

Disentangling Representation Evolution in Transformers through Directional Decomposition

Shwai He^{1,*} Haichao Zhang² Shen Yan³

¹University of Maryland, College Park ²Northeastern University ³ByteDance

shwaihe@umd.edu zhang.haich@northeastern.edu sheny@bytedance.com

Abstract

Transformer layers evolve representations through learned additive transformations that combine direction-preserving and direction-changing effects. We study this geometry by decomposing learned updates into parallel and perpendicular components. Across pretrained models, we find substantial learned parallel components, showing that direction-preserving behavior is not confined to the residual identity path. To test whether these components are redundant or behaviorally important, we apply the decomposition to residual-space updates and to value aggregation inside attention. Inference-time edits reveal a space-dependent asymmetry: value-space parallel edits preserve behavior more effectively than residual-space or perpendicular counterparts, especially when the direct self message is fixed and only the non-self aggregate is scaled. The decomposition also provides a compact, component-resolved view of compression error. During from-scratch pretraining, using parallel suppression as part of the forward computation lowers validation-loss trajectories and improves downstream averages in our evaluated runs, with the value-space variant strongest. Overall, this geometry connects update direction to editing robustness, compression diagnostics, and training-time intervention. Code is available in the [project repository](#).

1 Introduction

Transformer layers evolve representations through learned additive updates (Vaswani et al., 2017; Elhage et al., 2021) on top of residual paths (He et al., 2016). Geometrically, each update has two roles relative to the current representation: a parallel component confined to its span and a perpendicular component that redirects it. Decomposing learned

updates into these roles lets us test the functional sensitivity of each role through targeted editing.

Measurements across pretrained models show substantial components parallel to the current hidden state: beyond the exact carryover of the residual identity path, the learned updates themselves add collinear mass (He et al., 2016; Elhage et al., 2021).

These measurements alone do not show whether a collinear component preserves the current token’s own route or participates in context aggregation. Attention makes the distinction explicit: its value aggregate combines the direct self message with values routed from other positions, and decomposing the complete aggregate conflates the two. We therefore separate them in two settings. In *residual space*, the intervention edits the attention or MLP update added back to the hidden state. In *value space*, the primary exclude-self intervention keeps the direct self message fixed and decomposes only the non-self aggregate relative to the current token’s value. Scaling either component then tests the directional role of context aggregation.

These edits reveal a space-dependent asymmetry. Over the tested interval from full removal to the scale-one no-op, exclude-self value-space scaling remains closest to baseline and is the most robust tested intervention. Residual-space and perpendicular edits produce larger behavioral changes. Value-space edits act before the output projection relative to the current token’s value, while residual-space analogues act after projection relative to the hidden state. Robustness is therefore not a generic property of parallel components; it depends on the representation space, reference direction, and what the edit preserves.

Editing carries the central interventional analysis; we evaluate the same geometry beyond inference-time edits as well. For compres-

*Work done while Shwai He was an intern at ByteDance.

sion, we decompose the error between baseline and compressed-model updates, complementing prior work on post-training compression and layer-dropping analyses of transformer redundancy (Frantar et al., 2023; Lin et al., 2024; Sun et al., 2024; Men et al., 2024; He et al., 2026). Perpendicular-error rankings closely track total-error rankings, while parallel error is less discriminative; a norm-matched isotropic control reproduces much of this alignment. The decomposition therefore resolves compression error by direction while explaining why total magnitude remains a strong baseline. For training, we suppress parallel attention components in residual and value spaces during from-scratch pretraining. Across three small-model configurations, the value-space variant follows lower validation-loss trajectories. The 1.4B and 2.7B settings also improve average downstream scores, with the value-space variant giving the largest gain. The contributions are:

- This work reveals substantial direction-preserving components in transformer updates and maps their robustness across spaces.
- For attention updates, the value-space decomposition measures the aggregate against the current token’s own value, isolating the part routed from other tokens and explaining the stability gap between value-space and residual-space edits.
- Extending this geometry to compression and training provides component-resolved analysis and shows that self-value-parallel suppression changes validation-loss trajectories across training configurations.

2 Related Work

Geometric Control of Residual Updates. Residual connections preserve the input representation while allowing each block to add a learned transformation (He et al., 2016). In transformers, this creates a residual stream that carries information across layers and serves as the workspace for attention and feed-forward updates (Elhage et al., 2021). This preservation path is also a practical target for architectural scaling mechanisms: ReZero (Bachlechner et al., 2021), LayerScale (Touvron et al., 2021), and deep-transformer initialization or normalization schemes (Xiong et al., 2020; Wang et al., 2024) adjust residual branch scale to improve optimization and stability. Residual paths also mitigate degeneracy in deep self-attention, including rank

collapse (Dong et al., 2021). These works establish the importance of residual pathways and residual magnitude. We study a complementary geometric property of the learned update itself: whether it preserves the current direction or rotates the representation into a perpendicular component.

Context Modeling through the Attention Mechanism. Attention explicitly aggregates information across tokens (Vaswani et al., 2017), but attention mass also concentrates on special or persistent positions, as shown by attention-sink behavior in long-context inference (Xiao et al., 2024). Gated attention variants mitigate such sink behavior by modulating attention outputs (Qiu et al., 2025). However, even when attention sinks are reduced, large diagonal attention weights still route substantial mass back to the current token itself. Exclusive Self-Attention (XSA), introduced by Zhai (2026), directly addresses this self-directed pathway by modifying value flow aligned with the current token. We revisit this issue from a parallel/perpendicular geometric perspective, focusing on when attention updates preserve the current direction and when they introduce direction-changing information.

3 Background

3.1 Parallel and Perpendicular Decomposition

Parallel and perpendicular decomposition separates a vector into components along and orthogonal to a reference direction in the same representation space. Let $\Delta \in \mathbb{R}^d$ denote the vector being analyzed and let $\mathbf{z} \in \mathbb{R}^d$ be the reference vector. The unique components are:

$$\Delta_{\parallel} = \frac{\Delta \cdot \mathbf{z}}{\|\mathbf{z}\|^2} \mathbf{z} = \alpha \mathbf{z}, \quad \alpha = \frac{\Delta \cdot \mathbf{z}}{\|\mathbf{z}\|^2}, \quad (1)$$

$$\Delta_{\perp} = \Delta - \Delta_{\parallel}, \quad \Delta_{\perp} \perp \mathbf{z}. \quad (2)$$

The parallel component $\Delta_{\parallel}$ lies along the reference direction $\mathbf{z}$, whereas $\Delta_{\perp}$ lies in its $(d - 1)$ -dimensional orthogonal subspace and carries the direction-changing degrees of freedom.

This is the basic decomposition used throughout the paper. In residual space, Δ is a contribution added to the hidden state; in value space, it is the value aggregate formed inside attention.

