \sum_j P_{ij}(O_{ij}x_j - x_i). \tag{5}$$

This transported difference generalizes the graph generator $P - I$. Its negative, $I - T$, is the corresponding Laplacian-like operator. The native attention walk remains T, so its identity baseline must be restored when reconstructing the block.

Choose positive token weights π, normalized to $\sum_i \pi_i = 1$. These weights define the reference geometry through $\langle x, y\rangle_\pi = \sum_i \pi_i x_i^\top y_i$. The adjoint of T in this geometry is

$$(T^{*\pi}x)_i = \sum_j \frac{\pi_j}{\pi_i} P_{ji} O_{ji}^\top x_j. \tag{6}$$

The reference specifies the geometry of the split; it does not change the forward connection.

Proposition 1 (Unique fixed-reference split)**.** *For a realized T and fixed positive π, exactly one split $T = I + \mathcal{A}_\pi + \mathcal{D}_\pi$ has $\mathcal{A}_\pi^{*\pi} = -\mathcal{A}_\pi$ and $\mathcal{D}_\pi^{*\pi} = \mathcal{D}_\pi$. It is*

$$\mathcal{S}_\pi = \tfrac{1}{2}(T + T^{*\pi}), \qquad \mathcal{A}_\pi = \tfrac{1}{2}(T - T^{*\pi}), \qquad \mathcal{D}_\pi = \mathcal{S}_\pi - I. \tag{7}$$

Proof. The displayed terms have the required adjoint properties and sum to T. Taking the adjoint of any such split and then adding and subtracting the two identities recovers these terms uniquely. □

Skew-adjointness gives $\langle x, \mathcal{A}_\pi x\rangle_\pi = 0$. Diffusion has nonpositive quadratic action only when $I - \mathcal{S}_\pi \succeq_\pi 0$; stationarity and orthogonal edge transport give one sufficient regime. That regime does not hold automatically for causal softmax decoders. Here *diffusion* names the self-adjoint coordinate relative to I, not a universal smoothing effect or a finite-residual-step energy law (Appendix A.2). Restoring the native scale in row convention,

$$\Delta X_\ell^{\mathrm{attn}}(X) - b_\ell = H_\ell(I + \mathcal{A}_{\ell,\pi}[X] + \mathcal{D}_{\ell,\pi}[X])(Z_\ell(X)). \tag{8}$$

3.2 THE EXACT SEQUENTIAL BLOCK

Set $\psi_\ell(U) = \mathrm{FFN}_\ell(N_{2,\ell}(U))$ and retain the native sequence

$$\begin{aligned} a_\ell(X) &= H_\ell \mathcal{T}_\ell[X](Z_\ell(X)) + b_\ell, \\ X^a_\ell(X) &= X + a_\ell(X), \qquad R_\ell[X] = b_\ell + \psi_\ell(X^a_\ell(X)). \end{aligned} \tag{9}$$

Theorem 2 (Exact sequential full-block ADR system)**.** *For every state X where the native sequential pre-norm block of Equation 1 is defined, every positive token reference π, and evaluation-time row-stochastic attention, its output is*

$$\boxed{\Phi_\ell(X) = X + H_\ell\big(I + \mathcal{A}_{\ell,\pi}[X] + \mathcal{D}_{\ell,\pi}[X]\big)(Z_\ell(X)) + R_\ell[X].} \tag{10}$$

The attention sublayer may have native affine value/output terms; normalizations and the FFN are tokenwise. The coordinates are those of Proposition 1, and $R_\ell[X]$ is evaluated at the post-attention state of Equation 9.

Proof. The native block returns $X + a_\ell(X) + \psi_\ell(X + a_\ell(X))$. Substitute Equation 8 for $a_\ell(X) - b_\ell$ and collect $b_\ell + \psi_\ell(X + a_\ell(X))$. □

The theorem reconstructs the block update. The identity action is $H_\ell Z_\ell$, not the residual bypass X, and reaction contains the attention bias once. The token-local ψ_ℓ acts on a state that already contains cross-token information. Query/key dependence remains in $T_\ell[X]$; no FFN linearization or continuum limit is used. The components are analytical, not separate model branches.

3.3 DEPTHWISE SYSTEM

Along the native trajectory the theorem gives

$$X_{\ell+1} - X_\ell = H_\ell(I + \mathcal{A}_{\ell,\pi}[X_\ell] + \mathcal{D}_{\ell,\pi}[X_\ell])(Z_\ell) + R_\ell[X_\ell] =: F_\ell(X_\ell). \tag{11}$$

This is a non-autonomous discrete system: later blocks have different parameters and encounter states produced by earlier blocks. Its component actions can be measured along that sequence of states. To compare the blocks themselves at a common input, we can instead give two blocks the same saved state and let each recompute its own normalization, routing, and reaction.

4 WHAT DOES ADR MAKE IDENTIFIABLE?

The exact block representation turns the geometric split into measurable component actions. At a fixed positive token reference, the split is unique (Proposition 1). We now examine which modifications preserve causal attention, which measurements remain unchanged under a rescaling that preserves the block function, and how to compare actions across blocks.

At a block input X_ℓ, the theorem gives four actions in one update:

$$\begin{aligned} F_\ell &= u_{I,\ell} + u_{A,\ell} + u_{D,\ell} + u_{R,\ell}, \\ u_{I,\ell} &= H_\ell Z_\ell, \quad u_{A,\ell} = H_\ell \mathcal{A}_{\ell,\pi} Z_\ell, \quad u_{D,\ell} = H_\ell \mathcal{D}_{\ell,\pi} Z_\ell, \quad u_{R,\ell} = b_\ell + \psi_\ell(X^a_\ell). \end{aligned} \tag{12}$$

Identity acts on normalized Z_ℓ, not the residual carrier. Advection is skew-adjoint and has zero same-field quadratic form; diffusion is self-adjoint relative to identity, with conditional sign (Section 3.1). Reaction is token-local at X^a_ℓ, whose input already contains transported information.

4.1 GEOMETRY YIELDS CAUSAL CONTROL FAMILIES

The geometric coordinates cannot generally be adjusted independently while preserving causal support. For a realized connection, equal scalar gains on advection and diffusion give $I + \gamma(\mathcal{A}_\pi + \mathcal{D}_\pi) = (1-\gamma)I + \gamma T$, a mixture of the identity and the forward connection. The exact restriction is:

Proposition 3 (Causal coupling of advection and diffusion)**.** *Let the realized connection T be block lower triangular and π positive. For scalar gains $\alpha, \delta \in \mathbb{R}$,*

$$I + \alpha\mathcal{A}_\pi + \delta\mathcal{D}_\pi = (1-\delta)I + \tfrac{\alpha+\delta}{2}T + \tfrac{\delta-\alpha}{2}T^{*\pi}. \tag{13}$$

If T has a nonzero strictly lower-triangular block, this operator is causal on arbitrary fields if and only if $\alpha = \delta$.

The adjoint term otherwise introduces future-token blocks. Separately causal skew- or self-adjoint changes are instead token-diagonal under this reference, offering within-token controls for the two geometric directions. The proposition fixes the operator coefficients and tests causality on arbitrary fields. An online implementation must also construct those coefficients causally, while normalization may restrict the fields the block actually encounters.

For $0 \le \gamma \le 1$, equal gains interpolate between the identity and the realized attention connection while preserving causal support. Gains outside this interval also preserve support, although the combination is no longer convex. A learned layerwise gain could be evaluated by recomputing the FFN after the modified attention update and comparing held-out prediction at matched model cost. Appendix A.3 provides the proof and the qualification for restricted normalized fields.

The same cancellation matters when interpreting native actions. With a uniform token reference and unchanged computations on the original prefix, appending a suffix can change the separate advection and diffusion actions at earlier tokens through their adjoint terms. The separate components therefore describe the completed sequence, even though these changes cancel in the causal attention update. Appendix A.3 derives this dependence on sequence length. A causal implementation must preserve the cancellation or use separately causal token-diagonal changes.

4.2 Baseline-resolved actions reveal what can be compared

The diffusion coordinate includes an identity baseline. Define the symmetric connection action $u_S = H\mathcal{S}_\pi Z$. Then

$$u_D = -u_I + u_S, \qquad F = u_S + u_A + u_R. \tag{14}$$

This regrouping separates symmetric transport from the opposing identity terms in u_I and u_D. Those terms arise from the decomposition rather than two separate operations executed by the model. The regrouping also makes a depthwise question precise: which action magnitudes change, and which directions remain similar when neighboring blocks act on the same state? A pointwise bound in Appendix A.3 quantifies when the identity baseline can dominate a diffusion angle. Its conditions apply to each action pair; average magnitude ratios along a trajectory do not establish them.

The coordinates depend on the checkpoint parameterization even with π fixed. To exhibit that dependence, multiply the first normalization's output scale and offset (where present) by $c > 0$, and divide the query, key, and value projection weights by c, retaining their biases. Queries, keys, values, and the native block output are unchanged, although $Z' = cZ$ and the connection expressed on that field is $T' = T/c$. At fixed reference the realized actions transform as

$$\begin{aligned} &u'_I = cu_I, \quad u'_D = u_D + (1-c)u_I, \\ &u'_A = u_A, \quad u'_S = u_S, \qquad\qquad u'_R = u_R, \quad F' = F. \end{aligned} \tag{15}$$

Thus an unchanged block function can have different identity and diffusion strengths. Symmetric transport, advection, and reaction provide comparisons stable under this specified rescaling. In particular, the advection and reaction trends studied below cannot be induced solely by shifting the identity baseline in this way. The guarantee is specific to this rescaling at fixed reference and does not extend to arbitrary feature-basis changes or token references. Appendix A.4 gives the construction and its assumptions.

We use two complementary measures of a realized action:

$$m_{c,\ell} = \frac{\|u_{c,\ell}\|_\pi}{\|F_\ell\|_\pi}, \qquad p_{c,\ell} = \frac{\langle u_{c,\ell}, F_\ell\rangle_\pi}{\|F_\ell\|_\pi^2} = m_{c,\ell}\cos_\pi(u_{c,\ell}, F_\ell). \tag{16}$$

The ratio m_c measures action size, while p_c accounts for its direction relative to the complete update. The regrouped projections satisfy $p_S + p_A + p_R = 1$ for every observation, but can be negative. They describe how the vectors add; causal or predictive importance requires an intervention.

4.3 TRANSPORT CHANGES CONTINUE THROUGH REACTION

The theorem also fixes where to evaluate an approximation. At one block input, replace its realized connection by $\mathcal{T}'$ while preserving the attention bias and native reaction map. Write $\delta_t = H(\mathcal{T}' - \mathcal{T})(Z_\ell)$ and $\delta_r = \psi_\ell(X_\ell^a + \delta_t) - \psi_\ell(X_\ell^a)$. For $\|\delta_t\| > 0$,

$$\delta_F = \delta_t + \delta_r, \qquad \frac{\|\delta_F\|^2}{\|\delta_t\|^2} - 1 = 2\frac{\langle \delta_t, \delta_r \rangle}{\|\delta_t\|^2} + \frac{\|\delta_r\|^2}{\|\delta_t\|^2}. \tag{17}$$

For $\|\delta_r\| > 0$, the response attenuates the block-output error exactly when $\cos(\delta_t, \delta_r) < -\|\delta_r\|/(2\|\delta_t\|)$. If $\delta_r = 0$, the block-output error equals the attention error. This identity follows from sequential composition and applies beyond ADR. It shows why an attention approximation must be evaluated after the FFN responds: attention error alone does not determine block-output error. For an executable approximation, local fidelity must be followed by tests of held-out prediction, operation count, memory, and end-to-end latency.

5 EXPERIMENTS

We study how component actions change as a model processes an input, how neighboring blocks compare on the same input, and how the FFN changes an attention approximation's error. The depth profiles and common-input comparisons use seven fixed pretrained decoders—GPT-2 Small, Medium, Large, and XL (Radford et al., 2019); Qwen3-1.7B and Qwen3-4B (Team, 2025); and Llama-3.2-3B (Team, 2024)—on 64 nonoverlapping 64-token windows per checkpoint from the WikiText-2 validation set (Merity et al., 2017). We weight tokens uniformly and retain each checkpoint's trained weights and feature basis. The separate reaction intervention uses checkpoints selected before held-out testing. Methods and complete contrasts are in Appendix C.3.

5.1 HOW DO ADR ACTIONS CHANGE THROUGH DEPTH?

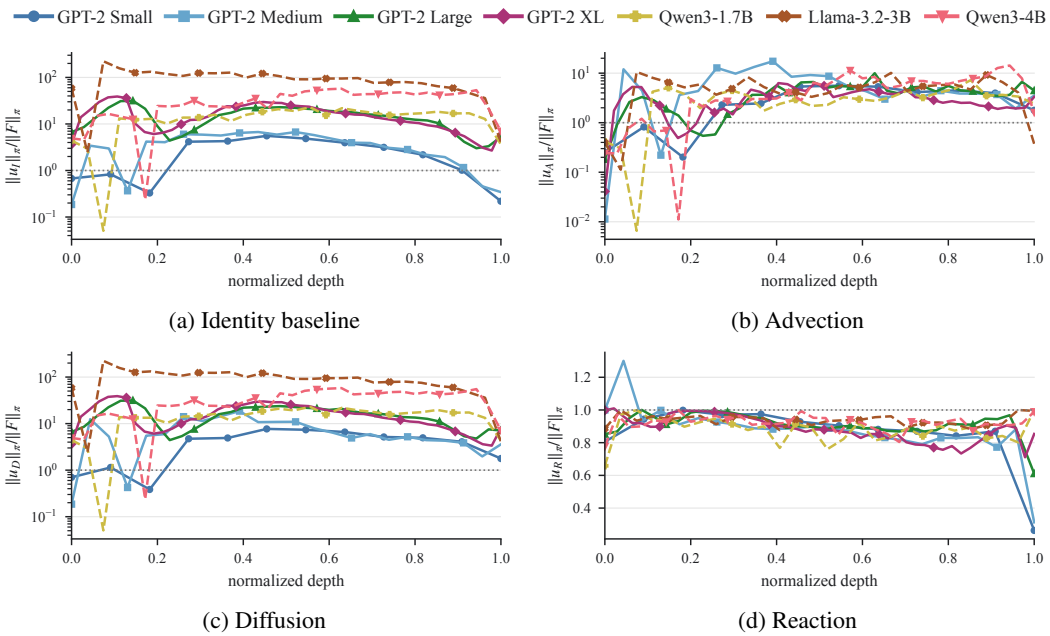
The norm ratio $m_c = \|u_c\|_\pi / \|F\|_\pi$ asks how large each realized action is relative to the complete update at the same state. Early and late bands are $[0.1, 0.4)$ and $[0.6, 0.9)$ of normalized depth; we average within a window before comparing windows. Advection's relative magnitude rises in six checkpoints and falls in GPT-2 Medium. Reaction's relative magnitude falls in all seven by 0.030–0.160, with simultaneous window-bootstrap intervals below zero. Identity and diffusion have different directions of change across checkpoints (Figure 2). The decomposition thus reveals distinct depth profiles within attention and in the reaction that follows it.

