RESIBAND: RECOVERING EVIDENCE CONTRAST UNDER LOCAL COMMON-MODE DOMINANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Many language-model tasks require identifying and combining a few pieces of relevant evidence in distractor-rich contexts. Models may use this evidence yet fail to distinguish it reliably from competing information. We investigate this difficulty in the residual stream. Controlled evidence-relocation experiments show that support–distractor differences remain present but are small relative to a dominant component shared by nearby residual-stream states. This imbalance is associated with weaker evidence discrimination in subsequent attention layers. We introduce RESIBAND, a causal residual operator that subtracts a short-horizon exponential average, smooths the remaining variation over a longer horizon, and maps the filtered signal into a learned low-rank residual correction. The fixed filter exposes variation relative to the local background, while the learned mapping adapts the correction to the task with pretrained weights frozen. Under a shared configuration across backbones, RESIBAND improves average F1 across three multi-hop QA benchmarks over LoRA by 19.30, 10.95, and 1.93 percentage points on Qwen3-1.7B, Qwen3-8B, and Llama-3.1-8B, respectively, while requiring fewer trainable parameters. Mechanistic analyses and residual interventions indicate that the learned corrections improve discrimination between supporting evidence and distractors. These results support temporal filtering and learned residual correction for better use of available evidence.

1 INTRODUCTION

Large language models increasingly need to integrate evidence from retrieved passages, multi-turn interactions, and intermediate reasoning traces. In these settings, a small part of the available context may determine the correct answer. Reliable generation requires the model to retain that evidence and distinguish it from nearby context and plausible distractors. This distinction matters in grounded question answering and multi-document reasoning, where the answer must remain tied to information supplied earlier in the context (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Lewis et al., 2020).

Causal self-attention provides access to preceding positions (Vaswani et al., 2017), but does not ensure use of relevant evidence. A support passage may be separated from the query and surrounded by plausible but irrelevant information (Liu et al., 2024; Hsieh et al., 2024; Modarressi et al., 2025). To influence the answer, its representation must remain distinguishable from competing context during subsequent computation. We therefore ask whether the residual stream provides a sufficiently discriminative signal for relevant evidence to gain an advantage over matched distractors.

We study this question with matched evidence controls in the residual stream. Nearby token states share a substantial component associated with their local context, while the measured difference between gold support and a matched distractor is small relative to that component. This imbalance is consistent with a conditioning problem for subsequent evidence routing. The support can remain in the context and contribute to the answer, even when its representation offers only a weak advantage over a comparable distractor.

Two complementary diagnostics motivate this account. On annotated multi-hop question answering examples, removing distant support K/V states reduces gold-answer likelihood more than removing matched irrelevant states. The supports therefore remain behaviorally important. At the same time, the raw residual stream has strong local common-mode concentration and weak gold-versus-control

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.

contrast, while moving identical support content farther from the query reduces its native routing advantage and final answer margin. The problem is thus not locality by itself. It is the loss of task-relevant evidence contrast beneath a dominant local background.

We introduce RESIBAND, a causal residual operator that addresses this conditioning problem. The operator subtracts a short-horizon causal average to reduce locally shared components. It then aggregates the remaining variation over a longer horizon to stabilize signals that persist across tokens. A learned low-rank map writes selected directions back into the residual stream. Subsequent frozen Transformer blocks process corrected states through native attention and feed-forward computations.

The two parts of the operator serve different purposes. The fixed filtering stage changes the signal available to the learned map, but it does not identify the answer span on its own. The learned write adapts that signal to the task without replacing native attention. We hypothesize that this combination improves the model's discrimination of relevant evidence against comparable distractors.

We test this hypothesis at three levels. First, we evaluate question answering under a shared all-layer configuration on Qwen3-1.7B, Qwen3-8B, and Llama-3.1-8B. We compare RESIBAND with the frozen Base model and a LoRA (Hu et al., 2022) baseline trained on the same data for the same number of steps. Second, operator ablations compare the complete filter with its high-pass and low-pass components and with standard causal attention pooling. Inference-time depth ablations examine the contributions of jointly trained branches across backbones. Third, matched interventions and routing measurements test whether improved signal conditioning is accompanied by changes in the learned write, native evidence competition, and the answer margin.

Our contributions are:

- We identify a residual-stream imbalance with matched evidence controls. Gold-support differences are small relative to locally shared components, and this imbalance is associated with weaker evidence discrimination in subsequent routing.
- We introduce RESIBAND, a causal residual operator that combines local subtraction, longer-horizon aggregation, and a learned low-rank write to correct task-relevant representations while keeping pretrained weights frozen.
- We demonstrate behavioral gains across three backbones and use matched diagnostics and interventions to examine the link between residual conditioning, native evidence routing, and answer-level evidence discrimination.

2 RELATED WORK

Evidence availability versus evidence use. Long-context models and benchmarks have made clear that a large nominal context window does not guarantee reliable use of its contents. Accuracy can fall when the same evidence is moved away from the query, placed in the middle of a prompt, or surrounded by plausible distractors (Liu et al., 2024; Hsieh et al., 2024; Modarressi et al., 2025). These results establish the behavioral gap, but do not by themselves distinguish forgotten evidence from evidence that remains available yet loses later competition. We target the latter case. Matched K/V deletion first verifies that nonlocal support is still load-bearing; controlled relocation then tests how its residual-state contrast changes relative to the query-local background. The resulting diagnosis is therefore not simply that distance hurts, but that required evidence can reach the answer pathway under poor conditioning.