3.2 Transformer Residual Updates

We consider transformer residual streams with L layers and hidden dimension d . Each residual sub-

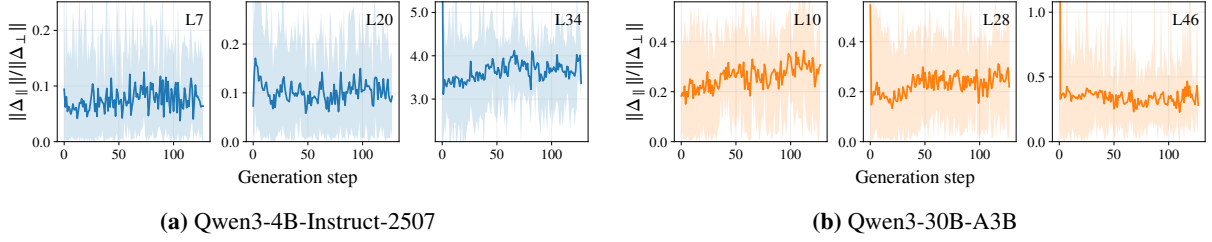


Figure 1: Parallel components persist across depth. For each accumulated attention-plus-MLP update, $\Delta_{t,||}^l$ is its projection along the incoming hidden state, whereas $\Delta_{t,\perp}^l$ is orthogonal to that state. Curves show $\|\Delta_{t,||}^l\|/\|\Delta_{t,\perp}^l\|$ at three sampled layers per model across generation steps; lines are means and bands span the observed range. Ratios above one indicate parallel dominance.

layer adds an update to the current representation:

$$\mathbf{z}^{l+1} = \mathbf{z}^l + \Delta^l. \quad (3)$$

The update Δ^l denotes the output of a multi-head self-attention (Attn) sub-layer, a feed-forward (FFN) sub-layer, or the accumulated attention-plus-MLP contribution. For an individual sub-layer, the reference is the hidden state immediately before its residual addition. For the joint view, the reference is the state entering attention and Δ^l is the net change after both attention and MLP. We write $\mathbf{z}_t^l$ for the reference state at token position $t \in \{1, \dots, T\}$ and Δ_t^l for the corresponding residual contribution.

For a residual update:

$$\Delta_{t,||}^l = \alpha_t^l \mathbf{z}_t^l, \quad \alpha_t^l = \frac{\Delta_t^l \cdot \mathbf{z}_t^l}{\|\mathbf{z}_t^l\|^2}. \quad (4)$$

Adding the parallel component therefore rescales the existing hidden state by $1 + \alpha_t^l$. The resulting representation is

$$\mathbf{z}_t^{l+1} = (1 + \alpha_t^l) \mathbf{z}_t^l + \Delta_{t,\perp}^l, \quad \Delta_{t,\perp}^l \perp \mathbf{z}_t^l. \quad (5)$$

To measure update geometry, we report the relative magnitude $\|\Delta_{t,||}^l\|/\|\Delta_{t,\perp}^l\|$ across layers. This ratio quantifies how much update magnitude lies along the current representation versus orthogonal directions; the sign of α_t^l indicates whether the parallel component points with or against the current representation.

Figure 1 reports this ratio across generation steps at three sampled layers for two Qwen3 models. The persistent parallel components motivate us to ask when they affect model behavior and in which representation spaces those effects are robust.

4 Method

We apply the decomposition from §3.1 to residual contributions and attention value aggregates. We

then scale their parallel and perpendicular components to test how each direction affects model behavior. Figure 2 summarizes the intervention sites and the associated diagonal view.

4.1 Residual-Level Decomposition

Instantiation. Set $\Delta = \Delta_t^l$ (the residual update of block f^l at token t) and $\mathbf{z} = \mathbf{z}_t^l$ (the pre-update hidden state). For layer and module profiles, we report calibration-corpus averages of the per-token residual parallel ratio:

$$r_t^l = r(\Delta_t^l, \mathbf{z}_t^l) = \frac{|\alpha_t^l| \|\mathbf{z}_t^l\|}{\|\Delta_{t,\perp}^l\|}, \quad (6)$$

$$\alpha_t^l = \frac{\Delta_t^l \cdot \mathbf{z}_t^l}{\|\mathbf{z}_t^l\|^2}.$$

This residual-level view serves as the shared observable for the rest of the paper. The resulting diagnostic is block-agnostic. We apply it to (i) the Attn sub-layer output, (ii) the FFN sub-layer output, and (iii) their joint effect at the transformer-block level.

4.2 Value-Space Decomposition

Inside attention, let $\mathbf{x}_s^{\text{attn}}$ denote the input to the value projection, including the model’s pre-attention normalization. Each source token s is mapped to a value vector $\mathbf{v}_s = W_V \mathbf{x}_s^{\text{attn}}$. For query token t , let $\mathcal{A}_{ts}$ denote the attention mixing operator on these values, which for standard attention is the scalar weight A_{ts} times the identity. The pre-output-projection aggregate is

$$\mathbf{o}_t = \sum_{s \leq t} \mathcal{A}_{ts} \mathbf{v}_s. \quad (7)$$

Within this space the natural reference is the current token’s own value $\mathbf{v}_t$: the direction $\mathbf{o}_t$ would take

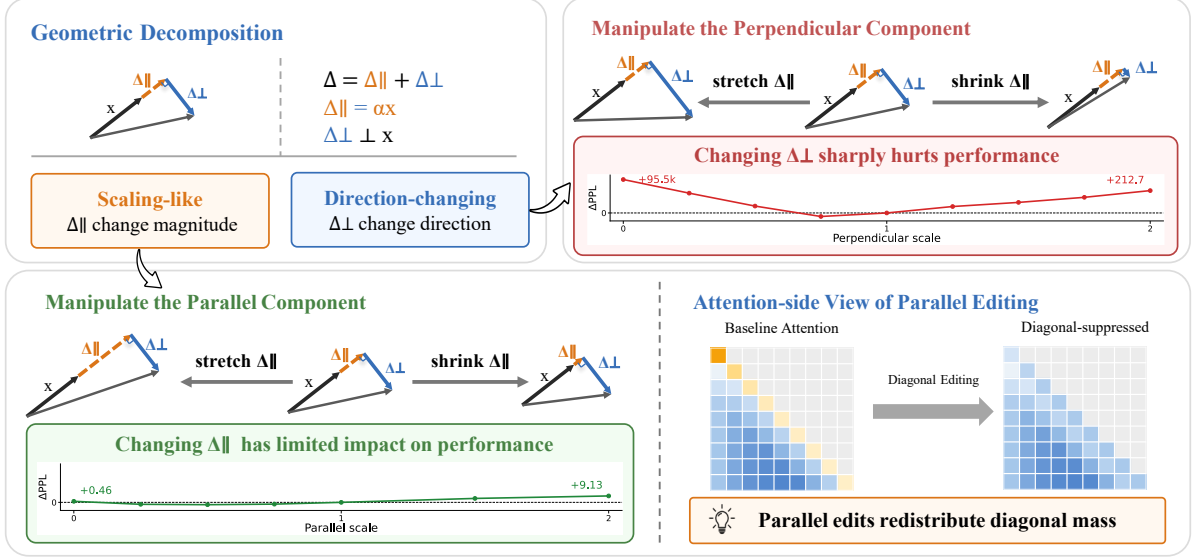


Figure 2: Geometric decomposition and component scaling. Residual updates and attention value aggregates are decomposed into parallel and perpendicular components and scaled independently; parallel-only scaling can be expressed as an attention-diagonal change. Labels report raw ΔPPL relative to the scale-one no-op (lower is better); curves are rescaled only for display.

if attention routed nothing from other positions.

$$\mathbf{o}_{t,\parallel} = \frac{\mathbf{o}_t^\top \mathbf{v}_t}{\|\mathbf{v}_t\|^2} \mathbf{v}_t, \quad \mathbf{o}_{t,\perp} = \mathbf{o}_t - \mathbf{o}_{t,\parallel}. \quad (8)$$

Residual-space geometry does not isolate this self-value-aligned structure because the output projection mixes the aggregate into the residual stream.