The relative reaction decline does not imply a common absolute decline: its descriptive early–late mean-log magnitude rises in five checkpoints and falls in GPT-2 Small and Large. In each of the five, the complete update grows faster than reaction. These changes in mean log magnitude describe a different statistic from the differences in mean m_R with adjusted intervals above. Signed projection provides a second distinction. Reaction's descriptive mean p_R falls in six checkpoints, but rises by 0.067 in Llama-3.2-3B as its relative norm falls by 0.030. Magnitude and alignment with the update therefore change differently; neither is a causal importance share (Appendix C.3).

The identity baseline explains the large identity and diffusion ratios. They can individually exceed the update norm by two orders of magnitude, yet $u_D = -u_I + u_S$. In the five larger checkpoints, native $\|u_S\|/\|u_I\|$ is only 0.054–0.249. The symmetric action measures the smaller transport left after the identity terms cancel. This cancellation occurs in the analytical decomposition, rather than between two costly model operations. The normalization rescaling in Section 4.2 changes identity and diffusion actions but preserves symmetric transport, advection, reaction, and the complete update at fixed reference.

5.2 WHICH ACTIONS REMAIN ORGANIZED ACROSS NEIGHBORING BLOCKS?

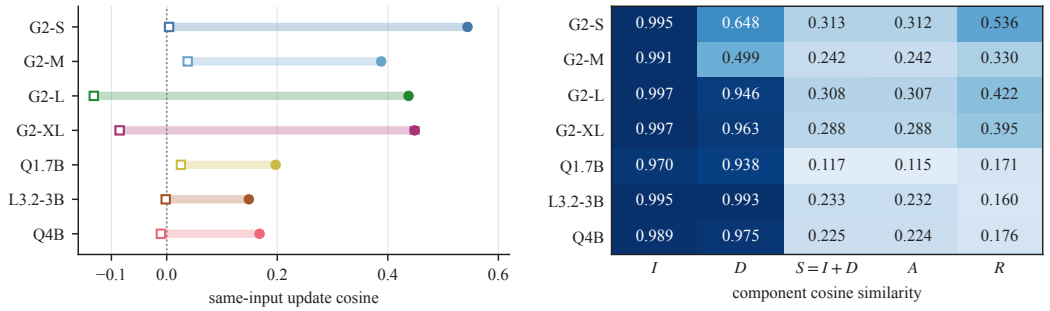
The depth profiles above compare blocks at different input states. To hold the input fixed, we save X_ℓ and run block ℓ and block $\ell + 1$ separately on that state. Each block recomputes its own normalization, routing, attention, and FFN. We compare update directions by their cosine. A value of one means the updates point in the same direction; zero means they are orthogonal. The adjacent complete-update cosine is 0.149–0.544 across the seven checkpoints: neighboring updates



(a) Identity baseline (b) Advection

(c) Diffusion (d) Reaction

Figure 2: From early to late depth, relative reaction falls in all seven checkpoints while advection rises in six. Curves show mean $\|u_c\|_\pi / \|F\|_\pi$ over 64 windows of 64 tokens per checkpoint, under uniform reference and native parameterization. These are descriptive norm ratios, not importance shares; adjusted contrasts are in Appendix C.3. Identity is HZ; reaction includes bias and the post-attention FFN. The first three axes are logarithmic; reaction is linear.



(a) Neighboring versus distant blocks (b) Directions of component actions

Figure 3: Neighboring updates align more than distant controls; diffusion alignment can mask changing symmetric transport. Left: connected mean cosines for neighboring blocks (filled circles) and fixed distant controls (open squares), with 95% window-bootstrap intervals. Right: mean neighboring-component cosines; $S = I + D$ removes diffusion's identity baseline. Row codes denote GPT-2 (G2), Qwen3 (Q), and Llama-3.2 (L). Each block recomputes all operations on the same input. Means average the middle 60% of layers, then 64 windows of 64 tokens per checkpoint.

remain directionally distinct. Yet each is more aligned than a fixed distant-block comparison (Figure 3). For GPT-2 Small and Qwen3-1.7B, three matched random initializations per architecture give a further check. The trained-minus-initialized cosine contrasts are $+0.548$ $[0.536, 0.562]$ and $+0.200$ $[0.197, 0.203]$, respectively; these 95% intervals resample both initialization seeds and text windows. These controls support a learned contribution to adjacent alignment in those two architectures. Across the seven checkpoints, neighboring updates are locally aligned on the saved inputs but remain distinct (Appendix C.3).

The component comparison reveals differences within these complete updates. Identity actions are closely aligned (cosine 0.970–0.997), while advection (0.115–0.312) and reaction (0.160–0.536)

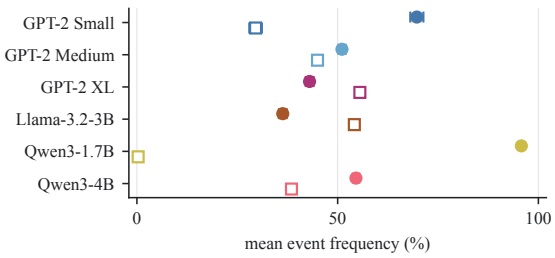


Figure 4: FFN opposition can still amplify attention-approximation error. Filled circles show opposition with block-error amplification; open squares show attenuation. Frequencies average eligible blocks within windows, then 128 held-out windows per checkpoint. Bars show descriptive 95% window-bootstrap intervals.

change direction substantially. Diffusion appears similarly persistent in the five larger checkpoints (0.938–0.993), but the baseline-resolved symmetric action has cosine only 0.117–0.308. Neighboring blocks therefore retain some directional order as complete updates while their geometric transport and reaction actions continue changing. The average $\|u_S\|/\|u_I\|$ along the original trajectory and the cosines on shared inputs summarize different measurements. Their averages do not establish the conditions for each pair of actions in the angle bound of Appendix A.3.

Low component cosines persist on a shared input, so differences between the inputs encountered during normal computation cannot alone explain them. The fixed distant-block comparison establishes relative alignment on those inputs, but distant blocks may receive states farther from their usual state distribution. The contrast can therefore reflect both the block maps and their compatibility with inputs from a particular depth. We have not repeated the comparison using the later block's usual input or several different distances between blocks. Matched initializations show that shared architecture alone does not account for the observed adjacent alignment in the two tested cases. The findings describe local alignment on the evaluated states; they do not establish similarity on arbitrary inputs, explain how training produces it, or show that computation can be reused.

5.3 Does reaction preserve a transport approximation's error?

To test the effect of sequential composition, a separate intervention approximates the attention output projection at fixed block input and measures $\kappa = \|\delta_F\|/\|\delta_t\|$, the ratio of block-output error to attention error after the same block's FFN responds.

Across six checkpoints selected before held-out evaluation and 128 new text windows each, the FFN response often points against the attention error yet the combined block error is *larger* in 36.3–95.8% of tested window–block pairs. Attenuation also occurs and varies by checkpoint. In all six, the log ratio of block-output to attention error is higher at the final block than at middle depths, with adjusted intervals above zero (Figure 4; Appendix C.5). The FFN therefore materially changes the local error that an attention-only evaluation would report. GPT-2 Large did not meet the prespecified selection criteria, so it has no held-out result. Adding the response to a predictor of held-out loss change also failed its prespecified improvement threshold. The intervention concerns an executable attention projection, not separate advection or diffusion changes or a demonstrated speedup.

The native and common-input studies show local depth order with changing action magnitudes and directions. Reaction's relative decline need not mean absolute weakening, and diffusion alignment need not imply persistent symmetric transport. The separate intervention shows that reaction can amplify an attention approximation's error despite opposing it. Native strength and perturbation response answer different questions, but both require the complete update for interpretation. Assigning semantic roles or removing computation requires task-level interventions.

6 Related Work

We study how the geometry of a native Transformer block relates to its behavior across depth and under modification. This question connects dynamical representations and layerwise explanations with work on computational reuse and responses to intervention.

Geometric dynamics and native block representations. Dynamical systems have guided stable residual architectures (Haber & Ruthotto, 2017; Ruthotto & Haber, 2020) and continuous-depth models (Chen et al., 2018), while interacting-particle accounts analyze Transformer attention dy-

namics and clustering (Geshkovski et al., 2023). Vector diffusion maps couple scalar affinities to orthogonal edge maps (Singer & Wu, 2012); neural sheaf diffusion learns maps between feature spaces (Bodnar et al., 2022), and bundle neural networks construct diffusion using vector-bundle geometry (Bamberger et al., 2025). ADR has also motivated learned graph architectures (Eliasof et al., 2024). For Transformers, Lin et al. (2026) represent attention as an exact connection walk and discuss an earlier block-level ADR interpretation. We extend this routing and feature-transport representation to the native sequential pre-norm block, retaining attention's scales and affine terms and the full nonlinear FFN at the post-attention state (Theorem 2).

From attention explanations to complete-block actions. Attention rollout and flow trace information across layers (Abnar & Zuidema, 2020). Norm-based analyses incorporate transformed values (Kobayashi et al., 2020), residual connections, and normalization (Kobayashi et al., 2021). Full-layer analyses include the FFN's effect on contextualization (Kobayashi et al., 2024), while DecompX propagates source-token contributions through complete layers to predictions (Modarressi et al., 2023). Our decomposition resolves the geometry within a complete update: skew and symmetric transport under a fixed reference, followed by nonlinear reaction at its native input. Resolving the identity baseline separates diffusion alignment from symmetric-transport alignment (Section 5).

Depth organization and computational reuse. Universal Transformers introduce recurrence across depth (Dehghani et al., 2019), ALBERT shares parameters across layers (Lan et al., 2020), and deep equilibrium models solve for fixed points of weight-tied transformations (Bai et al., 2019). Other analyses concern representation dynamics: pure attention without skip connections or MLPs exhibits rapid rank collapse (Dong et al., 2021), while spectral analyses identify conditions under which Transformer representations avoid oversmoothing (Dovonon et al., 2025). ShortGPT uses layer input/output similarity to guide pruning and evaluates the modified model (Men et al., 2025). Our common-input replay compares distinct pretrained blocks through their complete updates and component actions. It identifies local directional order on the evaluated states; reuse or pruning requires evaluation of the modified model.

Sequential response and model modification. The Hydra effect and subsequent self-repair studies show that downstream computations can compensate for ablations, with repair varying across inputs and components (McGrath et al., 2023; Rushing & Nanda, 2024). BRECQ and TesseraQ use block reconstruction for post-training quantization (Li et al., 2021; Li & Panda, 2024). Our intervention measures how the same block's FFN changes an attention approximation's hidden-state error. An opposing response can still amplify the error, so local fidelity must be assessed after reaction (Section 4.3). Prediction quality and execution cost remain separate evaluation endpoints.

7 Conclusion

We give an exact ADR representation of sequential pre-norm Transformer blocks, preserving attention's normalization, scaling, and bias and the full nonlinear FFN at the attention-updated state. Within a fixed token geometry, it constrains causal modifications and identifies measurements preserved by a specified normalization rescaling that leaves the block function unchanged. The evaluated decoders show changing component magnitudes through depth and local alignment between neighboring updates on shared inputs, while their transport and reaction directions remain distinct. Reaction's relative magnitude declines without a common absolute weakening. In the five larger checkpoints, high diffusion alignment coexists with changing symmetric transport once the identity baseline is separated. The approximation intervention further shows that an opposing FFN response can amplify attention error at the block output. Exact ADR connects these observations within the complete nonlinear computation, providing a quantitative account of depthwise organization on the evaluated states and concrete constraints for modifying a Transformer block.

AI USE STATEMENT

We used generative AI assistants for writing, literature retrieval and discovery, and research ideation and execution. Writing assistance included drafting and revising prose, improving clarity and organization, and refining captions and mathematical explanations. Literature assistance included identifying relevant work, developing search queries, summarizing retrieved sources, and comparing their assumptions and contributions with ours. Research assistance included developing and checking mathematical arguments, refining hypotheses and experimental comparisons, reviewing methodology, implementing and debugging analysis code, and interpreting experimental results.

AI-generated suggestions were treated as inputs to source checking and scientific review. Numerical claims are tied to retained model-run artifacts; reconstruction tests, independent recomputation, and targeted code checks support validation of the analyses. Synthetic algebra checks are kept distinct from empirical evidence about pretrained models. The authors retain responsibility for the final text, citations, mathematical claims, code, experimental results, and conclusions, including all material developed with AI assistance.

REPRODUCIBILITY STATEMENT

Sections 2–4 state the native computation and comparison objects; Appendices A.1–B give assumptions and proofs. Appendices C.1–C.5 specify populations, estimators, calibration and held-out evaluation splits, and validation criteria. Appendix D.2 gives the projection-approximation protocol and timing scope. Gaps in the experiment records are identified in Appendix C.1; execution records apply to the individual studies in which they were collected. Appendix D.1 specifies the constructed task, selection of prompts answered correctly before intervention, matched geometric interventions, and item-level uncertainty for the exploratory depth case study.

REFERENCES

Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.385. URL https://doi.org/10.18653/v1/2020.acl-main.385.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 4895–4901. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.298. URL https://doi.org/10.18653/v1/2023.emnlp-main.298.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 688–699, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/01386bd6d8e091c2ab4c7c7de644d37b-Abstract.html.

Jacob Bamberger, Federico Barbero, Xiaowen Dong, and Michael M. Bronstein. Bundle neural network for message diffusion on graphs. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=scI9307PLG.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Lió, and Michael M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December*

9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/75c45fca2aa416ada062b26cc4fb7641-Abstract-Conference.html.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6572–6583, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 2021. URL http://proceedings.mlr.press/v139/dong21a.html.

Gbètondji J.-S. Dovonon, Michael M. Bronstein, and Matt J. Kusner. Setting the record straight on transformer oversmoothing. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=HHI6qWLFF1.

Moshe Eliasof, Eldad Haber, and Eran Treister. Feature transportation improves graph neural networks. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 11874–11882. AAAI Press, 2024. doi: 10.1609/AAAI.V38I11.29073. URL https://doi.org/10.1609/aaai.v38i11.29073.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *CoRR*, abs/2312.10794, 2023. doi: 10.48550/ARXIV.2312.10794. URL https://doi.org/10.48550/arXiv.2312.10794.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *CoRR*, abs/1705.03341, 2017. URL http://arxiv.org/abs/1705.03341.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7057–7075. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.574. URL https://doi.org/10.18653/v1/2020.emnlp-main.574.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating residual and normalization layers into analysis of masked language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 4547–4568. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.373. URL https://doi.org/10.18653/v1/2021.emnlp-main.373.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Analyzing feed-forward blocks in transformers through the lens of attention maps. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mYWsyTuiRp.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Yuhang Li and Priyadarshini Panda. Tesseraq: Ultra low-bit LLM post-training quantization with block reconstruction. *CoRR*, abs/2410.19103, 2024. doi: 10.48550/ARXIV.2410.19103. URL https://doi.org/10.48550/arXiv.2410.19103.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: pushing the limit of post-training quantization by block reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Binbin Lin, Wei Chen, Yalun Li, Wenxiao Wang, Jieping Ye, and Xiaofei He. From self-attention to connection laplacian: A unified operator view of transformers, 2026. URL https://doi.org/10.48550/arXiv.2607.10677.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023. URL https://doi.org/10.48550/arXiv.2307.15771.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 20192–20204. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.1035. URL https://doi.org/10.18653/v1/2025.findings-acl.1035.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. Decompx: Explaining transformers decisions by propagating token decomposition. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 2649–2664. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.149. URL https://doi.org/10.18653/v1/2023.acl-long.149.