Representation geometry and evidence competition. Prior work has studied global representation degeneration and routing bottlenecks through rank collapse, oversmoothing, attention sinks, and over-squashing (Dong et al., 2021; Noci et al., 2022; Guo et al., 2023; Nguyen et al., 2023; Dovonon et al., 2024; Xiao et al., 2024; Barbero et al., 2024; 2025). Our focus is narrower and task-conditional. We do not require the hidden states to collapse globally; the relevant failure is that the gold–distractor difference becomes small relative to a component shared by the query's local context. This evidence-relative view directly motivates the statistic and matched controls in Section 3. Because attention weights alone need not reveal causal use (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), we follow the contrast through projected scores, normalized attention, and value delivery, and pair those measurements with answer-margin, deletion, and residual interventions. These measurements complement attention maps with intervention evidence.

Changing evidence routing versus conditioning its input. Methods for improving how models use contextual evidence often change how information is stored, routed, or aggregated. Sparse and structured attention modifies token interaction (Beltagy et al., 2020; Ding et al., 2023; Yuan et al., 2025); recurrence and compressed memory extend persistent state (Dai et al., 2019); state-space and hybrid models provide alternative sequence mixers (Gu et al., 2022; Gu & Dao, 2023; Lieber et al., 2024); and retrieval augmentation supplies evidence from an external store (Lewis et al., 2020). These approaches are essential when evidence is unavailable or the aggregation mechanism must be replaced. Our setting exposes a complementary failure: the evidence is already present and causally useful, but its task-specific variation is weak relative to the residual background consumed by the frozen model. RESIBAND therefore keeps the pretrained memory and attention machinery intact. It conditions their input by exposing band-limited temporal variation and learning a bounded residual write before subsequent native computation. This is the central distinction and the reason the intervention is needed even when the full context is already accessible.

3 LOCAL BACKGROUND OBSCURES EVIDENCE CONTRAST

Relevant evidence can remain available to a model yet compete weakly with distractors. We study this failure under controlled evidence relocation. When support moves beyond the query's local neighborhood, its residual-state contrast with matched distractors can become small relative to the component shared by query-local states. We first measure this imbalance, then use matched K/V deletion to test whether the displaced support remains causally useful, and finally trace how residual contrast relates to the model's native evidence-routing path.

3.1 A CONTROLLED DIAGNOSTIC OF NONLOCAL EVIDENCE

We construct a content-matched diagnostic using a frozen Qwen3-1.7B base model. Each of 256 independent HotpotQA sources is expanded into a fixed-length long prompt containing one annotated gold-support bundle and six equal-token distractor-control bundles. The seven bundles occupy predeclared slots ranging from approximately 0.23K to 8.93K tokens from the answer-query position. Across the seven variants of a source, the support exchanges slots with a control; the question, answer, token multiset, document multiset, total length, and all unaffected token spans remain fixed. We measure raw residual states at all 28 layers, yielding 1,792 matched prompt variants.

Let q be the position whose next-token distribution begins the answer, L_q the 64-token window ending at q, S the gold-support tokens, and C the equal-token distractor control occupying the matched slot. For a region U, write $\bar{h}_U^{(\ell)} = |U|^{-1} \sum_{i \in U} h_i^{(\ell)}$. We measure local common-mode concentration as

$$c_L^{(\ell)} = \frac{\|\bar{h}_{L_q}^{(\ell)}\|_2^2}{|L_q|^{-1} \sum_{i \in L_q} \|h_i^{(\ell)}\|_2^2}, \tag{1}$$

and support–control contrast as

$$d_{S,C}^{(\ell)} = \frac{\|\bar{h}_S^{(\ell)} - \bar{h}_C^{(\ell)}\|_2^2}{\frac{1}{2}\Big(|S|^{-1} \sum_{i \in S} \|h_i^{(\ell)}\|_2^2 + |C|^{-1} \sum_{i \in C} \|h_i^{(\ell)}\|_2^2\Big)}. \tag{2}$$

The ratio $R_{S,C}^{(\ell)} = d_{S,C}^{(\ell)} / (c_L^{(\ell)} + \epsilon)$ measures evidence-specific variation relative to the local common mode.

3.2 NONLOCAL EVIDENCE ENTERS A COMMON-MODE-DOMINATED REGIME

Figure 1 shows the effect of matched evidence relocation on residual-stream contrast. Averaged over 256 sources and 28 layers, moving support from the nearest slot (0.23K tokens) to the first nonlocal slot (1.33K tokens) increases c_L from 0.447 to 0.552. The support–control contrast $d_{S,C}$ remains nonzero, changing from 0.088 to 0.078, while $R_{S,C}$ falls from 0.201 to 0.141. The lower $R_{S,C}$ persists across the remaining nonlocal slots and is visible across layers (Figure 1(c)). Thus, support–control differences persist but weaken relative to the query-local common mode.