Direct Self-Message Preservation. The aggregate mixes two roles: the token’s own value passed straight through, and context routed from other positions. Editing both at once confounds them, so we split the aggregate and edit only the second:

$$\mathbf{o}_t = \underbrace{\mathcal{A}_{tt} \mathbf{v}_t}_{\mathbf{d}_t \text{ (self)}} + \underbrace{\sum_{s < t} \mathcal{A}_{ts} \mathbf{v}_s}_{\mathbf{c}_t \text{ (non-self)}} \quad (9)$$

$$\mathbf{c}_{t,\parallel} = \frac{\mathbf{c}_t^\top \mathbf{v}_t}{\|\mathbf{v}_t\|^2} \mathbf{v}_t, \quad \mathbf{c}_{t,\perp} = \mathbf{c}_t - \mathbf{c}_{t,\parallel}.$$

We scale only the parallel and perpendicular components of the non-self aggregate and then restore the unchanged self message:

$$\tilde{\mathbf{o}}_t^{\text{excl}} = \mathbf{d}_t + s^{(\parallel)} \mathbf{c}_{t,\parallel} + s^{(\perp)} \mathbf{c}_{t,\perp}. \quad (10)$$

We refer to this operation as *exclude-self* scaling: it neither masks the self-attention edge nor renormalizes the attention row.

Value-space decomposition applies only to attention aggregates. MLP intermediate activations do not form an attention aggregation space, so we analyze MLP outputs in residual space.

4.3 Component Scaling and Diagonal Form

To express both spaces in one form, let

$$\mathbf{y}_t = \sum_{s \leq t} \mathcal{A}_{ts} \mathbf{m}_s, \quad (11)$$

$$\mathbf{y}_{t,\parallel} = \frac{\mathbf{y}_t^\top \mathbf{r}_t}{\|\mathbf{r}_t\|^2} \mathbf{r}_t,$$

$$\mathbf{y}_{t,\perp} = \mathbf{y}_t - \mathbf{y}_{t,\parallel}.$$

where $\mathbf{m}_s$ is the aggregated message and $\mathbf{r}_t$ is the reference direction. The two attention-side instantiations are

$$\text{value: } \mathbf{y}_t = \mathbf{o}_t, \quad \mathbf{m}_s = \mathbf{v}_s, \quad \mathbf{r}_t = \mathbf{v}_t,$$

$$\text{res.: } \mathbf{y}_t = \Delta_t^{\text{Attn}}, \quad \mathbf{m}_s = W_O \mathbf{v}_s, \quad \mathbf{r}_t = \mathbf{z}_t. \quad (12)$$

Component scaling then uses the shared intervention

$$\tilde{\mathbf{y}}_t = s^{(\parallel)} \mathbf{y}_{t,\parallel} + s^{(\perp)} \mathbf{y}_{t,\perp}. \quad (13)$$

The no-op is $s^{(\parallel)} = s^{(\perp)} = 1$, and parallel removal sets $s^{(\parallel)} = 0, s^{(\perp)} = 1$. Exclude-self scaling applies the same form to $\mathbf{c}_t$ and restores $\mathbf{d}_t$ as in Eq. 10.

The same edit can also be read on the attention map itself, which connects it to work on self-directed attention mass (§2). For a parallel-only edit, the target change is the change in the projection of $\mathbf{y}_t$ along $\mathbf{r}_t$. Holding all off-diagonal message operators fixed, a diagonal update gives

$\tilde{\mathbf{y}}_t = \mathbf{y}_t + \Delta A_{tt} \mathbf{m}_t$. It realizes the requested scaling exactly when it satisfies the general operator constraint:

$$\mathbf{r}_t^\top \Delta A_{tt} \mathbf{m}_t = (s^{(\parallel)} - 1) \mathbf{y}_t^\top \mathbf{r}_t. \quad (14)$$

Thus the diagonal view is defined by its effect on the reference projection, not by a particular scalar parameterization. Restricting the diagonal operator to $\Delta A_{tt} = \Delta A_{tt} I$ gives

$$\Delta A_{tt} = (s^{(\parallel)} - 1) \frac{\mathbf{y}_t^\top \mathbf{r}_t}{\mathbf{m}_t^\top \mathbf{r}_t}. \quad (15)$$

Substituting Eq. 12 yields

$$\Delta A_{tt}^{\text{val}} = (s^{(\parallel)} - 1) \frac{\mathbf{o}_t^\top \mathbf{v}_t}{\|\mathbf{v}_t\|^2}, \quad (16)$$

$$\Delta A_{tt}^{\text{res}} = (s^{(\parallel)} - 1) \frac{(\Delta_{\text{Attn}}^\top)^\top \mathbf{z}_t}{(W_O \mathbf{v}_t)^\top \mathbf{z}_t}. \quad (17)$$

The value-space form is vector-exact because $\mathbf{m}_t = \mathbf{r}_t = \mathbf{v}_t$. The residual diagonal adjustment instead moves along $W_O \mathbf{v}_t$, not generally along $\mathbf{z}_t$; it matches the scalar projection change but is not vector-equivalent to residual component scaling. These expressions also show why equal scale coefficients do not define norm-matched interventions: the applied vector change depends on the message norm, the reference geometry, and, for value-space edits, the subsequent output projection. We therefore compare each edit with its own scale-one no-op and audit perturbation norms at a common post- W_O site. Appendix A gives the operator realization, removed-case forms, and the applied-output audit.

5 Experimental Setup

Inference-time interventions. For pretrained-model evaluations, edits operate only during the forward pass and leave model weights unchanged. Value-space results use the self-preserving scaling in Eq. 10 unless explicitly labeled otherwise; each experiment reports its edited layers and scales. Appendices A–B give implementation details.

Pretraining setup. For pretraining, we train three GPT-style configurations from scratch on standard next-token prediction, with baseline and full-aggregate parallel-removal variants. We additionally train 1.4B and 2.7B models for downstream evaluation (Table 3). Model configurations and the released settings are in Appendix E.

6 Directional Robustness of Updates

6.1 Component Scaling

Component-scaling results. Figure 3 sweeps $s^{(\parallel)}$ over $[-1, 3]$ and $s^{(\perp)}$ over $[0, 2]$ at three sites. On the parallel grid, exclude-self value-space scaling is stable within the unit interval: full removal changes PPL by only 0.46 and the largest absolute change is 1.32, whereas removing either residual-space parallel component increases PPL by about 21–22. The stable interval has a clear sign boundary: under reversal, the value-space delta reaches 188.77 at scale -1 . Perpendicular scaling produces larger changes: at scale 2, the deltas are 212.68 in value space, 14,757.67 for the residual attention branch, and 52.62 for the residual MLP branch. The sweep therefore shows both directional and representation-space asymmetries. Residual-space edits act on a complete learned branch, whereas exclude-self scaling isolates the non-self parallel component while preserving the remaining aggregate. The same directional asymmetry is already present inside the MLP: decomposing the post-gating carrier rather than the branch output leaves full parallel removal near baseline while perpendicular removal remains disruptive (Appendix D).

6.2 Value-Space vs. Residual-Space Editing

Robustness on language tasks. Table 1 reports the full-removal comparison on the seven non-generative tasks. Under the matched residual/diagonal protocol, Attn Para-Rem. changes the average by $-2.46 / -2.94$ points on the 1.7B/30B-A3B models, whereas Diag. Rem. changes it by $-22.44 / -32.97$ points. Under the value-space protocol, exclude-self removal stays within 1.51 points of the 1.7B baseline and 0.12 points of the 30B-A3B baseline, whereas editing the self contribution as well reduces the average by 8.2–10.0 points. Exclude-self removal is markedly more robust because it preserves the self edge and leaves the remaining attention normalization unchanged; hard diagonal removal is far more disruptive because it removes the self edge and renormalizes the remaining attention. Thus robustness depends on both the representation space and preservation of the direct self message.