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

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations. *J. Math. Imaging Vis.*, 62(3):352–364, 2020. doi: 10.1007/S10851-019-00903-1. URL https://doi.org/10.1007/s10851-019-00903-1.

Amit Singer and Hau-Tieng Wu. Vector diffusion maps and the connection laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. doi: 10.1002/cpa.21395.

Llama Team. The llama 3 herd of models, 2024. URL https://doi.org/10.48550/arXiv.2407.21783.

Qwen Team. Qwen3 technical report, 2025. URL https://doi.org/10.48550/arXiv.2505.09388.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: truncation-aware singular value decomposition for large language model compression. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LNYIUouhdt.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020. URL http://proceedings.mlr.press/v119/xiong20b.html.

Hao Yu and Jianxin Wu. Compressing transformers: Features are low-rank, but weights are not! In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 11007–11015. AAAI Press, 2023. doi: 10.1609/AAAI.V37I9.26304. URL https://doi.org/10.1609/aaai.v37i9.26304.

A Proofs and Algebraic Identities

A.1 Attention reconstruction and comparison bounds

All maps in the following norm bounds act on column token vectors; row-stacked model fields use the transpose convention of Section 2. For reconstruction, expand the native head sum:

$$\Delta X_i^{\mathrm{attn}} = \sum_{h,j} a_h(i,j)(Z_j W_{V,h} + b_{V,h}) W_{O,h} + b_O.$$

Each head has $\sum_j a_h(i,j) = 1$, so its value-bias contribution is $b_{V,h} W_{O,h}$. Also $HP_{ij}\alpha_h(i,j) = a_h(i,j)$ on nonzero edges, while both sides vanish on zero-routing edges. Substitution yields Equation 4. No invertibility or orthogonality is used.

For a nonzero native attention increment, we verify reconstruction using

$$r_\ell = \frac{\|\Delta X_\ell^{\mathrm{attn}} - H_\ell \mathcal{T}_\ell[X_\ell](Z_\ell) - b_\ell\|_F}{\|\Delta X_\ell^{\mathrm{attn}}\|_F} < 10^{-6}.$$

This checks that the extracted quantities reproduce the native attention increment; the denominator is its full affine output, not only the feature-dependent transport.

Grouped-query capacity. If query heads share value maps across G groups $\mathcal{H}_g$ (Ainslie et al., 2023), then

$$O_{ij}^\top = \sum_{g=1}^{G} W_{V,g} \left(\sum_{h \in \mathcal{H}_g} \alpha_h(i,j) W_{O,h} \right), \qquad \mathrm{rank}(O_{ij}) \le \min(d, G d_v),$$

where d_v is the value dimension per group. Each summand has rank at most d_v, while the sum can have larger rank. Ordinary multi-head attention has $G = H$; the normalized ceiling G/H additionally assumes $d = H d_v$; otherwise normalize the bound using the actual dimensions. The bound describes architectural capacity, not the effective rank or isometry of a realized edge map.

For row-stochastic P, the eigenvector $\mathbf{1}$ gives spectral radius at least one, while $\|P\|_\infty = 1$ gives spectral radius at most one. The induced-norm bound gives $1 \le \|P\|_2 \le \sqrt{\|P\|_1} \le \sqrt{n}$. A normal P would have spectral norm equal to spectral radius, so norm above one also implies non-normality. This does not make non-normality sufficient for any specified amount of expansion.

Set $S_{ij} = \|O_{ij}\|_2$ and $u_j = \|v_j\|_2$. The triangle inequality gives $\|(Tv)_i\|_2 \le \sum_j P_{ij} S_{ij} u_j$, hence $\|Tv\|_2 \le \|P \odot S\|_2 \|u\|_2 = \|P \odot S\|_2 \|v\|_2$. If every map is the same sQ with $Q^\top Q = I$, then $T = P \otimes (sQ)$, so its norm is exactly $s\|P\|_2$ for $s \ge 0$. Applying the same triangle inequality only to omitted sources proves the route-truncation bound. Explicitly, for a retained source set S_i and column field z,

$$\left\| (Tz)_i - \sum_{j \in S_i} P_{ij} O_{ij} z_j \right\|_2 \le \sum_{j \notin S_i} P_{ij} \|O_{ij}\|_2 \|z_j\|_2.$$

This bound deletes omitted messages without renormalizing the retained probabilities. Renormalized truncation also changes retained-message weights and requires an additional error term. Small omitted attention mass controls action error only together with transport gain and source magnitude. Scaled-orthogonal maps make each transport norm an explicit scalar gain; they do not establish held-out prediction tolerance or runtime savings.

Calibration-optimal output projection. Let Y stack all affine-free calibration attention outputs, and let $Q \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ satisfy $Q^\top Q = I$. The implemented replacement produces $YQQ^\top$ on those inputs. Its reconstruction objective is

$$\|Y - YQQ^\top\|_F^2 = \mathrm{tr}(Y^\top Y) - \mathrm{tr}(Q^\top Y^\top Y Q).$$

The second trace is maximized by the leading r eigenvectors of $Y^\top Y$: in its eigenbasis the projector assigns weights between zero and one that sum to r, so the largest eigenvalues receive unit weight

at the optimum. The minimum error is the sum of the remaining eigenvalues. In fact, if C stacks the affine-corrected projection inputs and $Y = CW$, then

$$\min_{\operatorname{rank}(\widetilde{W})\le r} \|C(W - \widetilde{W})\|_F^2 = \sum_{k>r} \sigma_k(Y)^2.$$

Every $C\widetilde{W}$ has rank at most r, so truncated-SVD output error is a lower bound. The implemented $\widetilde{W} = WQQ^\top$ attains it. This remains true when $C^\top C$ is singular. When it is positive definite, putting $L = (C^\top C)^{1/2}$ gives an equivalent minimizer $L^{-1}(LW)_r$, up to nonuniqueness at tied singular values. This is the activation-weighted low-rank objective used in feature-aware compression and whitening-based SVD (Yu & Wu, 2023; Wang et al., 2025). It optimizes a fixed linear map on calibration inputs, not arbitrary low-rank model changes or downstream prediction loss.

Causal routing and non-expansion.

Proposition 4 (Causal stochastic routing norm)**.** *Let $P \in \mathbb{R}^{n\times n}$ be nonnegative, row-stochastic, and lower triangular. Then $\|P\|_2 \le 1$ if and only if $P = I$. On this support, double stochasticity is also equivalent to $P = I$.*

Proof. Let $v = \mathbf{1}/\sqrt{n}$. Row stochasticity gives $Pv = v$. If $\|P\|_2 \le 1$, then $\|P^\top v\|_2 \le 1$ while $\langle P^\top v, v\rangle = \langle v, Pv\rangle = 1$. Equality in Cauchy–Schwarz gives $P^\top v = v$, so column sums also equal one. Since the first row of a lower-triangular stochastic matrix is $(1, 0, \dots, 0)$, nonnegativity and the first column sum force all entries below P_{00} to vanish. Repeating on the remaining principal submatrix yields $P = I$. The converse is immediate. The same induction proves the double-stochasticity statement directly. □

This is a Euclidean routing statement, not a lower bound on the gain of the joint connection on a native field. For example, a common scaled-orthogonal edge map gives $\|T\| = s\|P\|$, which can be below one when s is small. Without the causal support restriction, $P = \mathbf{1}\mathbf{1}^\top/n$ is nontrivial doubly stochastic routing with norm one.

A.2 Adjoint geometry and the dissipativity condition

The split and full-block proofs appear with Proposition 1 and Theorem 2. Here we verify the geometric condition used in Section 3.1. On column fields, the inner product gives

$$\langle x, Ty\rangle_\pi = \sum_{i,j} \pi_i P_{ij} x_i^\top O_{ij} y_j = \sum_j \pi_j \left(\sum_i \frac{\pi_i}{\pi_j} P_{ij} O_{ij}^\top x_i \right)^\top y_j,$$

which proves Equation 6. For every positive π, skew-adjointness then implies $\langle x, \mathcal{A}_\pi x\rangle_\pi = -\langle \mathcal{A}_\pi x, x\rangle_\pi = 0$ over the reals. For a finite step on the same field,

$$\|x + \epsilon \mathcal{A}_\pi x\|_\pi^2 = \|x\|_\pi^2 + \epsilon^2 \|\mathcal{A}_\pi x\|_\pi^2.$$

Thus the skew component has zero quadratic action on the same field without stationarity or orthogonal edge maps, while a finite residual addition can increase the norm.

Since $\mathcal{D}_\pi = \mathcal{S}_\pi - I$ is self-adjoint, $\langle x, \mathcal{D}_\pi x\rangle_\pi \le 0$ for every x holds exactly when $I - \mathcal{S}_\pi$ is positive semidefinite in this inner product. One sufficient condition is row-stochastic P with positive stationary reference π and orthogonal O_{ij} on every nonzero edge. Jensen's inequality gives

$$\|Tx\|_\pi^2 \le \sum_{i,j} \pi_i P_{ij} \|O_{ij} x_j\|_2^2 = \sum_j \pi_j \|x_j\|_2^2 = \|x\|_\pi^2.$$

Consequently $\langle x, \mathcal{S}_\pi x\rangle_\pi = \langle x, Tx\rangle_\pi \le \|x\|_\pi^2$ by Cauchy–Schwarz. This sufficient regime needs neither reversible routing nor $O_{ji} = O_{ij}^\top$. It is not necessary for a particular connection to have nonpositive quadratic action. Conversely, with $P = I$ and $O_{ii} = 2I_d$, one obtains $T = 2I$ and $\mathcal{D}_\pi = I$: self-adjointness alone does not enforce the sign. These are generator-level statements, not a smoothing law for the native residual update.

The geometric coordinates also depend on the token reference. For two positive references π, ν and the same realized T,

$$\mathcal{A}_\nu - \mathcal{A}_\pi = -\tfrac{1}{2}(T^{*\nu} - T^{*\pi}), \qquad \mathcal{D}_\nu - \mathcal{D}_\pi = +\tfrac{1}{2}(T^{*\nu} - T^{*\pi}). \tag{18}$$

Their sum and the native block remain unchanged. Uniqueness in Proposition 1 fixes the connection, positive reference, identity baseline, and head-average convention. For row-stacked fields the inner product is $\langle U, V\rangle_\pi = \sum_i \pi_i U_i V_i^\top$. Our primary depthwise reference is $\pi_i = 1/n$, and the one-step received mass $\nu_j = n^{-1}\sum_i P_{ij}$ supplies a sensitivity analysis. A layer-adaptive reference mixes changes of geometry into changes of individual coordinates, so the reference must be held fixed for an operator-only depth comparison.

For full causal softmax with finite allowed logits, $P_{ij} > 0$ for $j \le i$ and zero otherwise. Its only stationary probability is concentrated at position zero. Indeed, the last equation of $P^\top \pi = \pi$ gives $(1 - P_{n-1,n-1})\pi_{n-1} = 0$; since $P_{ii} < 1$ for every $i > 0$, descending induction eliminates all positions except zero. This stationary measure is degenerate over the sequence. Uniform and received-mass references are positive for this routing and define non-degenerate adjoints without assuming stationarity.

A.3 Causal support and coordinate cancellation

Proof of Proposition 3. The adjoint formula makes $T^{*\pi}$ block upper triangular. On $i < j$, both T_{ij} and I_{ij} vanish. Substituting Equation 7 gives Equation 13, whose strictly upper blocks vanish exactly when $\alpha = \delta$ if T has cross-token support. □

The necessity statement concerns the realized operator on arbitrary fields, not every nonlinear composition on a restricted normalization image. For example, let $e = \mathbf{1}/\sqrt{d}$, choose a nonzero $v \perp e$, and take $O_{ij} = ev^\top$ with nontrivial causal routing. For centered LayerNorm with unit gain and zero offset, $e^\top Z_i = 0$ at every token, so $T^{*\pi} Z = 0$. On these fields the unequal-gain map reduces to $(1-\delta)Z + (\alpha+\delta)TZ/2$, which is causal despite the modified frozen operator's nonzero upper blocks. Equal gains remain sufficient to preserve operator support without such restrictions.

For block lower-triangular T, the weighted adjoint is upper triangular. On every forbidden block $i < j$, the ADR coordinates satisfy

$$(\mathcal{A}_\pi)_{ij} = -\tfrac{1}{2}(T^{*\pi})_{ij}, \qquad (\mathcal{D}_\pi)_{ij} = +\tfrac{1}{2}(T^{*\pi})_{ij}. \tag{19}$$

These blocks cancel in $\mathcal{A}_\pi + \mathcal{D}_\pi = T - I$. The restriction in Proposition 3 also follows from a general property of causal support. Let C be block lower triangular and self-adjoint or skew-adjoint under a positive diagonal token metric. For $i > j$, its adjoint relation determines C_{ij} from $(\pi_j/\pi_i)C_{ji}^\top$, up to sign. Since $C_{ji} = 0$, every strictly lower block also vanishes. Thus a separately causal self-adjoint or skew-adjoint coordinate has only token-diagonal blocks. Such blocks can still rotate features.

For a two-token scalar illustration under uniform π, take

$$T = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad \mathcal{A} = \begin{pmatrix} 0 & -1/2 \\ 1/2 & 0 \end{pmatrix}, \qquad \mathcal{D} = \begin{pmatrix} 0 & 1/2 \\ 1/2 & -1 \end{pmatrix}.$$

Their sum satisfies $I + \mathcal{A} + \mathcal{D} = T$. Here $\mathcal{D}$ has eigenvalues $(-1 \pm \sqrt{2})/2$, so even this scalar causal routing example is not dissipative under uniform reference. Unequal scalar gains leave the forbidden upper entry $(\delta - \alpha)/2$, while equal gains cancel it. Another limiting check is zero transport: $T = 0$ implies $\mathcal{A} = 0$ and $\mathcal{D} = -I$. Large identity and diffusion coordinate actions can therefore cancel even when no forward attention action remains. These examples illustrate the algebra, not measured Transformer behavior or redundant native matrix operations.

Quantifying identity-baseline persistence.

Proposition 5 (Baseline-controlled diffusion angles)**.** *For two action triples in the same fixed geometry satisfying $u_{D,j} = u_{S,j} - u_{I,j}$ and $u_{I,j} \neq 0$, let $\epsilon_j = \|u_{S,j}\|_\pi / \|u_{I,j}\|_\pi < 1$ for $j = 1, 2$. If θ_D, θ_I are the respective between-action angles, then*

$$|\theta_D - \theta_I| \le \arcsin \epsilon_1 + \arcsin \epsilon_2. \tag{20}$$

For a nonzero identity action, write $a = -u_I$ and $b = u_D = a + u_S$. If $\epsilon = \|u_S\|_\pi / \|u_I\|_\pi < 1$, then $\langle a, b \rangle_\pi \geq (1 - \epsilon)\|a\|_\pi^2 > 0$. The angle θ between a and b is therefore acute. The distance from a to the line spanned by b is $\|a\|_\pi \sin\theta$ and cannot exceed $\|a - b\|_\pi = \|u_S\|_\pi$. Hence $\theta \leq \arcsin\epsilon$. Apply the angular triangle inequality to two such pairs in the same inner product to obtain Equation 20. This bound holds per field pair; an average ratio below one does not establish its hypotheses at every observation. Applying the bound empirically would therefore require checking each pair separately.