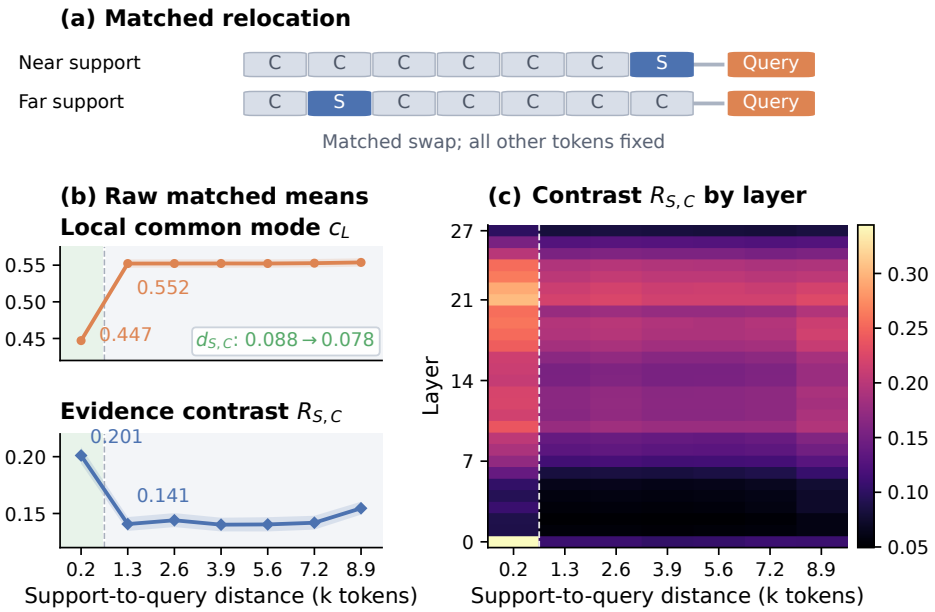

Figure 1: **Matched evidence relocation.** (a) Gold support swaps with an equal-token control; other tokens remain fixed. (b) Mean c_L and $R_{S,C}$ by support distance ($d_{S,C}$ in the callout). (c) $R_{S,C}$ by residual layer. Frozen Qwen3-1.7B, 256 sources; bands show 95% source-bootstrap intervals (10,000 resamples).

3.3 From residual-stream contrast to native evidence routing

We next examine the same relocation after the residual stream enters the model's native attention computation. Across the 256 sources used in Figure 1, moving support from the nearest to the first nonlocal placement changes the support-minus-control post-RoPE query-key score from 2.169 to -0.890, and normalized attention-mass contrast from 0.0305 to -0.0068. Thus, the support also loses its advantage over the matched control during native routing.

We then measure answer discrimination on the matched prompts. Let $y^\star$ and y^{ctr} denote the complete gold and equal-token counter answers. Under teacher forcing, define their per-token log-probability margin as

$$M_\theta(x) = \frac{\log p_\theta(y^\star \mid x)}{|y^\star|} - \frac{\log p_\theta(y^{\mathrm{ctr}} \mid x)}{|y^{\mathrm{ctr}}|}. \tag{3}$$

The margin falls from 2.165 to 2.078 at the first nonlocal placement, a paired change of -0.087 (95% CI $[-0.125, -0.050]$). This decrease comes primarily from lower gold-answer likelihood; counter-answer likelihood remains nearly unchanged.

The support is still useful at this placement: deleting its cached K/V reduces the same full-answer margin more than deleting the matched control (mean difference 1.049, 95% CI $[0.917, 1.179]$). Taken together, the measurements show useful nonlocal evidence with weaker advantages in native routing and answer scoring. This is the conditioning regime that motivates ResiBand's residual-stream intervention.

Design implication The diagnosis identifies a representation problem rather than a missing-memory problem: useful evidence remains available, but its distinction from matched distractors is weak relative to a shared residual-stream background. We therefore test an intervention that suppresses locally shared variation at each token and learns which remaining directions to write into the residual stream. Native attention remains responsible for retrieving distant evidence. Section 4 defines this intervention, and Section 5.5 tests whether it improves subsequent evidence competition.

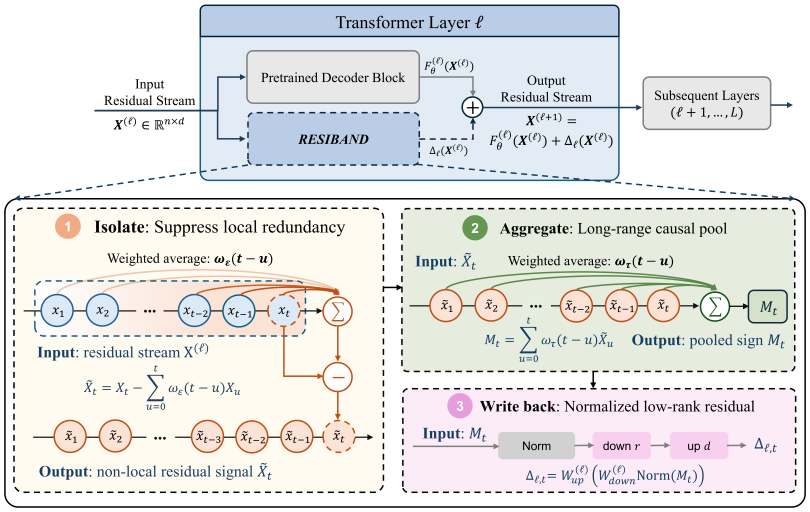

Figure 2: Architecture of RESIBAND. The module follows a three-stage path: local redundancy suppression via a causal local average, non-local residual aggregation via a long-range causal pool, and low-rank write-back into the residual stream.

4 METHOD

RESIBAND converts temporal variation relative to the local hidden-state background into a learned residual correction. At decoder layer ℓ, let $X^{(\ell)} \in \mathbb{R}^{n\times d}$ denote the input residual stream and $F_\theta^{(\ell)}$ the frozen Transformer block. We modify its output as

$$X^{(\ell+1)} = F_\theta^{(\ell)}\Big(X^{(\ell)}\Big) + \Delta_\ell\Big(X^{(\ell)}\Big). \tag{4}$$

4.1 CAUSAL RESIDUAL BAND

For a time constant $\rho > 0$, define the normalized causal exponential smoother

$$L_\rho(X)_t = \sum_{u=0}^{t} w_{\rho,t}(t-u)X_u, \qquad w_{\rho,t}(j) = \frac{\exp(-j/\rho)}{\sum_{v=0}^{t}\exp(-v/\rho)}. \tag{5}$$

The signal supplied to the learned branch is

$$Z^{(\ell)} = L_\tau\Big((I - L_\varepsilon)X^{(\ell)}\Big). \tag{6}$$

The short-horizon term $L_\varepsilon(X)$ estimates the component shared by nearby tokens. Subtracting it exposes variation relative to that local background; the longer-horizon smoother L_τ then integrates variation that persists across a wider causal context. Together they form a fixed causal band-pass along the token axis.