To test whether self-preserving value-space scaling remains effective as the available context grows, Table 2 extends the comparison from 4k to 12k. Retaining half remains comparatively stable with the self term fixed, and the exclude-self variant con-

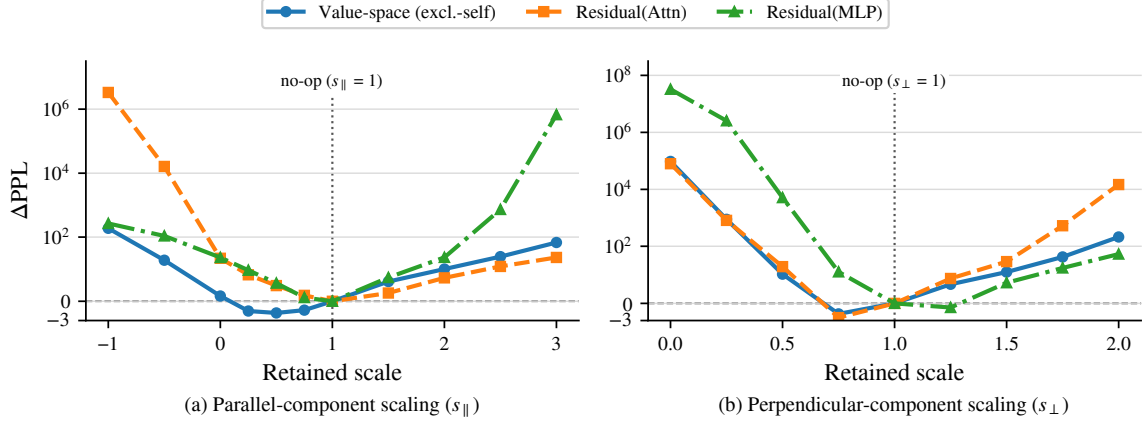


Figure 3: Perplexity under component scaling on Qwen3-1.7B. The left panel varies $s_{\parallel}$ with $s_{\perp} = 1$; the right panel varies $s_{\perp}$ with $s_{\parallel} = 1$. On WikiText-2, ΔPPL is edited minus the matched scale-one no-op, so lower is better. All decoder layers are edited; value-space curves use exclude-self scaling.

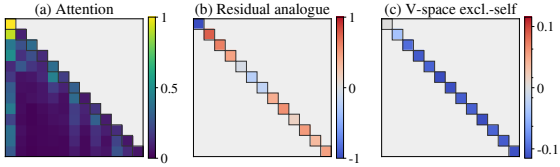


Figure 4: Effective diagonal views on Qwen3-4B at layer 19. For full parallel removal ($s_{\parallel} = 0$), the left panel shows head-6 attention weights; the middle shows the residual-space attention-map view, and the right shows the pre- W_O , head-6 exclude-self value-space change. The latter two color scales encode ΔA_{tt} ; off-diagonal entries are unchanged.

sistently outperforms full-aggregate scaling at the same retained scale. The gap to baseline nevertheless widens with context: under exclude-self scaling it grows from -0.22 at 4k to -2.42 at 12k when half the parallel component is retained, and from -11.27 to -17.27 under full removal. Self-preserving robustness therefore weakens as more context has to be aggregated. The same retained scale across seven models at 4k (Appendix Table 8) stays within 0.5 points of baseline on Gemma and Llama series.

Attention diagonal views for parallel editing. Figure 4 compares the value-space diagonal change with the residual-space attention-map view; Appendix Table 4 lists both coefficients.

The residual-space coefficients span -1.00 to 0.76 , whereas the selected head’s exclude-self value-space coefficients are non-positive, with the first-token coefficient exactly zero because there is no non-self context. The difference reflects their reference directions and the post- W_O mixing used

by residual-space editing. Coefficient magnitude is not intervention magnitude: a Qwen3-0.6B diagnostic measures both edits at the same post- W_O attention output and finds perturbation norms of 25.15% of the original branch norm for the residual-space edit, versus 49.10% for exclude-self value-space editing (Appendix Table 5). Residual-space fragility therefore does not follow from a larger applied perturbation in this comparison.

Having established the behavioral gap between value- and residual-space edits, we next inspect how their parallel-removal objectives appear as diagonal coefficient changes.

The all-layer averages also conceal substantial head and depth heterogeneity. Appendix C reports isolated-head sensitivities and joint all-head controls across four Qwen3 scales.

6.3 Compression Diagnostics

Compression modifies parameters rather than activations, inducing an error between baseline and compressed-model updates. We decompose this error to test whether its perpendicular component distinguishes compression settings more clearly than its parallel component.

At each layer, we define the compression error as $e = \Delta_{\text{comp}} - \Delta_{\text{base}}$ and decompose it relative to the baseline-update direction: $e_{\parallel} = \text{proj}_{\Delta_{\text{base}}} e$ and $e_{\perp} = e - e_{\parallel}$. Figure 6 reports the component norms normalized by $\|\Delta_{\text{base}}\|$ and averaged over fixed prompts.

Table 1: Task-level comparison of residual-, value-space, and diagonal controls. Entries are absolute scores (%); higher is better and Avg. is unweighted. Both value-space rows set $s_{||} = 0$. V-Para Rem. edits the full value aggregate; V-Excl.-self preserves $A_{tt}v_t$. Diag. Rem. zeros A_{tt} and renormalizes over non-self positions for $t > 0$; no edit is applied at $t = 0$. Attn Para-Rem. and Diag. Rem. share a separate matched evaluation, whose baseline averages are 61.61 and 73.65 for Qwen3-1.7B and Qwen3-30B-A3B, respectively; the displayed Baseline rows belong to the value-space comparison. Differences reported in the text are computed before rounding. Bold marks the best baseline/value-space result.

Model	Setting	OBQA	PIQA	Wino	ARC-C	HSwag	BoolQ	RTE	Avg.
Qwen3-1.7B	Baseline	37.20	72.47	60.69	53.07	60.16	77.65	70.04	61.61
	Attn Para-Rem.	38.20	72.31	56.59	50.09	58.74	71.38	66.79	<u>59.16</u>
	V-Para Rem.	34.40	69.48	54.85	41.38	54.04	66.27	53.43	<u>53.41</u>
	V-Excl.-self	39.20	71.76	60.54	49.15	59.88	74.46	65.70	<u>60.10</u>
	Diag. Rem.	26.20	51.36	49.17	24.49	26.66	40.70	55.60	<u>39.17</u>
Qwen3-30B-A3B	Baseline	44.80	80.69	70.24	69.97	77.79	88.53	81.95	73.42
	Attn Para-Rem.	44.00	78.73	66.85	65.10	73.87	88.04	78.34	<u>70.70</u>
	V-Para Rem.	41.20	76.22	64.56	59.04	69.57	70.89	62.45	<u>63.42</u>
	V-Excl.-self	44.60	79.87	72.45	68.34	78.13	86.36	83.39	<u>73.31</u>
	Diag. Rem.	30.60	57.67	48.46	26.96	34.78	38.59	47.65	<u>40.68</u>

Table 2: RULER on Llama-3.2-3B. Scores are means over all 13 tasks; parentheses give edited minus matched baseline, computed before rounding. Scale 0 removes the parallel component and 0.5 retains half. V-Full scales the whole value aggregate; V-Excl.-self preserves the direct self term and scales only the remainder.