Prefix extension and the generation frontier. Use stacked column fields and uniform reference. Suppose appending a suffix leaves prefix positions, masks, and all native prefix computations unchanged. Partition the realized connection and normalized field as

$$T_n = \begin{pmatrix} T_m & 0 \\ B & C \end{pmatrix}, \qquad z = (z_p, z_s).$$

The prefix rows of $T_n z$ equal $T_m z_p$, whereas those of $T_n^\top z$ equal $T_m^\top z_p + B^\top z_s$. Thus the native prefix output is unchanged, but its separate coordinate actions at native scale satisfy

$$(u_A^{(n)})_p - u_A^{(m)} = -\frac{H}{2} B^\top z_s, \qquad (u_D^{(n)})_p - u_D^{(m)} = +\frac{H}{2} B^\top z_s. \tag{21}$$

Uniform references at lengths m and n have the same within-prefix weight ratios, which is the required consistency for this calculation. A length-dependent received-mass reference need not have that property. At the final token $i = n - 1$, the strictly cross-token adjoint row vanishes:

$$(u_{A,\mathrm{cross}})_i = (u_{D,\mathrm{cross}})_i = \frac{H}{2} (T_{\mathrm{cross}} z)_i.$$

Only the cross-token actions are equated; diagonal feature transformations can differ. Thus a token's ADR coordinates can depend on later tokens even though its native output cannot. Combining rows from different prefixes need not give a globally skew- or self-adjoint operator.

A.4 Native block correspondence

Table 1 collects the native operations in Theorem 2. Normalization, affine terms, and sequential evaluation are retained explicitly.

Table 1: Native operations in the exact full-block ADR map.

Native operation	Role in the exact map
Residual additions	Preserve X across attention and add $\psi_\ell(X_\ell^a)$ after the attention substep.
Pre-attention normalization	Supplies the actual field $Z_\ell = N_{1,\ell}(X)$, including learned scale and any offset.
Queries, keys, positions, mask, score scaling	Construct the state-conditioned routing probabilities entering both $P_{\ell,ij}$ and the head mixture in $O_{\ell,ij}$.
Value/output projections	Supply the feature maps $M_{\ell,h}$ inside $T_\ell = P_\ell \circ O_\ell$; non-isometric transport stays in this connection.
Native head sum	Supplies the known multiplier H_ℓ outside $I + \mathcal{A}_{\ell,\pi} + \mathcal{D}_{\ell,\pi}$.
Attention affine terms	Give the broadcast addition b_ℓ, entering X_ℓ^a and included once in $R_\ell[X]$.
Post-attention normalization and FFN	Give $\psi_\ell = \mathrm{FFN}_\ell \circ N_{2,\ell}$ at the updated state, including its native gating, nonlinearities, and biases.

The residual bypass X, the identity action HZ, and token-diagonal attention $HT_{\mathrm{diag}}(Z)$ have different roles and are generally unequal; T_{diag} retains the blocks $P_{ii}O_{ii}$ only. The head multiplier H is distinct from attention-score scaling. Decomposing the native operator HT around a unit identity would give $H\mathcal{A}_\pi$ and $H\mathcal{D}_\pi + (H-1)I$. We retain H outside the normalized connection split, with $\mathcal{S}_\pi = I + \mathcal{D}_\pi$ denoting the symmetric connection action rather than diffusion.

Attention bias b_ℓ is added once through $R_\ell[X]$ and also conditions the FFN input through $X_\ell^a(X)$. Normalization offsets are already included in $Z_\ell(X)$. Query/key dependence makes the transport state-conditioned, while anisotropic or expansive edge maps remain part of its action. The algebraic theorem applies to every input in the native domain; reconstruction residuals separately certify that extracted model quantities implement it numerically.

Post-norm composition. For the conventional sequential post-norm ordering, denote the native normalizations after the two residual additions by N_{attn} and N_{ffn}. Attention reads X directly. At evaluation time its connection still reconstructs the affine attention increment as

$$a(X) = H\mathcal{T}[X](X) + b.$$

The block is the exact nested composition

$$\begin{aligned} U &= N_{\mathrm{attn}}(X + H(I + \mathcal{A}_\pi[X] + \mathcal{D}_\pi[X])(X) + b)\,, \\ \Phi^{\mathrm{post}}(X) &= N_{\mathrm{ffn}}(U + \mathrm{FFN}(U))\,. \end{aligned} \tag{22}$$

This follows by substituting the same fixed-reference connection identity into the native sequence. Both normalizations remain in their original positions, so Equation 22 is not the additive pre-norm formula of Theorem 2 with a relabeled reaction. In particular, the terminal normalization acts on the combined residual and FFN state. Its correction must not be silently absorbed into $R = b + \psi(X^a)$, which is defined for the pre-norm theorem. The equation records the architectural correspondence; it adds no post-norm empirical validation or generality to the decoder studies.

Function-preserving normalization rescaling. Suppose the first normalization has trainable output scale and, where present, offset, and its output is consumed only by the attention Q/K/V projections. Multiply that scale and offset by a scalar $c > 0$, and divide each consuming projection weight by c, preserving all projection biases. Then $Z' = cZ$ and every affine query, key, and value is unchanged. Positional operations and any subsequent Q/K normalization therefore receive the same inputs; routing, attention output, post-attention state, FFN output, and the complete block map are unchanged. Shared GQA projections are rescaled once per parameter, preserving their tying.

The value–output maps satisfy $M'_h = M_h/c$, so $T' = T/c$. At the same positive reference, $(T')^{*\pi} = T^{*\pi}/c$; substitution yields Equation 15. The identity baseline and diffusion can therefore change without any change to the native function. Their uniqueness is conditional on the parameterization as well as the reference and head-average convention. The same distinction applies to the sign condition: $\mathcal{D}'_\pi = \mathcal{S}_\pi/c - I$ is nonpositive precisely when $c \geq \lambda_{\max}(\mathcal{S}_\pi)$, for $c > 0$ at the fixed realization. Thus sufficiently large compensated scaling can enforce coordinate dissipativity without changing the native block. This equivalence holds in exact arithmetic for this scalar rescaling. It is untested as a checkpoint intervention and does not extend to general channel rescalings.

B Finite-Step and Continuous-Depth Boundaries

The primary object of Theorem 2 is a finite native block. To state a continuous-depth interpretation precisely, fix its maps $a_\ell(X)$ and ψ_ℓ from Equation 9 and define the residual-scaled family

$$X_\ell^{a,h}(X) = X + h\,a_\ell(X), \qquad \Phi_\ell^h(X) = X + h\,a_\ell(X) + h\,\psi_\ell(X_\ell^{a,h}(X)). \tag{23}$$

The member $h = 1$ is exactly the native sequential block and $h = 0$ is the identity. If ψ_ℓ is locally Lipschitz along the relevant segment and $a_\ell(X)$ is bounded there, then

$$\Phi_\ell^h(X) = X + h\big(a_\ell(X) + \psi_\ell(X)\big) + O(h^2).$$

Thus the post-attention evaluation of the reaction affects the finite step but enters only at second order in this scaled expansion.

A fixed-horizon limit needs more than reducing h for one checkpoint. It requires a hypothetical depth-refined model family with $L_h h = \tau$, an interpolation of the layer-dependent maps at $t_\ell = \ell h$, consistent convergence to a limiting field, and uniform regularity and stability bounds on the relevant trajectories. No such family is constructed here, and neither the exact identity nor the observed finite-depth profiles place pretrained checkpoints in a small-step regime. The token graph also remains discrete; a local spatial PDE would require additional assumptions beyond a continuous-depth limit.

C Experimental Definitions and Validation

C.1 Attention populations and estimators

The attention studies use fixed trained checkpoints and nonoverlapping text windows. Window resampling quantifies conditional text variation, not uncertainty over independently trained models.

No tokens, heads, edges, or layers are counted as independent replicates. The primary corpus is WikiText-2 validation at context 64. The matched initialization controls use three parameter draws per architecture. Nonoverlap does not imply independence across source documents: the intervals condition on the token-window sampling unit, without an article-clustered sensitivity analysis. Tokenizer-specific windows are not paired texts across architectures.

For the full-depth spectral, calibrated-alignment, routing-gain, and single-projection studies, WikiText-2 raw validation records are stripped; records shorter than ten characters are skipped. The native tokenizer adds no BOS/EOS wrapper. Records are joined with an EOS separator (token zero if no EOS is defined), divided into disjoint length-64 windows, and their start positions are shuffled with seed 1234 before selection. Positions restart within each window. This protocol samples a token stream rather than intact documents. The offline snapshots' upstream revision hashes and a document-membership map were not retained for every study; file hashes cannot recover the missing revision and document-membership information.

The joint kernel/action display uses 128 windows per checkpoint. Its cosine panel uses a separate 64-window directional study; plotted layers are restricted to the intersection of their middle-band coverage. The 64-window action replication uses a different seed, includes token zero, and uses a slightly narrower integer depth band than the 128-window study. Its 10,000-draw, 95% window-bootstrap intervals quantify replication uncertainty, not uncertainty on the displayed 128-window estimates.

For a retained common edge population $\mathcal{E}$, kernel drift is

$$D_K = \left(\frac{\sum_{(i,j)\in\mathcal{E}} \|K_{\ell+1,ij} - K_{\ell,ij}\|_F^2}{\sum_{(i,j)\in\mathcal{E}} \|K_{\ell,ij}\|_F^2} \right)^{1/2}.$$

Same-input action drift is $D_{TZ} = \|\mathcal{T}_{\ell+1}[X_{\ell+1}](Z_\ell) - \mathcal{T}_\ell[X_\ell](Z_\ell)\|_F / (\|\mathcal{T}_\ell[X_\ell](Z_\ell)\|_F + \varepsilon)$, with a numerical denominator guard. Each connection keeps routing realized at its own trajectory state. This is a common-field comparison, not full-block replay at X_ℓ. The adjacent-action norm ratio is $\|\mathcal{T}_{\ell+1}(Z_\ell)\|_F / \|\mathcal{T}_\ell(Z_\ell)\|_F$; it differs from the native-field gain defined below. Scale-normalized drift replaces each head map M_h by $M_h/\sqrt{\mu_h}$, with $\mu_h = \|M_h\|_F^2/d$, without changing routing. The strict full-width anisotropy is $\|O^\top O - \mu_O I\|_F / (\mu_O \sqrt{d})$ with $\mu_O = \|O\|_F^2/d$.

For routing, normalized depth is $x = \ell/(L-1)$. Compute late $[0.6, 0.9)$ minus early $[0.1, 0.4)$ means within each window before averaging 64 windows. Row entropy is $-\sum_j P_{ij} \log P_{ij} / \log(i+1)$ for $i > 0$; the forced first query is omitted. The squared self fraction is $\sum_i P_{ii}^2 / \|P\|_F^2$ and non-normality is $\|P^\top P - P P^\top\|_F / \|P\|_F^2$. Native-field gains are $\|PZ\|_F / \|Z\|_F$ and $\|\mathcal{T}(Z)\|_F / \|Z\|_F$. Eight families of measurements use 10,000 window-bootstrap draws with Bonferroni 99.375% intervals. Trained-minus-initialization comparisons are paired on windows and correct 48 metric/checkpoint/seed contrasts. Routing masks, sums, spectral bounds, and the causal skew/self identity are checked throughout.

For active transport, define $r_{99} = \min\{r : \sum_{k\le r} \sigma_k^2 \ge 0.99 \sum_k \sigma_k^2\}$. At this fixed rank, the primary defect is $\delta_{99} = \min_s \|\Sigma_{r_{99}} - sI\|_F / \|\Sigma_{r_{99}}\|_F$. The optimum is the mean retained singular value. If c is their coefficient of variation, $\delta_{99} = c/\sqrt{1+c^2}$; these are equivalent geometry summaries, not independent corroborations. Full-depth spectra use raw float64 factors and consider all adjacent pairs in the first 32 of 64 windows. For each window and pair, score eligible edges by $\min(P_{\ell,ij}, P_{\ell+1,ij})$ and keep the ten largest positive scores. Eligibility is $1 \le j \le i < n$: source zero and query zero are excluded, but self edges remain. Across the first 32 windows, rank edge positions by the number of windows in which they occur among those ten at least $\lceil (L-1)/2 \rceil$ times; break ties by mean score over candidate occurrences. Keep the first eight positions, including lower-consistency positions if needed to fill the shortlist. At each window/pair, compute spectra only for shortlist positions present in that window/pair's top-ten candidates. This is a shared shortlist with window-specific eligibility, not eight independently selected edges in every window. The threshold sweep is separate; the spectral shortlist has no fixed 0.05 cutoff or guaranteed fraction of retained routing mass.

The numerical rank threshold is $\sigma_k > 10^{-6}\sigma_1$. Checkpoint estimates take the median over edges within window-layer, then layers within window, then windows. Depth curves stop after the first median and take the median across contributing windows. GPT-2 Medium contributes observations

in only 13 spectral windows, with at least two at each displayed layer; the other trained checkpoints contribute all 32. The estimates condition on this selected edge population, not all causal edges.

The active-envelope statistic summarizes spectral shape using a different fitted rank. For sorted singular values σ, choose r_{fit} from head-dimension multiples to minimize $\|\sigma - s_r \mathbf{1}_{k\le r}\|_2/\|\sigma\|_2$, where $s_r = r^{-1}\sum_{k\le r}\sigma_k$. For an adjacent pair, fix $r = r_{\mathrm{fit},\ell}$ and form $v_{\ell,k} = \log(\sigma_{\ell,k} + 10^{-12}) - r^{-1}\sum_{a\le r}\log(\sigma_{\ell,a} + 10^{-12})$ for $k \le r$, and likewise at $\ell + 1$. The active-envelope drift is $\|v_{\ell+1} - v_\ell\|_2/\sqrt{r}$; raw map drift is $\|O_{\ell+1} - O_\ell\|_F/\|O_\ell\|_F$ on the same retained edge. Removing scale and orientation changes the question being asked. Their numerical ordering is descriptive, not a matched test of adjacent-specific spectral persistence.