4.2 LEARNED RESIDUAL WRITE

The fixed band-pass does not determine which variation is useful. Each installed layer therefore learns a rank-r projection of the filtered signal. For token t, let $z_t^{(\ell)} = (Z_{t,:}^{(\ell)})^\top \in \mathbb{R}^d$ be its filtered state written as a column. Its column-vector update is

$$\Delta_{\ell,t}(X^{(\ell)}) = W_{\text{up}}^{(\ell)} W_{\text{down}}^{(\ell)} \operatorname{Norm}\Big(z_t^{(\ell)}\Big), \tag{7}$$

where $W_{\text{down}}^{(\ell)} \in \mathbb{R}^{r\times d}$ and $W_{\text{up}}^{(\ell)} \in \mathbb{R}^{d\times r}$; row t of $\Delta_\ell(X^{(\ell)}) \in \mathbb{R}^{n\times d}$ is $\Delta_{\ell,t}(X^{(\ell)})^\top$. Only these projections are optimized. W_{up} is initialized to zero, so training begins from the original model. The task loss learns which directions in the residual band should be written back; subsequent frozen blocks then perform their native evidence routing on the corrected residual stream.

5 EXPERIMENTS

We evaluate RESIBAND as a trainable residual operator. We ask four questions. **RQ1:** Does RESIBAND improve answer-level performance across backbones and controlled evidence-distance settings? **RQ2:** Is the gain attributable to the band-limited residual operator rather than either filter branch alone, causal attention, or generic task adaptation? **RQ3:** At which depths does the trained operator contribute at inference? **RQ4:** How does RESIBAND alter the conditioning state diagnosed in Section 3, and do its learned updates causally improve native evidence routing and the answer?

5.1 EXPERIMENTAL SETUP

Across Qwen3-1.7B-Base, Qwen3-8B-Base (Yang et al., 2025), and Llama-3.1-8B (Grattafiori et al., 2024), we compare frozen Base, LoRA, and all-layer RESIBAND. The adapted arms use the same HotpotQA–GSM8K training mixture and fixed evaluation rows. We also compare single-branch and attention-pooling controls, mask layer regions at inference, and use matched interventions to probe the learned update. We report whole-completion F1 on MuSiQue (MQ), HotpotQA, and 2WikiMultiHopQA (2Wiki), numeric exact match on GSM8K (GSM) (Cobbe et al., 2021), and accuracy on MMLU (Hendrycks et al., 2021). The main comparison uses three training seeds; operator and depth ablations use seed 42. Full training and scoring details are in Appendix B, and mechanism protocols are in Appendix C.1.

5.2 RQ1: ANSWER-LEVEL PERFORMANCE

Table 1 reports the cross-backbone comparison with matched training exposure for the adapted arms.

Table 1: Cross-backbone results. Adapted rows show mean ± sample SD over three seeds; Base is evaluated once. MH Avg. averages MQ, HotpotQA, and 2Wiki; Δ is the RESIBAND−LoRA difference in points.

Backbone	Method	Multi-hop QA				Math / General	
		MQ	HotpotQA	2Wiki	MH Avg.	GSM	MMLU
Qwen3-1.7B	Base	2.00	11.36	7.01	6.79	7.48	33.05
	LoRA	6.67 ± 0.58	9.73 ± 0.62	7.83 ± 0.35	8.08 ± 0.37	49.54 ± 3.87	36.32 ± 0.65
	RESIBAND	16.10 ± 4.98	36.29 ± 6.40	29.75 ± 7.43	27.38 ± 6.10	51.21 ± 2.35	37.65 ± 0.30
	Δ vs. LoRA	+9.43	+26.56	+21.92	+19.30	+1.67	+1.33
Qwen3-8B	Base	10.99	17.04	15.96	14.67	74.42	36.62
	LoRA	10.20 ± 1.15	14.76 ± 3.45	13.28 ± 1.90	12.74 ± 2.09	72.29 ± 2.86	44.31 ± 0.86
	RESIBAND	15.30 ± 4.61	22.38 ± 7.02	33.39 ± 2.13	23.69 ± 4.37	72.38 ± 4.24	45.34 ± 0.55
	Δ vs. LoRA	+5.10	+7.62	+20.11	+10.95	+0.09	+1.03
Llama-3.1-8B	Base	2.77	7.28	8.06	6.04	18.13	34.52
	LoRA	6.14 ± 1.21	8.10 ± 2.37	10.24 ± 1.16	8.16 ± 1.53	34.82 ± 1.75	39.97 ± 0.89
	RESIBAND	8.76 ± 0.82	10.65 ± 0.93	10.87 ± 0.99	10.09 ± 0.79	39.40 ± 4.84	40.52 ± 0.24
	Δ vs. LoRA	+2.62	+2.55	+0.63	+1.93	+4.58	+0.55

Across all nine multi-hop model–benchmark pairs, RESIBAND exceeds both LoRA and frozen Base. Its multi-hop averages improve over LoRA by 19.30, 10.95, and 1.93 points on Qwen3-1.7B, Qwen3-8B, and Llama-3.1-8B, respectively. On Qwen3-8B, LoRA falls below Base (12.74 versus 14.67), whereas RESIBAND reaches 23.69. GSM8K and MMLU also match or exceed LoRA on every backbone. Thus the shared configuration substantially improves evidence-heavy QA without a comparable cost on the other evaluated tasks.

Controlled evidence-distance validation. We return to the matched HotpotQA relocation protocol of Section 3 to test generated answers as support moves away from the query. An outcome-blind subset of 100 sources supplies five fixed support positions. We show the setting with six matched distractor bundles, where relevant evidence competes with similar content; the remaining neutral filler is natural text. Frozen Base, the Qwen3-1.7B seed-42 Q/V LoRA checkpoint included in Table 1, an adapter-only control, and RESIBAND receive identical prompts.