Method	4k	8k	12k
Baseline	86.91	82.09	79.30
V-Full, $s_{ }=0.5$	84.81 (-2.10)	78.80 (-3.29)	74.76 (-4.53)
V-Excl.-self, $s_{ }=0.5$	86.69 (-0.22)	80.80 (-1.29)	76.88 (-2.42)
V-Full, $s_{ }=0$	60.99 (-25.92)	48.61 (-33.48)	39.17 (-40.12)
V-Excl.-self, $s_{ }=0$	75.64 (-11.27)	67.97 (-14.11)	62.02 (-17.27)

7 Training-Time Intervention

Perpendicular-error curves separate the four settings more clearly, while parallel-error curves interleave, and perpendicular error closely tracks total-error rankings. A norm-matched isotropic control reproduces much of that agreement, so the decomposition explains why total-error ranking is strong, most of its ordering being carried by perpendicular error, rather than displacing total magnitude as a ranking signal; Weighting the Wanda score by the residual-perpendicular component rather than by total magnitude stays within -0.24 to $+0.38$ points of the standard score across five models (Appendix Table 14), so the directional view refines the diagnostic without replacing the pruning criterion. Appendix F gives the branch-wise and norm-matched controls.

Post-hoc edits introduce a train–inference mismatch. We instead train from scratch with residual-

Table 3: Post-training benchmark summary. Avg. is the unweighted mean accuracy over ARC-Easy, BoolQ, HellaSwag, OpenBookQA, PIQA, and WinoGrande; Δ Avg is relative to each model’s matched baseline. Attn and V-Para denote residual-space attention and value-space parallel removal.

Model	Setting	Avg	Δ Avg
1.4B	Baseline	58.5	0.0
	Attn Para-Rem.	58.8	+0.3
	V-Para Rem.	59.2	+0.7
2.7B	Baseline	60.2	0.0
	Attn Para-Rem.	60.9	+0.7
	V-Para Rem.	61.7	+1.5

or value-space parallel removal active in both optimization and evaluation, letting the model adapt to the modified forward computation and testing whether optimization reallocates learning pressure across the retained updates. Figure 5 shows an early separation from baseline that persists through the final checkpoint across all three configurations. Table 3 shows that both interventions improve the downstream average, while V-Para Rem. gives the strongest result at both scales, with gains of 0.7 and 1.5 points. Together with the loss curves, these results show that parallel suppression affects both the training trajectory and the resulting checkpoint, with a larger downstream change for the value-space variant.

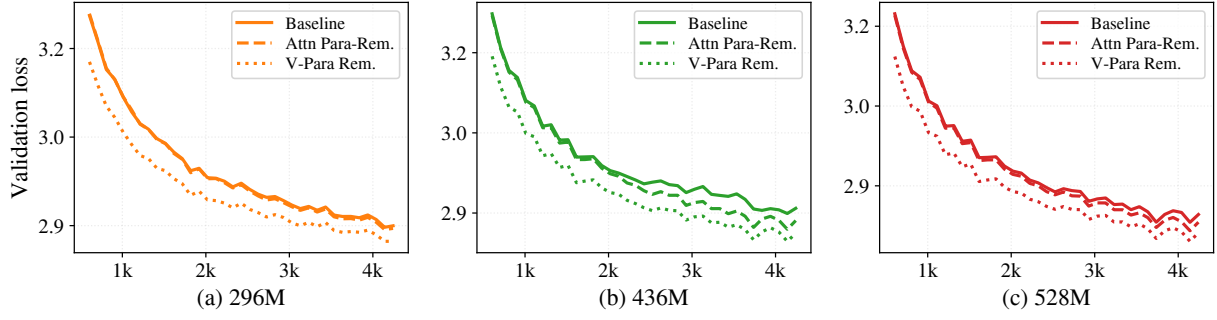


Figure 5: Training-time parallel removal lowers validation loss. OpenWebText (Gokaslan et al., 2019) loss trajectories compare one matched run of baseline, Attn Para-Rem., and V-Para Rem. Both edits set $s_{\parallel} = 0$ and retain the perpendicular component.

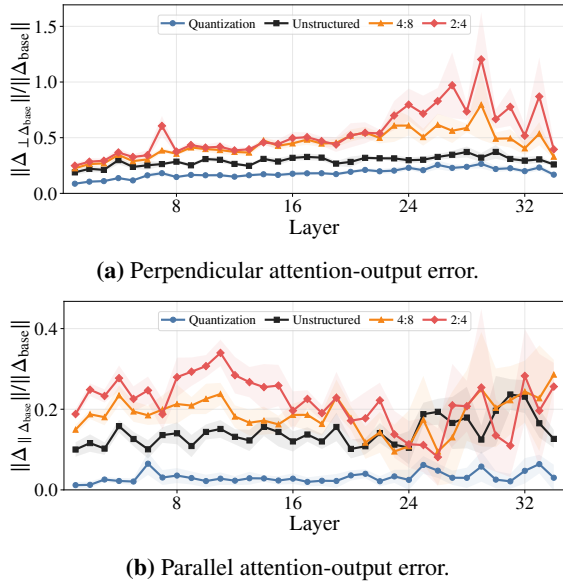


Figure 6: Directional compression error. Curves are layerwise means and bands are ± 1 SD over 12 fixed prompts, comparing compressed and baseline Qwen3-4B attention updates. The upper and lower panels show perpendicular and parallel error, each normalized by the full baseline-update norm (higher is more distortion). Settings are 4-bit AWQ and 50% Wanda pruning with unstructured, 4:8, or 2:4 masks.

Appendix E gives task-level scores and a gated-scaling control, where fixed removal retains the lowest training loss.

8 Discussion

This geometry supports compression diagnosis and training-time allocation beyond the intervention analysis. Across attention and MLP controls, parallel robustness depends on the representation space, reference direction, and preserved computation. Exclude-self editing keeps the direct self message fixed while isolating contextual aggregation; MLP

controls extend the analysis inside the branch.

Compression diagnostics. Directional decomposition complements total error by showing how compression changes each layer’s update direction. Here the perpendicular component gives the clearest separation among methods.

Training-time allocation. Training from scratch uses the modified forward computation during both optimization and evaluation, allowing adaptation from initialization and avoiding a train–inference mismatch. Both variants improve the downstream average, with the largest gains from V-Para Rem.

9 Conclusion

Directional editing shows how robustness depends on representation space, reference direction, and preservation of the direct self message. Exclude-self value-space scaling probes contextual aggregation, and MLP controls extend the geometry inside the branch. In from-scratch pretraining, parallel suppression lowers validation loss and improves downstream averages, most strongly in value space. Directional geometry therefore provides a unified tool for editing and training under a modified forward computation.

Limitations

Our analysis and experiments focus on the models, language tasks, and intervention settings studied here. Although the same geometric picture is consistent across the evaluated models, its behavior under other architectures, training regimes, or multimodal task distributions remains to be established. We also treat the proposed edits primarily as analysis tools; turning them into practical training or inference methods requires further validation.

Ethical Considerations

Directional decomposition provides a concrete tool for auditing representation change and screening compression or interventions before downstream evaluation. This screening requires no retraining and is designed to complement task-specific testing and human review. Responsible deployment should evaluate prompt- and context-specific behavior as well as bias, privacy, security, and high-stakes use. Access controls and downstream monitoring are appropriate when editing capabilities could be used for targeted manipulation.