Normalization-aware tangent control. Let $J_\ell = D\Phi_\ell(X_\ell)$ be the full native block Jacobian and $\widehat{T}_\ell(v) = T_\ell[X_\ell](DN_{1,\ell}(X_\ell)[v])$ its normalization-lifted, frozen connection action. For a Gaussian tangent v, compare

$$a = \widehat{T}_{\ell+1}(J_\ell v), \qquad b = J_\ell(\widehat{T}_\ell(v)), \qquad d_{\mathrm{comm}} = \frac{\|a - b\|_F}{\sqrt{(\|a\|_F^2 + \|b\|_F^2)/2}}.$$

The next normalization derivative is evaluated at $X_{\ell+1}$. Replacing J_ℓ by the identity supplies the comparison baseline. Three probes per window are averaged, not counted as replicates. The retained populations are eight adjacent pairs on 64 windows for GPT-2 Small and sixteen pairs on 24 windows each for Qwen3-1.7B and Llama-3.2-3B, at contexts no greater than 64. Pair-mean defects of 1.25–1.43 remain far from zero, with near-zero path cosine and no material improvement over the identity baseline. This rejects the tested one-step tangent intertwining, not all nonlinear changes of coordinates. Jacobian products use float32 and connection actions float64; finite-difference checks validate the normalization lift.

Paired-output alignment. For aligned actions, use 32 windows and adjacent pairs whose lower layer lies in the middle 20–80% band. In row convention set $B = \mathcal{T}_\ell(Z_\ell)$, $C = \mathcal{T}_{\ell+1}(Z_\ell)$, with each connection frozen at its realized routing. Define $r_{\mathrm{id}} = \|C - B\|_F$, $r_{\mathrm{rot}} = \min_{Q^\top Q = I}\|C - BQ\|_F$, $r_0 = \|C\|_F$, and $f = 1 - r_{\mathrm{rot}}/r_{\mathrm{id}}$. No centering or scalar fitting is applied. The exact Procrustes residual satisfies $r_{\mathrm{rot}}^2 = \|B\|_F^2 + \|C\|_F^2 - 2\|B^\top C\|_*$ and is computed through small QR factors. Per pair, 64 null draws preserve C's spectrum while independently randomizing its token and feature frames; a separate control shuffles token rows. Window means of $f - q_{.95}(f_{\mathrm{null}})$ define the margins. Three-checkpoint Bonferroni intervals require the independent-frame lower bound above 0.05, shuffled-row lower bound above zero, and r_{rot}/r_0 upper bound below 0.95. A feature-only rotation leaves the optimal residual invariant.

Single-projection protocol. For compression, the selected zero-indexed layers are 6/18/14/18 for GPT-2 Small/Large and Qwen3-1.7B/4B. Their row-map output-projection dimensions are 768×768, 1280×1280, 2048×2048, and 4096×2560. Thirty-two windows fit the affine-free output second moment; the next 64 are held out, each with 63 next-token predictions. Candidate ranks are rounded down to multiples of eight at matrix-MAC fractions 0.25/0.5/0.75, giving 96/192/288, 160/320/480, 256/512/768, and 392/784/1176. The two-factor cost is $r(d_{\mathrm{in}} + d_{\mathrm{out}})$ versus native $d_{\mathrm{in}}d_{\mathrm{out}}$ per token; affine vector arithmetic is outside that count. Weight-SVD, magnitude coordinates, and three random orthonormal bases use the same ranks. No held-out basis or rank fitting occurs.

The primary loss estimand is per-window compressed-minus-native NLL in nats/token. Ten thousand paired window-bootstrap draws produce intervals corrected for 18 conditions within each checkpoint; bounded loss requires upper bound at most 0.02. Data-minus-magnitude and data-minus-mean-random comparisons correct six contrasts. There is no additional four-checkpoint family correction or statistical weight-SVD superiority test. Local action error removes the same exact affine term from compressed and native attention outputs before taking their relative Frobenius difference. Native no-op action/logit checks and exact NLL restoration pass on all evaluation windows; saved factors reproduce the in-memory projection. Maximum compression reconstruction residual is 1.964×10^{-15}.

C.2 ATTENTION RUNTIME MEASUREMENT CONFIGURATION

Compression timings use float64 eager prefill at context 64 on a 96-GiB PPU-ZW810 accelerator, accessed through the PPU backend's PyTorch `torch.cuda` interface. The measured checkpoints are GPT-2 Small/Large and Qwen3-1.7B/4B. Whole-model ratios compare median candidate forward time with the adjacent restored-native median; projection time uses 50 synchronized repeats. Working memory counts active projection/input storage and incremental peak allocation. Unused native weights retained for restoration are excluded; this endpoint is not deployment memory. Timing ratios have no independent-repeat confidence intervals. Other studies used different accelerator environments; their timings are not pooled with these measurements. The compression records retain source-code and token-window hashes, but not complete checkpoint, tokenizer, dataset, and runtime fingerprints. The all-layer reaction study retains those fingerprints for its own execution. Those records do not describe the compression timing environment.

C.3 COMPLETE-BLOCK COMPARISONS AND COMPONENT ACTIONS

For nonzero actions a, b in one fixed reference geometry, common-state replay reports relative drift, gain, and cosine:

$$d(a,b) = \frac{\|b-a\|_\pi}{\|a\|_\pi}, \qquad g(a,b) = \frac{\|b\|_\pi}{\|a\|_\pi}, \qquad c(a,b) = \frac{\langle a,b\rangle_\pi}{\|a\|_\pi\|b\|_\pi}. \tag{24}$$

They satisfy $d^2 = (g-1)^2 + 2g(1-c)$. Candidate blocks each recompute their own normalization, routing, attention, and reaction at the same input state; the comparison measures their actions there, rather than a global distance between block maps.

At every trajectory state X_ℓ, candidate block k recomputes $F_{k|\ell} = \Phi_k(X_\ell) - X_\ell$ with its own normalization, query/key routing, attention, and sequential FFN. The adjacent comparison uses $a = F_{\ell|\ell}$ and $b = F_{\ell+1|\ell}$. For this full-update comparison, write $d_F = d(a,b)$, $g_F = g(a,b)$, and $c_F = c(a,b)$, suppressing the window and layer indices. Alongside Equation 24, symmetric drift is

$$d^{\mathrm{sym}}(a,b) = \frac{\|b-a\|_F}{\sqrt{(\|a\|_F^2 + \|b\|_F^2)/2}}.$$

A fixed half-depth cyclic derangement selects the nonadjacent block control. Both comparisons use X_ℓ; the control is not a globally matched function-distance baseline, and more distant blocks can receive more atypical states. Statistics first average eligible layers within each window, then use 10,000 window-bootstrap draws. The 64-window trained population uses context 64; matched initialization controls use 16 windows for each of three seeds in GPT-2 Small and Qwen3-1.7B.

Directional distinction requires middle-band interval lower bounds $d_F > 0.5$ and directional fraction > 0.5, and cosine upper bound $c_F < 0.8$. Practical stabilization requires simultaneous late-band upper bounds $d_F < 0.10$, $|g_F - 1| < 0.10$, and $1 - c_F < 0.01$. The directional fraction is $2g_F(1-c_F)/d_F^2$ at each observation. Replaying the native block in float64 gives reconstruction error below 6×10^{-15}; connection reconstruction remains below 10^{-6}.

The separate native component profiles retain $F = u_I + u_A + u_D + b + q = u_S + u_A + b + q$, where $q = \psi(X^a)$ and $u_S = u_I + u_D$. They use 64 windows per trained checkpoint and uniform token reference. Component norm squares and every pairwise twice-inner-product are retained at each window and layer before aggregation. They support action magnitudes and interactions, not independent importance shares.

Reaction norm is reconstructed as $\|u_R\|^2 = \|b\|^2 + \|q\|^2 + 2\langle b, q\rangle$. The late-minus-early contrast first averages the relative action $m_c = \|u_c\|/\|F\|$ within each window over $[0.6, 0.9)$ and $[0.1, 0.4)$, respectively. We use 50,000 window-bootstrap draws and Bonferroni correction over all 28 coordinate/checkpoint contrasts, giving 99.8214% intervals. Table 4 reports their point estimates and intervals. Depth curves and middle-band component cosines are descriptive window means, without fitted alignment or additional confidence intervals. The expanded Gram sum reconstructs the update norm with maximum relative error 8.29×10^{-11} across 13,568 native observations; the common-state summaries cover 13,120 adjacent-pair observations. Bias-free checkpoints retain the same reaction definition with $b = 0$.

For the native branch comparison, we use the Gram matrix of u_S, u_A, b, and q to recover attention $a = u_S + u_A + b$ and FFN q norms without subtracting the large identity baseline. To separate reaction from its denominator, define $\Delta_{\mathrm{EL}} f$ as the mean within-window late-minus-early contrast of f over the same bands. At each positive-norm observation,

$$\log(\|u_R\|/\|F\|) = \log \|u_R\| - \log \|F\|.$$

Table 3 applies this identity before averaging. These are descriptive log-scale contrasts, not the arithmetic ratio contrasts in Table 4; they have no new confidence intervals. All 13,568 observations are retained, with maximum log-identity residual below 1.8×10^{-15}. The stable Gram reconstructs the update norm square to relative error below 7.2×10^{-13}.

The same Gram matrix supplies the signed projections $p_c = \langle u_c, F\rangle_\pi / \|F\|_\pi^2$ for $c \in \{S, A, R\}$. With $u_R = b + q$, its cross terms include the bias/FFN interaction. The sum $p_S + p_A + p_R = 1$ holds per observation to within 7.2×10^{-13}. Table 2 reports descriptive early–late means using the same windows and bands as the norm-ratio contrasts, without new confidence intervals. Their sum is zero before rounding. These are signed coefficients along the complete update, not nonnegative or causal importance shares. Because $p_c = m_c \cos_\pi(u_c, F)$ pointwise, magnitude and alignment both enter; separately averaged ratios must not be used to recover a mean cosine.

Table 2: Reaction's signed projection rises in Llama as its norm ratio falls. Descriptive early–late changes in p_c average 64 context-64 windows per checkpoint under uniform reference.

Checkpoint	Δp_S	Δp_A	Δp_R
GPT-2 Small	$+0.1475$	$+0.0169$	-0.1644
GPT-2 Medium	$+0.0761$	$+0.0160$	-0.0921
GPT-2 Large	$+0.0542$	$+0.0402$	-0.0944
GPT-2 XL	$+0.1016$	$+0.0849$	-0.1865
Qwen3-1.7B	$+0.0149$	$+0.0024$	-0.0173
Llama-3.2-3B	-0.0289	-0.0380	$+0.0669$
Qwen3-4B	$+0.0387$	-0.0050	-0.0338

Table 3: Reaction grows in mean-log magnitude in five checkpoints while the update grows faster. Entries are descriptive late-minus-early log contrasts over 64 context-64 windows per checkpoint; the ratio column equals reaction minus update before rounding.

Checkpoint	$\log \|u_R\|$	$\log \|F\|$	$\log(\|u_R\|/\|F\|)$
GPT-2 Small	-0.5756	-0.4476	-0.1280
GPT-2 Medium	$+0.1591$	$+0.2729$	-0.1138
GPT-2 Large	-0.0360	$+0.0394$	-0.0754
GPT-2 XL	$+0.0677$	$+0.2518$	-0.1841
Qwen3-1.7B	$+2.4838$	$+2.5511$	-0.0672
Llama-3.2-3B	$+0.6903$	$+0.7225$	-0.0322
Qwen3-4B	$+0.6031$	$+0.6511$	-0.0480

C.4 Reference geometry and cross-token controls

These controls use separate 16-window, context-64 populations in each checkpoint, with three architecture-matched initialization seeds per checkpoint. The two positive references are uniform weighting and one-step received mass $\pi_j = n^{-1} \sum_i P_{ij}$. The reference defines both the adjoint and the norm used to measure it. Adjoint amplification is $\|T^{*\pi} Z\|_\pi / \|TZ\|_\pi$, not the Frobenius ratio obtained after changing the adjoint alone. The middle-depth trained ranges are 13.8–31.9 and 2.44–8.60 under the respective references, versus 1.11–1.43 at initialization. Symmetric and skew connection actions strongly oppose one another under uniform reference, but their near-equal magnitudes also occur at initialization. Their cosine is algebraically determined by adjoint amplification and $\cos_\pi(TZ, T^{*\pi} Z)$, so it is not independent evidence of a learned cancellation mechanism. Changing reference changes both the adjoint and the measuring norm, leaving the native update unchanged; this comparison does not isolate coordinate transfer at a fixed measuring metric.

Table 4: Reaction's relative magnitude falls in every checkpoint; the other components vary. Entries are mean late-minus-early norm-ratio changes over 64 context-64 windows, with 99.8214% window-bootstrap intervals adjusted across 28 contrasts.

Checkpoint	Δm_I	Δm_A	Δm_D	Δm_R
GPT-2 Small	0.169	3.006	2.230	−0.117
	[0.133, 0.206]	[2.926, 3.086]	[2.155, 2.307]	[−0.128, −0.107]
GPT-2 Medium	−1.420	−4.882	−4.961	−0.099
	[−1.496, −1.338]	[−5.070, −4.675]	[−5.173, −4.724]	[−0.110, −0.088]
GPT-2 Large	−1.176	3.258	−0.387	−0.070
	[−1.387, −0.963]	[3.128, 3.372]	[−0.588, −0.187]	[−0.075, −0.065]
GPT-2 XL	−4.077	1.068	−3.848	−0.160
	[−4.979, −3.098]	[0.928, 1.206]	[−4.751, −2.872]	[−0.174, −0.146]
Qwen3-1.7B	4.851	0.591	4.807	−0.059
	[4.580, 5.132]	[0.472, 0.717]	[4.518, 5.106]	[−0.066, −0.053]
Llama-3.2-3B	−48.927	0.538	−48.832	−0.030
	[−49.815, −48.034]	[0.367, 0.710]	[−49.715, −47.941]	[−0.033, −0.026]
Qwen3-4B	26.346	4.933	26.700	−0.043
	[25.934, 26.776]	[4.745, 5.094]	[26.277, 27.139]	[−0.047, −0.040]

To separate source geometry from the first-position sink, let B retain the blocks of T with $i \neq j$ and source $j \neq 0$, and let $S_B = (B + B^{*\pi})/2$. The directional statistic is

$$c_B(Z) = \frac{\langle Z, S_B Z \rangle_\pi}{\|Z\|_\pi \|S_B Z\|_\pi}.$$

Its numerator equals $\langle Z, BZ \rangle_\pi$, so the ADR interpretation does not require this scalar to be computed through the adjoint. Field controls preserve the operator and compare the realized field against Gaussian fields with its token-population mean and full observable feature covariance. Additional tests remove the token mean and then the leading centered principal direction.

Operator controls hold the realized field and reference fixed. Within each head and row, causal route permutation rearranges source positions $1, \ldots, i-1$ while preserving the first source, diagonal, causal support, row mass, and weight multiset. Signed-permutation orthogonal conjugation changes feature orientation while preserving head-map ranks, spectra, and Frobenius norms. Four draws are averaged within each observation, not treated as replicates.

Middle-band means first average layers within a window. Trained intervals resample windows; initialization contrasts resample seeds with a common window resample across seeds. The retained paired 95% intervals are positive for all 28 directional operator-control contrasts and their training excesses; this is a collection of pointwise intervals, not an additional familywise guarantee. The raw cosine exceeds the covariance-matched control in all 14 checkpoint/reference cells. The corresponding signed-form advantage holds in only 26 of the 28 operator-control cells. The learned effect is therefore a directional relationship; it does not establish a universal increase in signed action or a smoothing law. The source-excluded symmetric action remains positively aligned with the normalized field. Holding the field fixed, causal route permutation reduces cosine by 0.027–0.067 under uniform reference and 0.037–0.075 under received-mass weighting; spectrum-preserving feature-map conjugation also reduces alignment. Together with the covariance-matched field comparisons and training excesses, this identifies learned directional pairing beyond the first-position sink and the tested low-order field statistics. It is a separate within-connection result, not a test of repeated computation across complete blocks.