Table 2: Answer F1 by support distance on Qwen3-1.7B.

Distance (K)	Base	LoRA	Adapter only	RESIBAND
0.23	0.040	0.112	0.217	**0.485**
1.33	0.042	0.123	0.202	**0.417**
2.58	0.033	0.149	0.183	**0.436**
3.94	0.032	0.122	0.186	**0.401**
8.92	0.035	0.160	0.168	**0.494**
Mean	0.036	0.133	0.191	**0.447**

Averaged over the five distances with matched distractors, RESIBAND reaches answer F1 0.447, compared with 0.191 for the adapter-only control and 0.133 for LoRA. Its paired source-level gains are $+0.256$ over the adapter-only control (95% bootstrap CI $[+0.191, +0.321]$) and $+0.314$ over LoRA ($[+0.249, +0.379]$). RESIBAND has the highest mean F1 at every tested distance, including 0.494 at 8.92K tokens versus 0.168 for adapter only (Table 2).

5.3 RQ2: WHICH RESIDUAL OPERATOR MATTERS?

Table 3 compares residual operators under the same all-layer reference protocol on Qwen3-1.7B. The adapter-only control uses the same rank-8 write, parameter count, training examples, and optimizer budget as RESIBAND, but receives the unfiltered residual stream. Its multi-hop average is 12.00, compared with 22.41 for RESIBAND. Causal attention and LP alone reach 7.36 and 7.57, respectively, while HP alone reaches 20.93. Combining the two fixed temporal scales gives RESIBAND the best multi-hop average and MMLU score. On GSM8K, the same-data adapted arms are close: adapter only scores 55.31, LoRA 53.99, and RESIBAND 53.28.

Table 3: Qwen3-1.7B operator ablation. Bold and underlined values mark the best and second-best results per column.

Method	Multi-hop QA				Math / General	
	MQ	HotpotQA	2Wiki	MH Avg.	GSM	MMLU
Frozen Base	2.00	11.36	7.01	6.79	7.48	33.05
LoRA	6.21	9.29	7.43	7.65	<u>53.99</u>	36.58
Causal attention	6.11	8.85	7.13	7.36	50.24	<u>37.67</u>
Low-pass only	6.06	9.05	7.60	7.57	50.02	35.99
High-pass only	<u>11.11</u>	<u>29.39</u>	<u>22.30</u>	<u>20.93</u>	51.36	37.38
Adapter only	9.73	12.48	13.78	12.00	**55.31**	37.11
RESIBAND	**13.90**	**30.65**	**22.69**	**22.41**	53.28	**37.77**

5.4 RQ3: DEPTH-WISE CONTRIBUTION

Starting from the all-layer-trained checkpoint, we freeze every parameter and change the active layers at inference. The fixed early/middle/late thirds are 0–8/9–18/19–27 for Qwen3-1.7B, 0–11/12–23/24–35 for Qwen3-8B, and 0–10/11–21/22–31 for Llama. These rows diagnose how branches learned jointly in the all-layer model contribute when they are selectively enabled at inference.

Among the three region masks, early and middle give the top two multi-hop averages on every backbone, while late is weakest (Table 4). Early-only raises the multi-hop average from 22.41 to 45.31 on Qwen3-1.7B and from 19.86 to 63.76 on Qwen3-8B; Llama-3.1-8B instead peaks with middle-only (13.74 versus 10.55).

Earlier writes leave more frozen blocks to integrate evidence, plausibly explaining the weak late-only results. On Qwen, early-only also beats all-layer execution, suggesting that later RESIBAND writes can reshape an effective early correction. Llama's peak lies one region later. All-layer remains best on Qwen MMLU and Llama GSM8K. Overall, multi-hop gains favor early-to-middle writes, with a backbone-dependent peak.

Table 4: Inference-region ablation with fixed all-layer-trained weights.

Active region	Multi-hop QA				Math / General	
	MQ	HotpotQA	2Wiki	MH Avg.	GSM	MMLU
Qwen3-1.7B	*(Early 0–8; Middle 9–18; Late 19–27)*					
Early	**34.34**	**60.13**	**41.45**	**45.31**	**57.51**	36.31
Middle	20.00	51.64	23.87	31.83	49.00	36.95
Late	1.81	6.12	6.94	4.96	21.54	32.97
All layers	13.90	30.65	22.69	22.41	53.28	**37.77**
Qwen3-8B	*(Early 0–11; Middle 12–23; Late 24–35)*					
Early	**49.82**	**76.91**	**64.55**	**63.76**	73.87	41.36
Middle	15.00	25.10	21.32	20.47	**74.19**	43.31
Late	11.19	17.14	12.85	13.73	63.73	36.68
All layers	13.13	15.18	31.28	19.86	73.50	**45.86**
Llama-3.1-8B	*(Early 0–10; Middle 11–21; Late 22–31)*					
Early	7.96	11.04	10.75	9.92	33.79	**41.45**
Middle	8.43	**20.99**	**11.82**	**13.74**	10.29	36.84
Late	3.99	8.12	8.56	6.89	12.77	34.37
All layers	**8.48**	11.49	11.67	10.55	**39.81**	40.66

5.5 RQ4: How ResiBand reconditions evidence routing

Section 3 establishes the frozen-model conditioning problem. Here we test how the ResiBand filter conditions the signal, whether the learned write changes native evidence routing, and whether the answer favors relevant evidence. All diagnostics use Qwen3-1.7B and prompts disjoint from the behavioral test rows. The three panels of Figure 3 use the same 100 matched 8K HotpotQA evidence-relocation sources. In the downstream diagnostics, On and Off enable or disable the trained ResiBand modules in the same checkpoint. We test write-direction specificity separately on controlled passcode prompts in Appendix C.9.