References

- Thomas C. Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, G. Cottrell, and Julian McAuley. 2021. [ReZero is all you need: Fast convergence at large depth](#). In *Conference on Uncertainty in Artificial Intelligence*.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. [PIQA: Reasoning about physical commonsense in natural language](#). In *AAAI Conference on Artificial Intelligence*.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). *ArXiv*, abs/1905.10044.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *ArXiv*, abs/1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *ArXiv*, abs/2110.14168.
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *ICLR*.
- Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. [Attention is not all you need: Pure attention loses rank doubly exponentially with depth](#). In *ICML*.
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *International Conference on Learning Representations*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Gemma Team. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Gemma Team. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. [Openwebtext corpus](#). <http://Skyllion007.github.io/OpenWebTextCorpus>.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *CVPR*.
- Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. 2026. [Uncovering the redundancy in transformers via a unified study of layer dropping](#). *Transactions on Machine Learning Research*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krman, Shantanu Acharya, Dima Rekeshe, Fei Jia, and Boris Ginsburg. 2024. [RULER: What’s the real context size of your long-context language models?](#) *ArXiv*, abs/2404.06654.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration](#). In *Proceedings of Machine Learning and Systems*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Llama Team. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. [ShortGPT: Layers in large language models are more redundant than you expect](#). *arXiv:2403.03853*.

- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. [Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free](#). *Preprint*, arXiv:2505.06708.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [An adversarial winograd schema challenge at scale](#). In *AAAI Conference on Artificial Intelligence*.
- Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. [A simple and effective pruning approach for large language models](#). In *International Conference on Learning Representations*.
- Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. 2021. [Going deeper with image transformers](#). In *ICCV*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NeurIPS*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Black-boxNLP@EMNLP*.
- Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2024. [DeepNet: Scaling transformers to 1,000 layers](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6761–6774.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. [Efficient streaming language models with attention sinks](#). In *ICLR*.
- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. [On layer normalization in the transformer architecture](#). In *International Conference on Machine Learning*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Annual Meeting of the Association for Computational Linguistics*.
- Shuangfei Zhai. 2026. [Exclusive self attention](#). *Preprint*, arXiv:2603.09078.

A Intervention Details

Attention-diagonal editing. Equation 14 defines the diagonal operator by its projection change. One minimum-Frobenius-norm realization is

$$\Delta \mathcal{A}_{tt}^* = (s^{(\parallel)} - 1) \frac{\mathbf{y}_t^\top \mathbf{r}_t}{\|\mathbf{m}_t\|^2 \|\mathbf{r}_t\|^2} \mathbf{r}_t \mathbf{m}_t^\top.$$

For a structured operator family, the same construction uses the Frobenius projection of $\mathbf{r}_t \mathbf{m}_t^\top$ onto the admissible operator subspace. Table 4 gives the resulting scalar forms. For the exclude-self variant used in our primary value-space results, Eq. 10 instead applies the projection to $\sum_{s < t} A_{ts} \mathbf{v}_s$ and then restores the unchanged direct term $A_{tt} \mathbf{v}_t$.

Table 4: Attention-diagonal forms for component scaling. The value-space form is exact; the residual form matches only the scalar projection constraint. “Removed case” sets $s^{(\parallel)} = 0$ while retaining the perpendicular component.

Quantity	Form
<i>Full-aggregate value space</i>	
Diagonal change	$\Delta A_{tt} = (s^{(\parallel)} - 1) \frac{\mathbf{v}_t^\top \mathbf{v}_t}{\ \mathbf{v}_t\ ^2}.$
Removed case	$\tilde{A}_{tt} = - \sum_{s < t} A_{ts} \frac{\mathbf{v}_s^\top \mathbf{v}_t}{\ \mathbf{v}_t\ ^2}.$
<i>Residual space</i>	
Diagonal change	$\Delta A_{tt} = (s^{(\parallel)} - 1) \frac{(\Delta_t^{\text{Attn}})^\top \mathbf{z}_t}{(W_O \mathbf{v}_t)^\top \mathbf{z}_t}.$
Removed case	$\tilde{A}_{tt} = A_{tt} - \sum_{s < t} A_{ts} \frac{(W_O \mathbf{v}_s)^\top \mathbf{z}_t}{(W_O \mathbf{v}_t)^\top \mathbf{z}_t}.$

Applied strength of the two edits. Table 5 audits both Full No-Para edits at a common post- W_O site. The exclude-self value-space edit applies the larger perturbation on every measure, so its smaller behavioral cost does not follow from a gentler intervention.

Table 5: Applied strength of attention Full No-Para interventions. Full No-Para sets $s_{\parallel} = 0$. Ratios are measured at the summed attention output on Qwen3-0.6B; value-space attention preserves the self term and is mapped through W_O . Perturbation and retained norm compare $\|\tilde{y} - y\|$ and $\|\tilde{y}\|$ with $\|y\|$; energy is the sample average of $\|\tilde{y} - y\|^2 / \|y\|^2$.

Geometry	Perturbation norm	Perturbation energy	Retained norm
Attention residual-space	25.15%	9.323%	95.01%
Attention exclude-self value-space	49.10%	25.89%	83.36%

Forward-hook implementation. We apply Eq. 13 through forward hooks that replace only the selected tensor; all subsequent computation is unchanged. Table 6 reports the measured cost.

Table 6: Measured pre- W_O projection overhead. Wall-clock increases relative to the unedited baseline are measured on one RTX 6000 Ada. BF16 inference uses batch 2: prefill has length 2,048 and decode generates one token from a 512-token static cache with CUDA Graph replay. Training uses the 0.7B model with batch 1 and length 2,048. The three inference values are ordered as Qwen3-0.6B/4B/14B.

Workload	Optimized overhead
Prefill	2.47% / 1.25% / 1.11%
Static graph decode	0.83% / 0.36% / 0.25%
Training (0.7B)	3.39%

The fused kernel combines the dot product, norm, and projection update with FlashAttention-2 (Dao, 2024).

B Evaluation Details

Evaluated models. The editing and diagnostic experiments use Qwen3 (Qwen Team, 2025), and individual figures and tables name the evaluated scale. Qwen3-4B results use the base checkpoint, except Figure 1, which uses Qwen3-4B-Instruct-2507. Long-context evaluation uses Llama-3.2-3B (Llama Team, 2024).

Benchmark evaluation. Benchmark and RULER tasks are executed through the Language Model Evaluation Harness (Gao et al., 2024). RULER (Hsieh et al., 2024) provides the long-context evaluation. Task definitions follow the original sources for CoQA (Reddy et al., 2019), DROP (Dua et al., 2019), TruthfulQA (Lin et al., 2022), and the remaining benchmarks (Clark et al., 2018, 2019; Zellers et al., 2019; Mihaylov et al., 2018; Bisk et al., 2020; Wang et al., 2018; Sakaguchi et al., 2020; Hendrycks et al., 2021; Cobbe et al., 2021). Table 7 gives the default few-shot setting and metric for each task.

Table 7: Default evaluation task settings. Acc. denotes accuracy; Norm denotes length normalization.

Task	Few-shot	Metric
ARC-Challenge	25	Acc. (Norm)
ARC-Easy	0	Acc. (Norm)
BoolQ	0	Acc.
CoQA	0	F1
DROP	0	F1
GSM8K	5	Exact match
HellaSwag	10	Acc. (Norm)
MMLU	5	Acc.
OpenBookQA	0	Acc. (Norm)
PIQA	0	Acc. (Norm)
RTE	0	Acc.
TruthfulQA-MC1/2	0	Acc.
TruthfulQA generation	0	Rouge-L
WinoGrande	5	Acc.

C Additional Editing Results

Zero-shot parallel editing. Table 8 gives the complete cross-model RULER-4k comparison. The Gemma-3 models use the released instruction-tuned checkpoints (Gemma Team, 2025).

Table 8: Cross-model retained-scale RULER-4k evaluation. V-Excl.-self preserves the direct self contribution and sets $s_{||} = 0.5$ for the non-self aggregate; Δ is edited minus matched baseline and is computed before display rounding.