C.5 Same-block reaction response to a transport approximation

This study evaluates Equation 17 at every block, holding its input fixed and replacing only its attention output projection. The affine-preserving map in Appendix D.2 retains at most 25% of projection matrix MACs. Ranks are 96, 128, 160 and 200 in the four GPT-2 checkpoints, 384 in Llama-3.2-3B, and 256 and 392 in Qwen3-1.7B/4B. The latter has a 4096×2560 row-map projection and retains 24.883% after rank rounding. All downstream native computation is recomputed; unchanged prefix states may be reused, with exact agreement checked against full forwards.

Write $s = \|\delta_r\|_F/\|\delta_t\|_F$, $c = \cos(\delta_t, \delta_r)$, $\kappa = \|\delta_F\|_F/\|\delta_t\|_F$, and $e_t = \|\delta_t\|_F/\|H\mathcal{T}(Z)\|_F$. Here δ_t is also the pre-FFN residual change, and δ_r the FFN increment change; the attention affine term is preserved. The criterion $\kappa^2 - 1 = 2sc + s^2$ is evaluated per observation. For $s > 0$, attenuation holds exactly when $c < -s/2$; it is impossible for $s \geq 2$. Negative c alone does not certify attenuation. These criteria apply to individual observations, not separately averaged s and c.

We exclude all 352 previously inspected windows from this study. Per checkpoint, 64 new context-64 WikiText-2 windows fit streamed double-precision calibration moments, 64 select and fix the evaluation conditions and predictors, and 128 are held out for final evaluation. Six equal-cost families share each rank: the calibration-output basis, weight SVD, magnitude coordinates and three seeded random bases. Development decisions for all seven checkpoints are fixed before held-out evaluation. A checkpoint proceeds to that evaluation only if it passes numerical validation and one of three criteria: a corrected depth contrast of absolute mean at least 0.05, a corrected positive predictive increment of at least 5% for the calibration-output family, or a corrected natural-minus-frozen loss contrast of absolute mean at least 0.002. GPT-2 Large passes numerical validation but none of these criteria, so its held-out split is not evaluated. The other six qualify through the depth contrast.

Table 5 reports the calibration-output family's held-out response. Joint opposition/amplification and attenuation indicators are averaged over layers within each window, then over windows. This directly measures the joint event rather than inferring it from separate mean signs. GPT-2 Small has mean $\kappa < 1$ at blocks 1, 2 and 6; Qwen3-1.7B has no such layer. Intervals in this table are descriptive 95% window-bootstrap intervals.

Table 5: FFN opposition can accompany amplification; attenuation varies by checkpoint. The calibration-output approximation uses a quarter projection-MAC budget and 128 held-out windows per selected checkpoint. κ ranges span layerwise mean error ratios. Event frequencies average layers within windows; brackets give descriptive 95% window-bootstrap intervals.

Checkpoint	Mean-κ range	Opposes and amplifies (%)	Attenuates (%)
GPT-2 Small	0.778–1.592	69.73 [68.10, 71.35]	29.56 [27.99, 31.12]
GPT-2 Medium	0.742–9.206	51.07 [50.16, 52.02]	44.99 [44.08, 45.90]
GPT-2 XL	0.958–6.330	43.00 [42.06, 43.99]	55.53 [54.52, 56.49]
Llama-3.2-3B	0.856–3.664	36.33 [35.38, 37.30]	54.16 [53.32, 54.94]
Qwen3-1.7B	1.024–1.904	95.76 [95.48, 96.04]	0.31 [0.14, 0.50]
Qwen3-4B	0.925–4.433	54.56 [53.82, 55.36]	38.50 [37.85, 39.06]

Depth contrasts first average $\log \kappa$ within each window's normalized early $[0, 0.25)$, middle $[0.25, 0.75)$ and late $[0.75, 1]$ bands. The terminal block is excluded from late and middle. Table 6 uses 100,000 paired window-bootstrap draws with correction across all 14 planned checkpoint/contrast combinations, including those not evaluated on held-out data. The terminal rebound is consistent across selected checkpoints, while the interior contrast changes sign; neither comparison is a monotonicity test.

Table 6: Terminal blocks have higher error-propagation ratios than middle depths in all six selected checkpoints. Entries are mean within-window $\log \kappa$ contrasts for the calibration-output family, using 128 held-out windows per checkpoint and Bonferroni intervals across 14 planned comparisons.

Checkpoint	Late minus early	Terminal minus middle
GPT-2 Small	$+0.020$ [0.015, 0.026]	$+0.140$ [0.127, 0.153]
GPT-2 Medium	-0.273 [-0.278, -0.269]	$+0.106$ [0.099, 0.114]
GPT-2 XL	-0.127 [-0.129, -0.125]	$+0.023$ [0.021, 0.025]
Llama-3.2-3B	-0.150 [-0.177, -0.122]	$+0.284$ [0.280, 0.289]
Qwen3-1.7B	-0.155 [-0.167, -0.145]	$+0.018$ [0.007, 0.030]
Qwen3-4B	-0.242 [-0.264, -0.221]	$+0.125$ [0.116, 0.134]

Separately within each approximation family, a frozen least-squares baseline predicts loss change from $\log e_t$ and layer indicators; an augmented model adds $\log \kappa$. Development uses eight-fold window-grouped cross-validation before fitting and freezing coefficients on all 64 development windows. Final performance is the mean window-relative MSE reduction. With correction across all 42 planned checkpoint/family combinations, none of the 36 observed comparisons attains a 5% increment: the largest corrected upper bound is 3.593%. In the calibration-output family, mean gains

range from -1.595% to $+0.199\%$, and every corrected upper bound is below 2.63%. This excludes the specified increment for these frozen predictors under conditional window uncertainty. Layer conditioning does not rule out a between-layer association, and no allocation policy was tested.

Table 7 supplies the absolute scale from the same held-out observations and fixed coefficients, without refitting. Let B_w and A_w be the layer-mean squared errors in window w. The declared increment is $\mathbb{E}_w[1 - A_w/B_w]$, not $1 - \mathbb{E}_w[A_w]/\mathbb{E}_w[B_w]$. For Llama the former is -1.595% while the latter is $+0.725\%$: the ratio of means gives greater weight to windows with larger baseline error. Baseline RMSE ranges from 0.00732 to 0.505 nats/token. Small incremental gains therefore need not imply small absolute prediction errors. These post-hoc summaries have no confidence intervals.

Table 7: Absolute loss-prediction error varies across checkpoints. The baseline uses attention error and layer; the augmented predictor adds reaction response. For the calibration-output family, coefficients are frozen on development data and MSE averages layers within windows, then 128 held-out windows. Units are $(\mathrm{nats/token})^2$.

Checkpoint	Baseline MSE	Augmented MSE
GPT-2 Small	6.87925×10^{-4}	6.86788×10^{-4}
GPT-2 Medium	1.58818×10^{-4}	1.58801×10^{-4}
GPT-2 XL	5.36030×10^{-5}	5.36583×10^{-5}
Qwen3-1.7B	2.55267×10^{-1}	2.55269×10^{-1}
Llama-3.2-3B	8.35689×10^{-4}	8.29626×10^{-4}
Qwen3-4B	6.83913×10^{-4}	6.84900×10^{-4}

Only Qwen3-1.7B meets the selection criterion for a held-out diagnostic that adds δ_t to the native block output while keeping the FFN increment at its native value. Its natural-minus-frozen loss contrast is $+0.0674$ $[0.0458, 0.0911]$ nats/token, with layers averaged within window and correction across seven planned comparisons. The natural reaction response raises loss relative to this constructed replay. The replay requires the native FFN increment from an unmodified forward pass. This is neither a deployable FFN removal nor a mediation decomposition. The frozen-reaction diagnostic is not evaluated on held-out windows for the other checkpoints.

Every evaluated window passes native reconstruction and an unchanged-projection replacement check at every layer. Across 135,168 final interventions, reconstruction is below 4.3×10^{-15}, local closure below 9.5×10^{-14} and the kappa identity below 1.3×10^{-11}; block-input changes and restoration loss are zero. Factor and input identities are verified. Independent reaggregation agrees within 4.2×10^{-16}, without constituting another statistical replicate. The six held-out checkpoint results are conditional on development selection; tokenizer-specific windows can share documents and are not paired across architectures. One corpus, context and budget with single-block interventions does not establish training specificity, multi-block composition or runtime benefit.

C.6 Local reaction intervention

The FFN attenuation uses 32 context-64 WikiText-2 windows per checkpoint and middle-band sites. It replaces the native additive increment q by $0.9q$, then recomputes normalization and routing for the following attention module. Relative response is the Frobenius change in the bias-subtracted attention action divided by its native bias-subtracted norm. Statistics average tested layers within each window and use 10,000 window-bootstrap draws for percentile 95% intervals. Across the seven checkpoints, learned-to-random response ratios are 0.86–1.69 under per-token norm-matched random directions. Against token-permuted increments, learned-to-shuffled ratios are 0.832–0.978 with the entire 95% interval below one in six checkpoints; GPT-2 Medium instead has 1.023 [1.015, 1.033]. These controls establish coupling but not universal learned-direction specificity. Complete FFN removal is a different, out-of-distribution intervention and is not the local attenuation estimand.

D Extended Results and Controls

D.1 ADR-guided task sensitivity across depth

This case study asks what the decomposition contributes to a controlled task-level intervention. It distinguishes three objects: the strength of a native action, predictive sensitivity to its direction, and the change in that sensitivity when local reaction responds. We study all 28 zero-based layers of Qwen3-1.7B on the same 32 constructed entity-pair/event items. The profile is exploratory: layers 6, 13 and 20 were examined first, and that comparison did not meet the prespecified 0.05-nat primary effect-size threshold. The complete-depth study extends that comparison on the same development population, without held-out confirmation of the effects described below.

Task and fixed population. Each item supplies two role assignments, two queries, active and passive voices, and role and copy tasks, giving 16 prompts. For example, "Alice thanked Bob" and "Bob was thanked by Alice" support agent/patient questions; the copy task instead asks who was mentioned first or second. Both tasks use the same two answer names and fixed four-demonstration completion format. The readout is the next-token correct-minus-incorrect candidate logit margin, in nats. These are forced-choice predictions, not unrestricted generation accuracy. Query wording and difficulty differ between tasks, so the contrast concerns this controlled comparison rather than an isolated semantic faculty.

Within each item and voice, retain assignment/query pairs only when both tasks have strictly positive native margins. Freeze this mask before intervention and retain it across layers, directions and reaction modes, including predictions made incorrect by intervention. All 32 items have nonempty masks in both voices: 111 active and 61 passive matched pairs, or 344 of the original 512 prompts. Average selected prompts within each task/voice, voices equally within items, and items equally. The two voices have different masks; their comparison is not an isolated causal voice effect.

Geometry defines the intervention. In stacked-column convention, let $z = Z_t$ at the last observed query token, and let $\mathcal{T}_{tt}$ be its diagonal connection block. Define

$$K = \tfrac{1}{2}(\mathcal{T}_{tt} - \mathcal{T}_{tt}^{\top}), \qquad L = \tfrac{1}{2}(\mathcal{T}_{tt} + \mathcal{T}_{tt}^{\top}), \qquad v_A = HKz, \quad v_S = HLz.$$

Here $S = I + D$ separates symmetric transport from the identity baseline in D. A diagonal skew modification changes A alone; a diagonal symmetric modification changes D with I fixed. Both preserve causal support and leave the query token's cross-token contribution unchanged. Diagonal $v_A + v_S$ covers only part of the attention action and does not isolate cross-token semantic pathways.

Write the native attention increment as $a = H\mathcal{T}Z + b$ and $U = X + a$. For each item/layer, use the common dose

$$\rho = 0.25 \min_{\text{all 16 native prompts}} \{\|v_A\|, \|v_S\|, \|a_t - b\|\}, \qquad p_c = -\rho v_c / \|v_c\|, \quad c \in \{A, S\}.$$

Two fixed random controls per direction (seeds 91411 and 91412) have norm ρ. Skew controls are orthogonal to z; symmetric controls preserve the native perturbation's component parallel to z and randomize its perpendicular component at fixed norm. Thus controls respect the realizable diagonal geometry instead of comparing with unconstrained random displacements. No selected prompt/site has a degenerate symmetric-control direction.

Let p denote the perturbation embedded at token t and $\psi(U) = \mathrm{FFN}(N_2(U))$. The responding-reaction arm executes $X'_{\mathrm{out}} = U + p + \psi(U + p)$; the frozen-reaction arm executes $X'_{\mathrm{out}} = U + \psi(U) + p$. Only the intervention site's realized reaction is frozen; all downstream blocks and final logits are recomputed in both arms. This comparison changes reaction's evaluation point without deleting the FFN or treating ADR coordinates as native executable branches.

Controlled readout and uncertainty. For component c, task q and reaction mode m, let $B_{c,q,m}$ be the change in correct-answer margin under the native-direction perturbation minus its mean change under the two matched controls. Define, within each item and voice before the equal-weight averages above,

$$V_{c,m} = B_{c,\mathrm{role},m} - B_{c,\mathrm{copy},m}, \qquad \Delta_m = V_{A,m} - V_{S,m}, \qquad \Gamma = \Delta_{\mathrm{responding}} - \Delta_{\mathrm{frozen}}.$$

Positive Δ means a more positive control-adjusted role-minus-copy effect for the opposite-native A direction than for S. It does not mean that A is beneficial, more important, or a larger native logit contributor. Γ measures how allowing local reaction to respond modifies this endpoint contrast, not an additive attribution of the native prediction.

All 896 item/layer sites pass the numerical checks: relative reconstruction below 10^{-6}, double-precision local vector closures below 10^{-10}, and no-op/replay endpoint agreement below 10^{-6}. Uncertainty uses 10,000 jointly paired item-bootstrap draws (seed 91521). For each bootstrap draw, take the maximum absolute centered deviation across layers after dividing by each layer's original item standard error. Its 95th percentile gives the common multiplier for symmetric bands around the layer means; zero-variance coordinates have degenerate bands. Coverage is across depth for each curve separately, not across all quantities. Paired depth-change and correlation intervals are 95% item-bootstrap intervals. Prompts, controls and layers are not independent replicates.

Depth-specific directional sensitivity. Table 8 reports every layer, including sign reversals and small pooled means. Layer 9 has $\Delta_{\text{frozen}} = 0.94992$ with simultaneous band $[0.66346, 1.23638]$; its constituents are $V_A = 0.57222$ and $V_S = -0.37769$, and both voices give positive contrasts. The neighboring layer 10 has -0.19755 with band $[-0.23665, -0.15844]$; negative contrasts recur at layers 23 and 26. The paired layer-6-to-20 increase is 0.04344 with interval $[0.03164, 0.05511]$, but the complete profile identifies nonmonotonic sensitivity rather than a global transition from negative to positive effects.