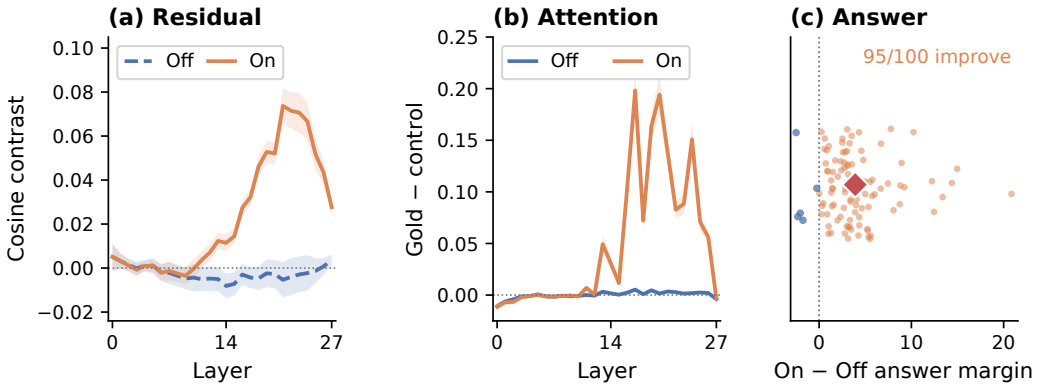

Figure 3: **Mechanism measurements for RQ4.** (a) Query-relative residual-stream cosine contrast $\cos(h_q, \bar{h}_S) - \cos(h_q, \bar{h}_C)$; (b) support–control attention; (c) correct–counter full-answer margin. Lines show source means; bands are 95% source-bootstrap intervals. Dots denote sources and diamonds denote means.

Residual-stream contrast. With all trained modules On, the 28-layer mean query-relative residual-stream contrast $H_{S,C}$ increases by $+0.0261$ over the same checkpoint with the modules Off (95% paired-bootstrap CI $[+0.0227, +0.0294]$; 96/100 sources improve). The two runs use identical questions and measure the same raw residual-stream location at each layer (Figure 3a). The increase appears mainly in later layers; this all-layer trajectory does not assign the gain to writes at those same layers (cf. RQ3).

Native evidence routing. On increases the far gold-minus-matched-control attention-mass contrast through the later frozen layers (Figure 3b). Averaged across all 28 layers, the paired gain is $+0.049$ with 95% bootstrap CI $[+0.045, +0.054]$ on 100 matched 8K pairs (100/100 positive). The same 100 pairs also provide residual-stream contrast, post-RoPE QK, attention, and weighted-value trajectories in Figure 8. A complementary local-gradient diagnostic in Appendix C probes task alignment at the all-writes-Off operating point.

Answer discrimination. Figure 3c shows paired On-minus-Off gains for the same 100 far 8K prompts used in panels (a) and (b). The mean full-answer margin M_θ rises from 4.133 to 8.050, a gain of $+3.917$ (95% paired-bootstrap CI $[+3.193, +4.670]$; 95/100 sources improve). The three panels therefore show residual-stream, native-attention, and answer-level changes on the same questions.

Across the 100 shared sources, gains in residual-stream cosine contrast and attention contrast are positively associated (Spearman $\rho = 0.467$, 95% source-bootstrap CI $[0.286, 0.613]$). Together, the matched panels show stronger evidence contrast, more attention to relevant evidence, and a larger correct-answer margin. Further content-selectivity and learned-write interventions are reported in Appendices C.8 and C.9.

5.6 Efficiency and limitations

Each rank-r module adds $2dr$ parameters: $32{,}768$ on Qwen3-1.7B, or 0.065% of one decoder block. RQ3 shows that the highest multi-hop averages come from early layers on the Qwen backbones and middle layers on Llama. We therefore profile deployment with only the corresponding region active. The native block and ResiBand branch read the same layer input, allowing us to execute them on separate CUDA streams and synchronize before residual addition. On an A800 with batch size one, generating 32 output tokens from a 4096-token input adds 2.7%, 3.1%, and 2.9% inference latency for Qwen3-1.7B (early nine layers), Qwen3-8B (early twelve layers), and Llama-3.1-8B (middle eleven layers), respectively, relative to their matched frozen Base models.

6 Conclusion

We study distractor-rich evidence integration in decoder-only Transformers, where a small set of task-relevant facts must compete with plausible but irrelevant information. Our results identify a specific failure mode: supporting evidence can remain causally available while its task-relevant variation is weak relative to a locally shared component of the residual stream. Matched analyses link this poorly conditioned signal to weaker query-relative evidence separation through projected scores, attention selection, and value delivery, while K/V deletion shows that the same evidence remains causally load-bearing for the answer. This distinction between evidence availability and competitive use motivates ResiBand, which suppresses local common-mode structure, stabilizes the remaining temporal variation, and maps it into a learned low-rank residual update while leaving the pretrained backbone frozen.

Across three decoder backbones, ResiBand improves multi-hop question answering over both the frozen model and a same-contract LoRA baseline. Operator ablations and causal interventions connect these gains to a coherent mechanism: fixed filtering improves evidence contrast, the learned write selects task-aligned directions, and subsequent native computation gives supporting evidence a stronger advantage over matched distractors. The gain does not require a uniform redistribution of attention toward query-separated tokens. Together, these findings establish residual-stream conditioning as a complementary design axis for evidence integration: when the required information is present but poorly expressed for downstream competition, reconditioning its signal can improve how the model uses it.