Model	Baseline	V-Excl.-self, $s_{ }=0.5$	Δ
Gemma-3-1B-IT	68.0	62.8	-5.3
Gemma-3-12B-IT	94.0	94.1	+0.1
Llama-3.2-3B	86.9	86.7	-0.2
Qwen3-0.6B	79.8	75.1	-4.8
Qwen3-1.7B	88.1	84.9	-3.2
Qwen3-4B	92.6	92.4	-0.2
Qwen3-8B	94.2	93.7	-0.5

Head-level localization. We test localization across every decoder layer of four Qwen3 scales by comparing all isolated query-head edits with a joint all-head edit. Figure 7 includes every head without significance-based selection.

Layer 0 gives the largest positive joint effect at every scale. Beyond layer 0, the joint effect is strongly depth-specific and often extends beyond the corresponding single-head range.

D MLP-Internal Geometry

Residual-space editing observes only the final MLP output, so it does not show whether the same directional asymmetry is already present before the down projection. Inside a gated MLP, the post-gating carrier and down-projected output are

$$\begin{aligned}
 h &= \phi(W_{\text{gate}x}) \odot W_{\text{up}}x, \\
 y &= \sum_g y_g, \\
 y_g &= W_{\text{down}}[:, I_g]h_g,
 \end{aligned} \tag{18}$$

where $h_g = h[I_g]$. We choose residual-width groups, so h_g , y_g , and x have the same dimensionality. This exposes two internal editing sites: each carrier h_g is decomposed relative to x before the down projection, and each projected contribution y_g is decomposed relative to its carrier h_g before the group contributions are summed. These controls distinguish geometry already present in the gated intermediate representation from geometry introduced by the down projection. Table 9 reports fixed-token loss on Qwen3-0.6B. At the two internal sites, full parallel removal changes loss by at

most 0.0301, whereas perpendicular removal increases it by more than 13 points.

Table 9: MLP-internal component removal. Values are edited-minus-no-op loss on fixed C4 tokens (lower is better). Full No-Para or No-Perp sets the named retained scale to zero; every no-op matches baseline.

Geometry	Full No-Para	Full No-Perp
Residual MLP (y relative to x)	+0.1165	+15.9567
Grouped h relative to x	+0.0301	+13.4184
Grouped y_g relative to h_g	+0.0037	+13.3324

Table 10 extends the check beyond fixed-token loss. Full No-Para preserves the multiple-choice averages at both scales and stays within 0.91 points of baseline on GSM8K, DROP, and CoQA; TruthfulQA-generation Rouge-L drops by 3.37 points on Qwen3-0.6B but is unchanged on Qwen3-1.7B.

Table 10: Downstream controls for MLP-internal geometry. Core-MC averages nine 0-shot benchmarks; MMLU is 5-shot. GSM8K-S/F are strict/flexible exact match, TQA R-L is TruthfulQA-generation Rouge-L, and DROP/CoQA use F1.

(a) Multiple-choice controls					
Model	Setting	Core-MC	MMLU		
Qwen3-0.6B	Baseline	50.94	52.40		
	Full No-Para	50.98	52.66		
	Full No-Perp	36.62	26.89		
Qwen3-1.7B	Baseline	55.19	60.21		
	Full No-Para	55.28	60.39		
	Full No-Perp	37.35	24.40		
(b) Generation and question-answering controls					
Model	Setting	GSM8K-S	GSM8K-F	TQA R-L	DROP CoQA
Qwen3-0.6B	Baseline	49.43	50.42	18.98	7.76 74.24
	Full No-Para	49.96	50.72	15.61	7.66 74.04
Qwen3-1.7B	Baseline	68.76	68.92	48.34	7.32 77.06
	Full No-Para	68.08	68.01	48.32	7.44 77.69

E Pretraining Configuration Details

Table 11 resolves the summary averages into task-level scores. The trajectories in Figure 5 use OpenWebText, and the 1.4B/2.7B runs use FineWeb100BT from FineWeb (Penedo et al., 2024). In every setting, the full-aggregate value-space intervention is applied at every layer.

The OpenWebText configurations labeled 296M, 436M, and 528M use (layers, heads, width) of (14, 16, 1024), (16, 18, 1152), and (18, 20, 1280); the 1.4B and 2.7B configurations use (24, 16, 2048)

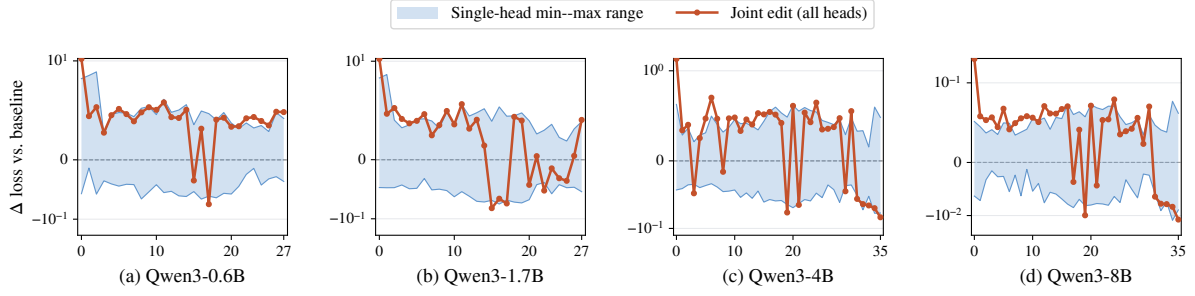


Figure 7: All-layer head sensitivity across Qwen3 scales. At each layer, orange shows joint full removal from all heads and the blue band spans the minimum and maximum isolated single-head effects; it is not a confidence interval. Δ loss is edited minus baseline over 128 fixed C4 texts (positive is worse), shown on symmetric-log axes.

Table 11: Task-wise post-training benchmark scores. Avg. is unweighted and Δ Avg uses the same-scale baseline. Base, Attn, and V-Para denote baseline, residual-space attention parallel removal, and full-aggregate value-space parallel removal.

Model	Setting	ARC-E	BoolQ	HSwag	OBQA	PIQA	WinoGr	Avg	Δ Avg
1.4B	Base	56.0	65.1	60.5	34.7	75.8	58.7	58.5	0.0
	Attn	57.1	63.9	61.3	35.3	76.0	59.1	58.8	+0.3
	V-Para	58.3	62.8	62.1	35.9	76.3	59.8	59.2	+0.7
2.7B	Base	58.4	61.2	66.0	37.1	76.5	61.8	60.2	0.0
	Attn	59.3	62.8	66.7	37.7	77.0	62.1	60.9	+0.7
	V-Para	60.4	64.4	67.1	38.2	77.6	62.6	61.7	+1.5

and (32, 20, 2560). The larger runs use eight DDP processes, approximately 104.9B tokens, 2,048-token contexts, 200K updates, and global batch 256. For the 1.4B model, the per-process micro-batch is 2, the accumulation factor is 128, and the peak/minimum learning rates are $4 \times 10^{-4}/4 \times 10^{-5}$; the corresponding settings for the 2.7B model are 1, 256, and $3 \times 10^{-4}/3 \times 10^{-5}$. Shared settings are a 2K-step warm-up with cosine decay, AdamW $\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping at 1.0, no dropout, seed 1337, and 200-batch evaluation every 1K updates.

Gated parallel-scaling diagnostic. Figure 8 compares baseline, fixed V-Para Rem., and gated parallel-scaling in the 1.4B run. Fixed removal has the lowest training-loss trajectory among the three settings.