Table 8: Task contrasts change sign through depth in an exploratory Qwen3-1.7B study. All rows use the same 32 matched items and native-correct selection; units are candidate-margin nats. Frozen Δ equally weights active and passive voices; Γ is responding minus frozen. Layers are zero-indexed.

Layer	Frozen	Active	Passive	Responding	Γ
0	0.11192	0.09909	0.12474	0.06315	-0.04877
1	-0.02867	-0.00281	-0.05454	-0.02818	0.00049
2	-0.04760	-0.03598	-0.05923	-0.02681	0.02079
3	0.00814	0.00486	0.01142	0.00629	-0.00184
4	-0.11658	-0.38809	0.15493	-0.14219	-0.02562
5	0.01015	0.06414	-0.04384	0.04239	0.03224
6	-0.01215	-0.00796	-0.01633	-0.04658	-0.03444
7	0.04648	-0.00699	0.09994	0.03278	-0.01370
8	0.01560	0.00654	0.02466	0.03529	0.01969
9	0.94992	0.74537	1.15446	0.96892	0.01900
10	-0.19755	-0.13525	-0.25984	-0.01078	0.18677
11	0.04312	0.43327	-0.34703	0.56546	0.52234
12	0.00011	0.02123	-0.02100	-0.01555	-0.01567
13	-0.00210	0.06202	-0.06623	-0.04253	-0.04043
14	0.06561	0.02942	0.10180	0.16720	0.10158
15	-0.01400	-0.09495	0.06695	0.02446	0.03846
16	0.04640	-0.01167	0.10448	0.06710	0.02069
17	0.06416	0.03838	0.08993	-0.03370	-0.09785
18	0.03906	0.12118	-0.04306	0.12767	0.08860
19	0.08881	0.10932	0.06829	0.12979	0.04099
20	0.03130	0.03685	0.02575	0.02924	-0.00206
21	0.02049	0.01396	0.02703	-0.01705	-0.03754
22	0.00736	-0.01717	0.03189	0.02536	0.01800
23	-0.17316	-0.03491	-0.31141	-0.09583	0.07733
24	0.01475	0.00836	0.02115	0.01025	-0.00451
25	0.02595	0.01016	0.04173	0.01172	-0.01423
26	-0.07567	-0.01970	-0.13164	0.00216	0.07783
27	0.08595	0.07587	0.09603	0.02607	-0.05988

Native strength and reaction answer different questions. We summarize native strength by the mean selected-prompt log ratio $\log(\|v_A\|/\|v_S\|)$ within each item/layer, weighting tasks and voices equally. Pooling adjacent within-item differences, its Pearson correlation with changes in Δ_{frozen} is

0.14965 $[0.12298, 0.17580]$. Dividing each item's contrast by its dose before computing changes gives 0.11135 $[0.07241, 0.14748]$. Layer 9 retains a positive dose-normalized peak of 0.38754 nats per unit norm with simultaneous band $[0.27113, 0.50395]$. Thus the strength balance tracks some variation, while direction-dependent task sensitivity requires its own measurement. This is descriptive association, not causal mediation; division by a finite dose is not a derivative or a complete removal of nonlinear dose effects. The normalized association also differs by voice: active $r = 0.00204$ $[-0.04799, 0.06074]$, passive $r = 0.15742$ $[0.13308, 0.18472]$.

At layer 11 the frozen voice contrasts, $+0.43327$ and -0.34703, largely cancel. Letting local reaction respond changes the pooled contrast from $+0.04312$ to $+0.56546$, giving $\Gamma = 0.52234$ with simultaneous band $[0.32457, 0.72012]$. At layer 9 the corresponding means are 0.94992 and 0.96892, with a Γ band containing zero. Reaction therefore modifies the task contrast differently at different layers; a small frozen pooled effect need not imply weak effects in each voice or after local reaction.

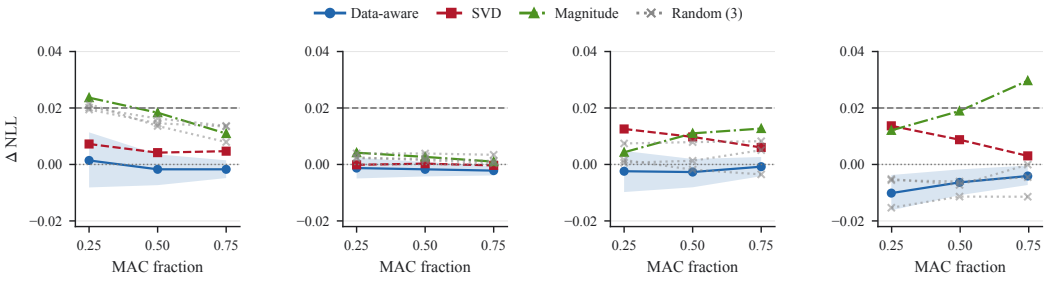
The practical value of ADR here is an intervention design that distinguishes geometrically constrained directions and their sequential consequences. It reveals task-sensitive structure hidden by native norms or a pooled score alone. The evidence is conditional on native-correct predictions in one checkpoint and one constructed development population; it supports this diagnostic use, not a universal division of linguistic labor between A and S.

D.2 Output-projection approximation and prediction tolerance

The transport spectra motivate a realizable test: can one attention output projection be approximated while preserving prediction? At the middle layer of GPT-2 Small/Large and Qwen3-1.7B/4B, fit an orthonormal rank-r output basis Q on 32 calibration windows, using the leading directions of the affine-free output second moment. For row-map weights W, replace the projection by

$$\widehat{f}(x) = (x - b_V)(WQ)Q^\top + b_V W + b_O.$$

This two-product map preserves the exact affine term. Q/K/V projections and GQA value groups remain native; b_V concatenates the native value biases at the projection input, and b_O is its output bias. This map attains minimum calibration output-reconstruction error over rank-at-most-r replacements of the fixed linear projection (Appendix A.1). It does not optimize prediction loss, which must be tested separately. The three ranks spend roughly 25%, 50%, or 75% of the original projection's matrix multiply–accumulates (MACs).



(a) GPT-2 Small (b) GPT-2 Large (c) Qwen3-1.7B (d) Qwen3-4B

Figure 5: Data-aware approximation of one output projection meets the loss bound at reduced matrix cost. Axes show the native projection-MAC fraction and mean compressed-minus-native NLL over 64 held-out windows. Data-aware uses calibration outputs; SVD uses weights; Magnitude selects coordinates; gray curves are three random bases, all at matched rank and cost. Data-aware bands are window-bootstrap intervals adjusted across 18 conditions per checkpoint. Dashed: 0.02-nat/token bound; dotted: unchanged loss. Lines connect tested ranks.

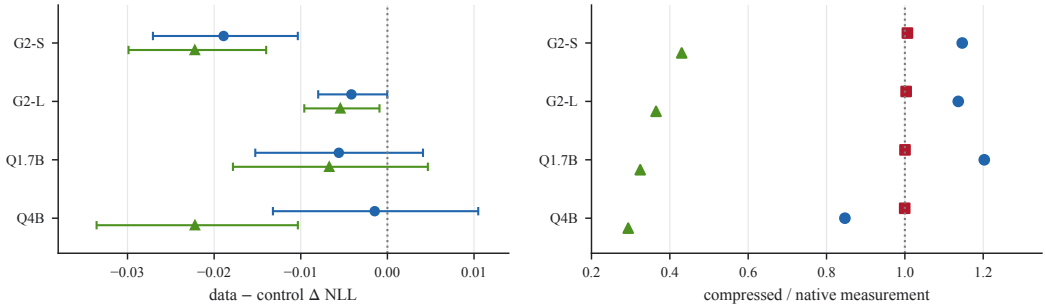
Figure 5 compares all methods on 64 disjoint held-out windows. The data-aware basis meets the declared upper loss-change bound of 0.02 nats/token at every budget in all four checkpoints. Equal-rank weight-SVD, magnitude-coordinate, and three random-subspace controls expose the interpretation: GPT-2 Small passes the stronger magnitude/random comparison at all three budgets, GPT-2

Large only at the quarter budget, and neither Qwen checkpoint establishes random-control advantage. Figure 6 gives the paired quarter-budget contrasts, including GPT-2 Large's narrow random-control margin.

At the quarter budget, local affine-free action error is 0.342–0.660 despite the small prediction-loss changes. Prediction can therefore tolerate substantial change in this attention output. Combined with the drift measurements, this distinguishes two statements: adjacent actions are directionally distinct, and reproducing one computation exactly is not required to meet the tested prediction-loss bound. The intervention does not remove the measured depthwise drift or an entire layer. The stronger advantage over generic equal-budget approximations is confined to the two GPT-2 cases. Affine-preserving zeroing and scalar attenuation were not evaluated. The result therefore does not distinguish selected-subspace tolerance from more general local dispensability in the Qwen cases.

D.2.1 Execution cost

Removing matrix MACs is distinct from accelerating execution. Figure 6 separates measured projection cost from whole-model latency at the quarter budget, where roughly 75% of the selected projection's matrix MACs are removed. Projection working memory falls by 57–71%, but the two-product implementation is slower in three checkpoints; only Qwen3-4B's projection runs faster. Whole-model latency ratios remain near one, establishing no meaningful whole-model speedup in the measured setting. Appendix C.2 specifies the acceleration-test configuration.



(a) Paired loss advantage

(b) Measured cost ratios

Figure 6: Quarter-budget projection savings yield no meaningful whole-model speedup. Left: data-minus-control NLL against mean random bases (blue circles) and magnitude coordinates (green triangles); negative favors the data-aware basis. Intervals bootstrap 64 held-out windows and adjust six comparisons per checkpoint. Right: whole-model latency (red squares), projection latency (blue circles), and projection working memory (green triangles), relative to native cost; below one is lower. Timing is descriptive; execution details are in Appendix C.2.

The equal-action-error comparison in Appendix D.4 separates prediction tolerance from action fidelity: no measured compressed data-aware point reaches relative action error 0.1. Cross-layer sharing, multi-layer compression, decoding, and lower-precision deployment require separate tests.

D.3 Head and mixture transport capacity

The weight-only head study forms each value–output product from double-precision factors. Its mean 99%-energy rank is 95.8–97.6% of head dimension across the checkpoints. Table 9 gives the separate full-depth active-mixture study, also using products and spectra in double precision. It normalizes rank by model width and uses the shared edge selection and window-respecting medians in Appendix C.1. The head means and edge medians are not a paired counterfactual: they distinguish head capacity from sampled-mixture capacity, without identifying a common compressible subspace or the effect of equalizing head weights.

In the separate middle-band study, strict relative-Frobenius full-width anisotropy is 2.6–3.3 across the seven context-64 checkpoints, with front-loaded spectral decay rather than a flat orthogonal block. Full-width anisotropy includes the effect of architectural rank ceilings; the fixed-active-rank test in Table 9 isolates unequal gains within the retained block.

Relative to the median of three initialization seeds, trained GPT-2 Small lowers r_{99}/d by 0.251 and increases δ_{99} by 0.097; Qwen3-1.7B lowers the rank fraction by 0.143 and increases the defect by 0.076. Every individual seed has the same ordering. These are controlled descriptive associations for two architectures, not a seven-checkpoint causal effect of training.

Table 9: Active mixtures concentrate energy in anisotropic subspaces. Entries are checkpoint medians with 95% window-bootstrap intervals; counts are eligible spectral windows. r_{99} retains 99% of squared singular-value mass; zero δ_{99} denotes scaled isometry within that subspace.

Checkpoint	Windows	r_{99}/d	Active defect δ_{99}
GPT-2 Small	32	0.377 [0.367, 0.391]	0.560 [0.548, 0.569]
GPT-2 Medium	13	0.349 [0.309, 0.492]	0.587 [0.566, 0.614]
GPT-2 Large	32	0.427 [0.383, 0.462]	0.590 [0.578, 0.600]
GPT-2 XL	32	0.409 [0.388, 0.437]	0.628 [0.620, 0.644]
Qwen3-1.7B	32	0.265 [0.260, 0.273]	0.533 [0.530, 0.535]
Llama-3.2-3B	32	0.217 [0.215, 0.222]	0.514 [0.510, 0.523]
Qwen3-4B	32	0.245 [0.237, 0.251]	0.520 [0.513, 0.525]

D.4 Attention alignment and functional fidelity

Spectral shape and functional change. In the three-checkpoint context/corpus endpoint grid, centered log-spectrum drift is 0.04–0.12 while raw edge-map drift is 1.27–1.38 in all twelve cells. The former removes scale and orientation and restricts the comparison to a fitted leading block. These different measurements distinguish spectral shape from the full map; they do not independently establish a learned spectral-stability effect. Across GPT-2 Small, GPT-2 Large, and Qwen3-1.7B at contexts 64, 128, and 256 on WikiText-2 and SlimPajama, D_K lies in 1.29–1.43 and D_{TZ} in 1.21–1.25. Scale-normalized action drift remains above one in all eighteen cells. The separate twelve-cell spectral comparison retains anisotropy and low effective rank alongside smaller changes in shape than function. These stress tests cover three checkpoints; they do not establish corpus invariance across all seven.

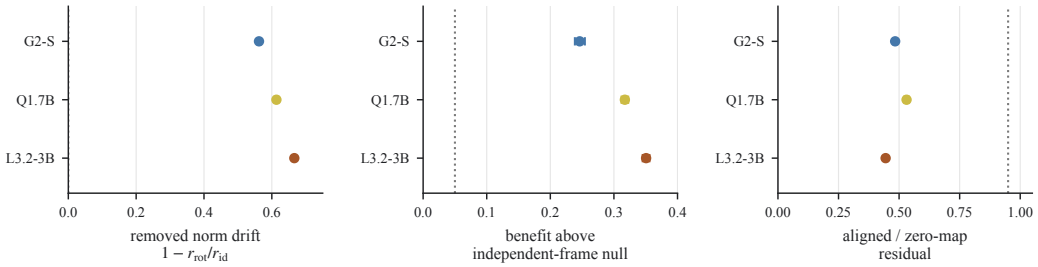
Within the head mixture that constructs O, changes in routing coefficients and head maps have comparable symmetric allocations, with a ratio of 0.82–1.04 and substantial negative interaction terms. Neither ingredient alone explains the joint drift. A normalization-aware full-block tangent map also fails to reconcile consecutive connections in the three tested checkpoints (Appendix C.1).

Calibrated paired-output alignment. For a common field, compare $B = \mathcal{T}_\ell(Z_\ell)$ and $C = \mathcal{T}_{\ell+1}(Z_\ell)$ through the unaligned residual $r_{\mathrm{id}} = \|C - B\|_F$ and the orthogonally aligned residual $r_{\mathrm{rot}} = \min_{Q^\top Q = I} \|C - BQ\|_F$. Figure 7 gives the calibrated action comparison. Alignment is fitted separately to each window and adjacent pair; it does not provide a parameter basis transferable to other inputs. A feature-only rotation leaves the optimal residual invariant, while the independent-frame null changes token geometry as well. Neither that null nor the shuffled-token control holds the common input field fixed while independently disrupting the feature maps. Shared-input token geometry therefore remains a possible explanation. The result establishes calibrated paired-output alignment, not training-specific shared operator geometry or a causal source of frame change.