AI USE STATEMENT

In this work, we used OpenAI Codex primarily to polish the language of the manuscript. Additionally, we used it to refine the presentation of figures and to draft and revise the disclosure statements. We have reviewed all AI-assisted material included in this paper. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix A specify the operator and its derivation. Section 5.1 and Appendix B describe the model backbones, training arms, data and evaluation protocol, hyperparameters, seeds, and compute resources. Appendix C.1 documents the matched diagnostic construction and statistical protocol used for the mechanism analyses.

REFERENCES

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G. M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In *NeurIPS*, 2024.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? In *COLM*, 2025.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. LongNet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2021.

Gbètondji J-S Dovonon, Michael M. Bronstein, and Matt J. Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022.

Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. In *ICLR*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *COLING*, 2020.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *TACL*, 2024.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In *ICML*, 2025.

Tam Nguyen, Tan M. Nguyen, and Richard G. Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. In *NeurIPS*, 2023.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *NeurIPS*, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP-IJCNLP*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

A METHOD DETAILS

This appendix expands the RESIBAND construction in Section 4. We give the kernel derivation behind the scale–shape parameterization and record implementation details used in the main experiments.

A.1 BAND-PASS KERNEL DERIVATION

RESIBAND applies a causal local subtraction followed by a causal long-range pooling operation along the token axis. Writing the operator composition explicitly, we define

$$K_{\tau,\varepsilon} = L_\tau \circ (\delta - L_\varepsilon), \tag{8}$$

where L_ρ denotes causal exponential smoothing with time constant ρ, and δ is the identity operator. The subtraction term $(\delta - L_\varepsilon)$ removes locally redundant components, while L_τ aggregates the remaining signal over a longer horizon.

First-order continuous-time proxy. To obtain an interpretable spectral characterization, we analyze a first-order continuous-time proxy of the discrete token-axis operator. Under this approximation, a causal exponential smoother with time constant ρ has frequency response

$$H_\rho(\omega) = \frac{1}{1 + i\omega\rho}. \tag{9}$$

Substituting this form into Eq. (8) gives

$$H_{\tau,\varepsilon}(\omega) = H_\tau(\omega)\big(1 - H_\varepsilon(\omega)\big) = \frac{i\omega\varepsilon}{(1 + i\omega\tau)(1 + i\omega\varepsilon)}. \tag{10}$$

The factor $1 - H_\varepsilon(\omega)$ suppresses low-frequency local components, whereas $H_\tau(\omega)$ attenuates high frequencies. Their product therefore yields a band-pass profile.

The squared magnitude is

$$|H_{\tau,\varepsilon}(\omega)|^2 = \frac{\omega^2\varepsilon^2}{(1 + \omega^2\tau^2)(1 + \omega^2\varepsilon^2)}. \tag{11}$$

Since the prefactor ε^2 does not affect the peak location, we introduce $z = \omega^2$ and consider

$$g(z) = \frac{z}{(1 + z\tau^2)(1 + z\varepsilon^2)}. \tag{12}$$

Differentiating with respect to z gives

$$g'(z) = \frac{1 - z^2\tau^2\varepsilon^2}{(1 + z\tau^2)^2\,(1 + z\varepsilon^2)^2}. \tag{13}$$

The unique positive stationary point satisfies

$$z_\star = \omega_\star^2 = \frac{1}{\tau\varepsilon}, \tag{14}$$

hence

$$\omega_\star = \frac{1}{\sqrt{\tau\varepsilon}}. \tag{15}$$

Moreover, $g(0) = 0$, $g(z) \to 0$ as $z \to \infty$, and $g'(z)$ changes sign from positive to negative at $z_\star$, so this stationary point is the unique positive maximizer. This identifies the characteristic token scale of the bridge as

$$s_{\tau,\varepsilon} \propto \frac{1}{\omega_\star} = \sqrt{\tau\varepsilon}. \tag{16}$$

Scale–shape reparameterization. Motivated by the peak analysis above, we reparameterize the two time constants as

$$s_{\tau,\varepsilon} = \sqrt{\tau\varepsilon}, \qquad r_{\tau,\varepsilon} = \frac{\tau}{\varepsilon}, \tag{17}$$

where $s_{\tau,\varepsilon}$ sets the characteristic scale and $r_{\tau,\varepsilon}$ controls the shape of the band-pass profile. The original parameters are recovered as

$$\tau = s_{\tau,\varepsilon}\sqrt{r_{\tau,\varepsilon}}, \qquad \varepsilon = s_{\tau,\varepsilon}/\sqrt{r_{\tau,\varepsilon}}. \tag{18}$$

This interpretation becomes explicit after introducing the normalized frequency $x = \omega s_{\tau,\varepsilon}$. Substituting the above relations into Eq. (11) yields

$$|H_{s,r}(\omega)|^2 = \frac{x^2/r}{(1+x^2r)(1+x^2/r)}. \tag{19}$$

Eq. (19) shows that, after normalization by the scale s, the response profile depends only on r. In particular, the peak occurs at $x_\star = 1$, confirming that s determines the center scale, while r determines the shape of the normalized response, including its bandwidth and peak sharpness. Accordingly, varying s shifts the bridge across token granularities, whereas varying r changes its selectivity at a fixed center scale.