F Extended Compression Geometry

We use the normalization from Figure 6. Figure 9 extends the breakdown to the accumulated update and the MLP output, and Table 12 reports the layerwise correlation between total and perpendicular attention-output error.

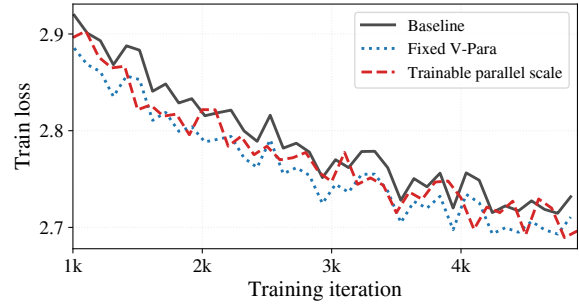


Figure 8: Parallel-scale adjustment during 1.4B pretraining. Single-run FineWeb100BT training loss (lower is better) is plotted against update. Fixed V-Para sets the full-aggregate $s_{||} = 0$; trainable scaling instead predicts a tokenwise parallel scale.

Table 12: Attention-output total error is mostly aligned with perpendicular error. On Qwen3-4B and 12 fixed prompts, r and ρ are layerwise Pearson and Spearman correlations between normalized total error $\|e\|/\|\Delta_{\text{base}}\|$ and its perpendicular or parallel component. Perp./Total and Para./Total are mean component-to-total ratios across layers.

Setting	$r(e, e_{\perp})$	$\rho(e, e_{\perp})$	$\rho(e, e_{ })$	Perp./Total	Para./Total
Quantization	0.998	0.990	0.596	0.980	0.168
Unstructured	0.970	0.909	0.810	0.891	0.439
4:8	0.986	0.986	-0.135	0.887	0.415
2:4	0.991	0.979	-0.432	0.875	0.425

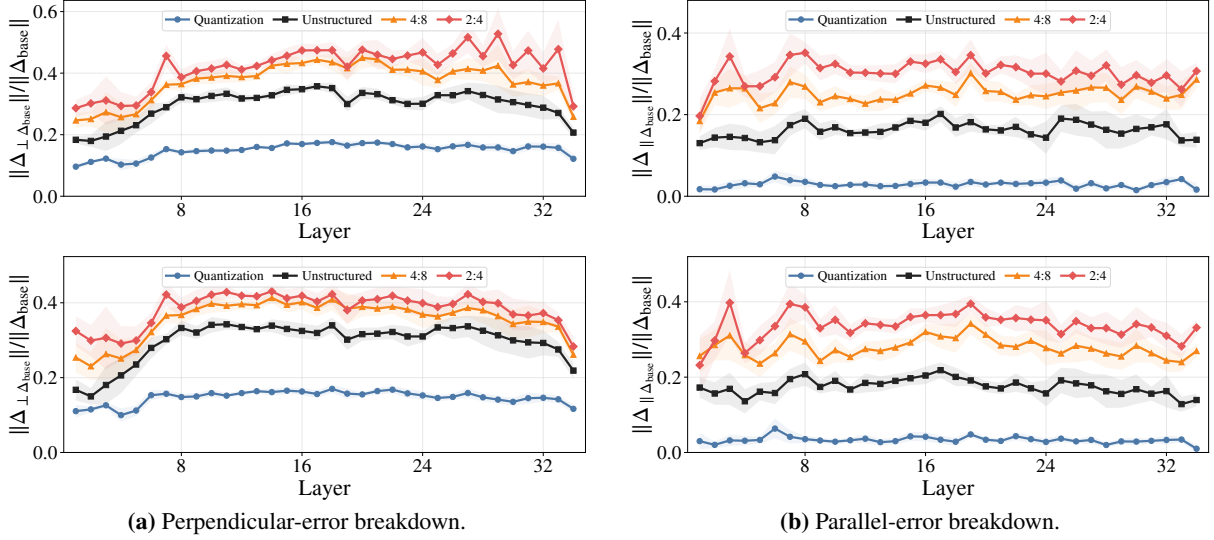


Figure 9: Compression-geometry breakdown across branches. The upper and lower groups show perpendicular and parallel compressed-minus-baseline error for the accumulated attention-plus-MLP update and the MLP output, each normalized by the full baseline-update norm. Curves are layerwise means and bands are ± 1 SD over 12 fixed prompts on Qwen3-4B. Settings are 4-bit AWQ and 50% Wanda pruning with unstructured, 4:8, or 2:4 masks.

Relation to Layer Dropping. The parallel/perpendicular view also decomposes a cosine-based diagnostic used for layer dropping (He et al., 2026). Let x be the incoming representation and let a layer update decompose as $\Delta = \Delta_{\parallel} + \Delta_{\perp}$, with $\Delta_{\parallel} = \alpha x$ and $\|\Delta_{\perp}\| = \beta \|x\|$. Then

$$\cos(x, x + \Delta) = \frac{1 + \alpha}{\sqrt{(1 + \alpha)^2 + \beta^2}}.$$

The diagnostic depends on $|\beta/(1 + \alpha)|$. In the positive-scale, small-angle regime, $1 - \cos(x, x + \Delta) \approx \frac{1}{2} (\beta/(1 + \alpha))^2$. Figure 10 reports the baseline update magnitudes.

Overall pruning-error rankings. Table 13 compares total- and perpendicular-error rankings across 21 model-space groups with uniform and norm-matched controls.

Table 13: Aggregate controls for total/perpendicular candidate overlap. Top- k is the fraction of coordinates shared by total-error and perpendicular-error rankings when selecting the largest 1% or 10%. Values average 21 model-space groups; controls use ten uniform-random seeds or a norm-matched isotropic direction.

Ranking comparison	Top-1%	Top-10%
Observed total vs. perpendicular	99.40%	99.86%
Uniform-random candidates	0.95%	10.01%
Isotropic norm-matched direction	99.94%	99.95%

For matched 50% global unstructured pruning, component-weighted Wanda multiplies each score by $|r_j|/\|\mathbf{r}\|$ or $\sqrt{1 - (r_j/\|\mathbf{r}\|)^2}$, averaged over

128 length-2,048 C4 sequences. Table 14 reports the eight-benchmark mean, including Gemma-2 (Gemma Team, 2024).

Table 14: Component-weighted Wanda quality. Eight-benchmark means after matched 50% global unstructured pruning; higher is better. Delta is residual-perpendicular minus Wanda before rounding.

Model	Wanda	Residual-perp.	Delta	Residual-para.
Gemma-2-2B	55.73	55.77	+0.04	54.72
Llama-3.2-1B	43.27	43.43	+0.16	42.67
Llama-3.2-1B-Inst.	45.38	45.14	-0.24	44.78
Qwen3-1.7B	54.17	54.55	+0.38	51.76
Qwen3-4B	61.45	61.56	+0.12	60.66

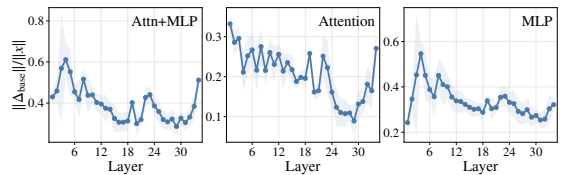


Figure 10: Baseline layer updates are small relative to the residual stream. Mean ± 1 SD of $\|\Delta_{\text{base}}\|/\|x\|$ across 12 prompts for Qwen3-4B layers 1–34, shown for the accumulated attention-plus-MLP update, attention output, and MLP output. Values below one indicate updates smaller than the incoming residual stream.