The shuffled-token alignment margin is the window mean of observed benefit minus the control's per-pair 95th percentile. All three intervals below are positive, with the same three-checkpoint correction as the primary alignment figure. The control rejects a generic token-pairing explanation; it does not identify which operator or common-field structure causes the observed alignment.

Checkpoint	Mean shuffled-token margin	98.333% interval
GPT-2 Small	0.1511	[0.1423, 0.1614]
Qwen3-1.7B	0.1633	[0.1590, 0.1678]
Llama-3.2-3B	0.2099	[0.2046, 0.2153]

The equal-action-error comparison below selects the lowest matrix-MAC data-aware rank among measured points satisfying the stated mean error bound. A dash means no point in the finite sweep meets it. The exact native projection has zero error and remains a feasible baseline at every bound.



(a) Alignment benefit (b) Null-calibrated margin (c) Residual against zero map

Figure 7: Fitted alignment reduces adjacent attention-action drift beyond the tested null. Left: fraction of norm drift removed. Middle: benefit above the 95th percentile of a spectrum-preserving independent token/feature-frame null, compared with the 0.05 margin. Right: aligned residual relative to $r_0 = \|C\|_F$, compared with 0.95. Means average middle-band pairs within 32 windows, then windows; bars are 98.333% window-bootstrap intervals adjusted across three checkpoints. Fits are specific to each window and pair.

The table therefore distinguishes loss tolerance from action fidelity without claiming an unmeasured continuous optimum.

Checkpoint	Error 0.05	0.10	0.20	0.30	0.50
GPT-2 Small	–	–	–	–	192
GPT-2 Large	–	–	–	–	320
Qwen3-1.7B	–	–	768	512	256
Qwen3-4B	–	–	–	1176	784

GPT-2 Large's quarter-budget data-minus-random interval has upper bound -0.0000194 nats/token in the primary bootstrap. A secondary million-draw precision check with unchanged population, statistic, seed, and quantiles gives upper bound -0.0001254. This checks Monte Carlo precision and does not add independent experimental evidence. The margin remains narrow. Neither Qwen checkpoint has a negative upper random-control bound at any budget.

D.5 Routing selectivity and depthwise change

Table 10: Worst-case routing gain rises while native-field routing gain falls. Entries are mean late $[0.6, 0.9)$ minus early $[0.1, 0.4)$ contrasts over 64 context-64 windows per checkpoint. Both action gains use the same normalized field; intervals are in Table 11.

Checkpoint	$\|P\|_2$	$\|PZ\|_F/\|Z\|_F$	$\|\mathcal{T}(Z)\|_F/\|Z\|_F$
GPT-2 Small	2.311	−0.2821	0.0618
GPT-2 Medium	1.105	−0.1379	0.0574
GPT-2 Large	2.340	−0.2715	0.0108
GPT-2 XL	1.955	−0.2658	0.0221
Qwen3-1.7B	0.916	−0.0420	−0.0153
Llama-3.2-3B	1.147	−0.0800	0.0006
Qwen3-4B	1.336	−0.0546	−0.0168

Table 11 gives the full uncertainty for the point estimates in Table 10. These retained gain contrasts use the original eight-metric correction, distinct from the four-quantity routing-dynamics analysis below.

The routing-dynamics comparison uses the same seven checkpoints and 64 context-64 windows per checkpoint as the routing-gain study. A matched float64 capture retains every head-mean routing matrix. Token-window hashes and all repeated entropy values agree with the earlier routing measurements; every native attention reconstruction residual is below 10^{-6}. Routing statistics and bootstrap summaries are independently recomputed from the saved matrices before use.

Table 11: Uncertainty for the gain contrasts in Table 10. The 99.375% window-bootstrap intervals adjust for eight metric families using 64 context-64 windows per checkpoint.

Checkpoint	$\|P\|_2$	$\|PZ\|_F/\|Z\|_F$	$\|\mathcal{T}(Z)\|_F/\|Z\|_F$
GPT-2 Small	$[2.261, 2.357]$	$[-0.2862, -0.2776]$	$[0.0575, 0.0666]$
GPT-2 Medium	$[1.009, 1.197]$	$[-0.1415, -0.1342]$	$[0.0526, 0.0622]$
GPT-2 Large	$[2.254, 2.418]$	$[-0.2755, -0.2675]$	$[0.0101, 0.0115]$
GPT-2 XL	$[1.865, 2.026]$	$[-0.2711, -0.2607]$	$[0.0202, 0.0239]$
Qwen3-1.7B	$[0.868, 0.962]$	$[-0.0444, -0.0395]$	$[-0.0159, -0.0148]$
Llama-3.2-3B	$[1.097, 1.190]$	$[-0.0815, -0.0785]$	$[0.0005, 0.0007]$
Qwen3-4B	$[1.274, 1.397]$	$[-0.0583, -0.0511]$	$[-0.0171, -0.0165]$

Normalized entropy E_ℓ retains the definition in Appendix C.1. A complementary concentration measure is

$$C_\ell = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{(i+1) \sum_j P_{\ell,ij}^2 - 1}{i}.$$

Uniform causal routing gives zero and single-source routing gives one. Unlike matrix spectral norm, this measure describes concentration within individual query rows. Figure 9 shows its full-depth profile.

For layers ℓ and k, define

$$D(\ell, k) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{2} \sum_j |P_{\ell,ij} - P_{k,ij}|.$$

Adjacent comparisons use midpoint depth $(\ell + 0.5)/(L-1)$; concentration profiles use layer depth $\ell/(L-1)$. Within-window means in early $[0.1, 0.4)$ and late $[0.6, 0.9)$ bands define the three contrasts. For each window, the late-band diameter is the largest $D(\ell, k)$ among distinct pairs with both layer depths in $[0.6, 0.9)$. We then average those window maxima; this is not the maximum over all windows.

Table 12 reports the four primary summaries with 98.75% intervals from 10,000 window-bootstrap draws, correcting four quantities within each checkpoint. All seven entropy contrasts are negative, and all squared-L_2 concentration contrasts are positive. Adjacent TV decreases in every checkpoint, but each lower interval for the mean late-band diameter exceeds the prespecified 0.05 TV scale. Thus no checkpoint passes the declared descriptive criterion of an upper interval below 0.05. This scale represents average reassigned routing mass, not predictive loss tolerance or an asymptotic convergence definition. The larger late-band distances also prevent small adjacent changes from being mistaken for an unchanged late routing pattern.

Source-specific measurements locate this concentration. Direct received mass shows the same qualitative depth trend in all seven checkpoints. From early to late depth, the off-origin excess mass delivered to the first position increases by 0.121–0.342. Over the same comparison, total-variation distance to a pure first-position sink decreases by 0.119–0.336, while distance from uniform causal routing increases. The intervals are clustered over windows and exclude zero in the stated directions. Architecture-matched initialization controls show much smaller effects, although one of six individual random seeds satisfies the within-run formation rule at a small magnitude.

Figure 8 shows this source specialization, including terminal departures from monotonicity. Mean row entropy falls and non-normality increases. The pure-sink distance expresses the received-mass result in distance form, not independent evidence; distance from uniform causal routing supplies the distinct baseline comparison. First-position concentration means that later queries increasingly read that source, not that their information is written into its causal state.

Table 12: Routing becomes more selective and changes less between neighbors, yet late-layer routes remain distinct. Entries are means and 98.75% window-bootstrap intervals over 64 context-64 windows per checkpoint. Δ denotes late minus early, using midpoint depth for adjacent TV and layer depth for concentration. Late diameter averages each window's largest late-band pairwise TV.

Checkpoint	ΔE	ΔC	Δ adjacent TV	Late diameter
GPT-2 Small	-0.200 $[-0.207, -0.193]$	0.284 $[0.276, 0.292]$	-0.113 $[-0.116, -0.109]$	0.136 $[0.132, 0.140]$
GPT-2 Medium	-0.113 $[-0.120, -0.105]$	0.142 $[0.128, 0.156]$	-0.077 $[-0.079, -0.075]$	0.206 $[0.197, 0.214]$
GPT-2 Large	-0.262 $[-0.271, -0.252]$	0.296 $[0.281, 0.310]$	-0.041 $[-0.043, -0.039]$	0.190 $[0.183, 0.197]$
GPT-2 XL	-0.223 $[-0.231, -0.214]$	0.262 $[0.247, 0.275]$	-0.042 $[-0.044, -0.040]$	0.151 $[0.146, 0.156]$
Qwen3-1.7B	-0.123 $[-0.130, -0.116]$	0.155 $[0.146, 0.163]$	-0.064 $[-0.066, -0.062]$	0.297 $[0.289, 0.304]$
Llama-3.2-3B	-0.192 $[-0.198, -0.185]$	0.224 $[0.214, 0.233]$	-0.027 $[-0.029, -0.025]$	0.126 $[0.120, 0.133]$
Qwen3-4B	-0.168 $[-0.175, -0.162]$	0.201 $[0.192, 0.210]$	-0.074 $[-0.077, -0.071]$	0.332 $[0.328, 0.336]$

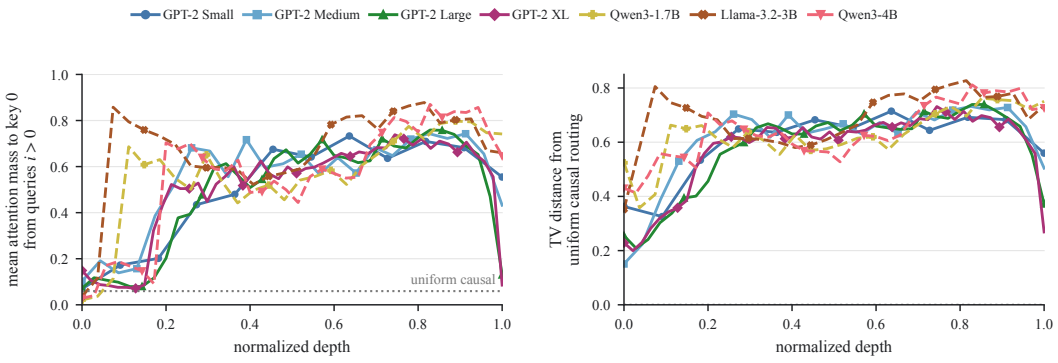


Figure 8: Routing increasingly selects the first position, with terminal reversals. Left: mean attention mass on key 0 from queries $i > 0$; the dotted line is uniform causal routing. Right: total-variation distance from the full uniform causal routing matrix. Curves average 64 context-64 WikiText-2 windows per checkpoint; solid lines denote MHA and dashed lines GQA.

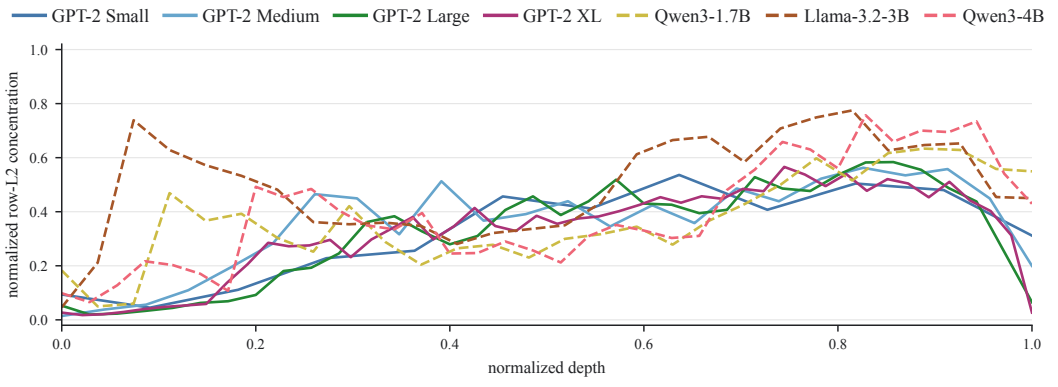


Figure 9: Routing concentration rises from early to late depth without changing monotonically. Curves show mean C_ℓ over 64 matched context-64 windows per checkpoint. Each row's squared-L_2 concentration is normalized by causal support size, then averaged over queries $i > 0$; zero denotes uniform routing and one single-source routing.

E DISCUSSION

The depthwise findings show why complete updates and component actions should be compared together. On common inputs, neighboring updates are more aligned than the fixed distant control, yet their transport and reaction directions remain distinct. Reaction's depth profiles also separate relative magnitude, absolute scale, and alignment with the complete update. ADR thus resolves changes within attention and its sequential interaction with the FFN that branch norms alone miss.

The normalization rescaling tests which measurements depend on how an unchanged block function is expressed. Identity and diffusion magnitudes can change, while symmetric transport, advection, and reaction remain unchanged under the specified rescaling at fixed reference. In the five larger checkpoints, the high diffusion alignment and low symmetric-transport alignment show why coordinate persistence alone does not establish persistent transport. Separating the identity baseline therefore matters when interpreting similarity across depth.

The causal result gives a structural test for proposed modifications. For a fixed realized causal connection, coupled scalar gains $\alpha = \delta = \gamma$ give $I+\gamma(\mathcal{A}_\pi+\mathcal{D}_\pi) = (1-\gamma)I+\gamma T$ and retain causal support. A prospective architecture could test a learned layerwise γ against the unchanged block, recomputing the FFN after the modified attention state and comparing held-out prediction at matched model cost. Its coefficients must be constructed causally as well. The proposition guarantees the support of the realized operator; prediction quality requires the proposed evaluation. Token-diagonal skew and symmetric controls allow separate geometric perturbations without adding future-token edges. Their task effects still need matched interventions, as the exploratory single-checkpoint study in Appendix D.1 illustrates.

The measured depth profiles do not specify which computation can be removed. Layer-pruning results such as ShortGPT's require evaluation of the modified model (Men et al., 2025). Identity, advection, and diffusion are components of the analysis, and norm ratios can change differently from absolute scale or alignment with the update. For an executable attention approximation, such as a reduced value/output projection, the relevant local error is δ_F, after the FFN responds to the altered attention state; δ_t alone can miss amplification, as the six-checkpoint intervention shows. A candidate that preserves this local action must then be tested on held-out prediction and measured operations, memory, and end-to-end latency. The present component measurements establish no general compression advantage or speedup.

The theoretical and empirical scope also determines how these findings can guide further work. The theorem preserves the full finite update of a sequential pre-norm block, including its nonlinear FFN. Predicting the FFN's response or assigning it a semantic role requires further analysis. Advection and diffusion comparisons depend on the token weights, parameterization, and feature basis; the construction does not provide a canonical alignment across models or arbitrary layerwise basis changes. Diffusion dissipates energy only under its stated sign condition.

The depth comparisons describe seven fixed pretrained decoders on 64-token windows, rather than a general law of training or independent training replicates. Comparing blocks on a saved input establishes similarity at that state: a distant block may encounter an atypical input, so the fixed comparison does not define a global distance between block functions. Appendix A.4 shows how the attention connection enters the nested computation of a post-norm block; the empirical findings concern pre-norm models. Broader architectural or task claims require comparisons that address these differences in inputs, geometry, and model populations.