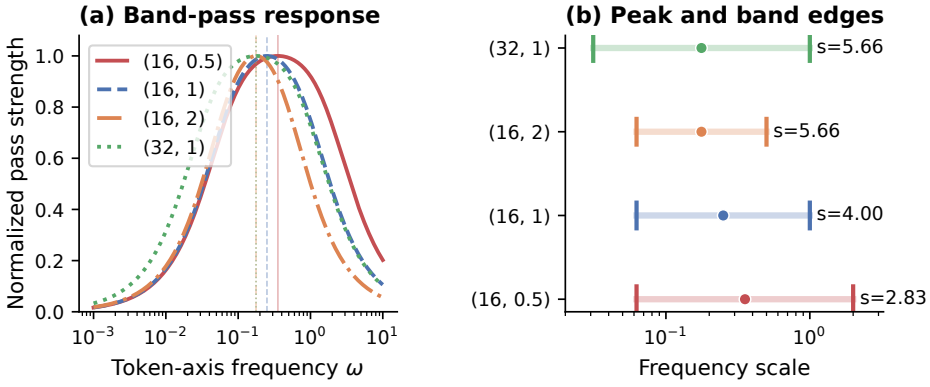

Figure 4: Frequency view of the RESIBAND band-pass kernel. The preferred scale is controlled approximately by $s_{\tau,\varepsilon} = \sqrt{\tau\varepsilon}$, while the shape ratio $r_{\tau,\varepsilon} = \tau/\varepsilon$ changes the band shape and edge locations.

A.2 CAUSAL RECURRENCE AND IMPLEMENTATION

Both token-axis smoothers admit an exact causal recurrence. Evaluating this recurrence avoids materializing the $n \times n$ exponential kernel while preserving the normalized causal weighting in Eq. (5). For a sequence $X_0, \ldots, X_{n-1}$, the smoother with time constant ρ obeys the following recurrence for $t = 0, \ldots, n-1$:

$$P_t^{(\rho)} = \lambda_\rho P_{t-1}^{(\rho)} + X_t, \qquad Z_t^{(\rho)} = \lambda_\rho Z_{t-1}^{(\rho)} + 1, \qquad L_\rho(X)_t = P_t^{(\rho)}/Z_t^{(\rho)}, \tag{20}$$

where $\lambda_\rho = \exp(-1/\rho)$, with $P_{-1}^{(\rho)} = 0$ and $Z_{-1}^{(\rho)} = 0$. The accumulator $P_t^{(\rho)}$ stores the exponentially weighted sum of previous token states, while $Z_t^{(\rho)}$ stores the corresponding normalization mass. Therefore $L_\rho(X)_t$ is exactly the normalized causal exponential average

$$L_\rho(X)_t = \frac{\sum_{u=0}^{t} \lambda_\rho^{t-u} X_u}{\sum_{u=0}^{t} \lambda_\rho^{t-u}}. \tag{21}$$

The computation is causal because the state at position t depends only on positions $u \le t$, and it requires only running accumulators rather than an explicit token-token kernel.

Although the exponential kernel assigns its largest weight to recent positions, a larger τ allows variations to persist in the running summary over a longer horizon. RESIBAND applies this smoother

to the high-pass residual $\widetilde{X}$, after locally shared components have been suppressed. The resulting signal emphasizes directions not already explained by the immediate neighborhood while damping isolated high-frequency fluctuations.

At each installed decoder layer, the implementation first computes the causal residual band in Eq. (6) and then writes its normalized output back through the low-rank projection in Eq. (7). The operator is therefore upstream of subsequent native attention computation, since its correction is added to the residual stream before later projections are evaluated.

The write-back projection $W_{\mathrm{up}}^{(\ell)}$ is initialized to zero, so $\Delta_\ell(X) = 0$ at initialization for every layer and input. As a result, installing RESIBAND does not change the pretrained model's initial function; all operator effects arise only through training.

A.3 PARAMETER ACCOUNTING

The causal smoothing operators L_τ and L_ε have no learned parameters in the reported experiments, and their time constants are fixed hyperparameters. The only trainable parameters introduced by RESIBAND are the low-rank write-back projections. For hidden size d, write rank r, and L_b installed decoder layers, the operator adds $2L_b dr$ trainable parameters, since each installed layer contains $W_{\mathrm{down}}^{(\ell)} \in \mathbb{R}^{r\times d}$ and $W_{\mathrm{up}}^{(\ell)} \in \mathbb{R}^{d\times r}$.

In the reported implementation, Norm has no learned affine parameters. Thus, the count above is the complete RESIBAND trainable budget. With $r = 8$, this corresponds to approximately 0.92M parameters for Qwen3-1.7B-Base, 2.10M for Llama-3.1-8B, and 2.36M for Qwen3-8B-Base. The operator therefore adds a small trainable budget and avoids attention-style pairwise token interactions.

B TRAINING AND EVALUATION DETAILS

B.1 TRAINING ARMS

The main comparison contains three arms. **Base** is the frozen pretrained decoder without task adaptation. **LoRA** trains rank-8 adapters on the native query and value projections with $\alpha = 16$. **ResiBand** freezes the same backbone and trains the rank-8 low-rank maps in Eq. (7) at every decoder layer. The fixed temporal filters introduce no trainable parameters. These controls use the same training examples, optimizer-step budget, effective batch size, and evaluation rows, so the comparison isolates the adaptation pathway under matched training exposure.

For L layers, hidden width d, and rank r, RESIBAND has $2Ldr$ trainable parameters. Query/value LoRA has $Lr[(d + d_Q) + (d + d_V)]$ parameters, where d_Q and d_V are the native projection output widths. Table 5 gives the exact counts used in the experiments. The LoRA baselines are therefore modestly *larger*, by 1.63–$1.75\times$; observed RESIBAND gains cannot be attributed to a larger trainable parameter budget.

Table 5: Exact trainable parameter counts for the main rank-8 arms. Backbone parameters are frozen and the temporal filters contain no learned weights.

Backbone	RESIBAND	Q/V LoRA	LoRA / RESIBAND
Qwen3-1.7B	917,504	1,605,632	$1.75\times$
Llama-3.1-8B	2,097,152	3,407,872	$1.63\times$
Qwen3-8B	2,359,296	3,833,856	$1.63\times$

LoRA is optimized with learning rate 10^{-4}; the RESIBAND maps use 3×10^{-4}. The rates are fixed by method rather than numerically matched because the two parameterizations have different optimization scales; neither rate is selected on formal test rows.

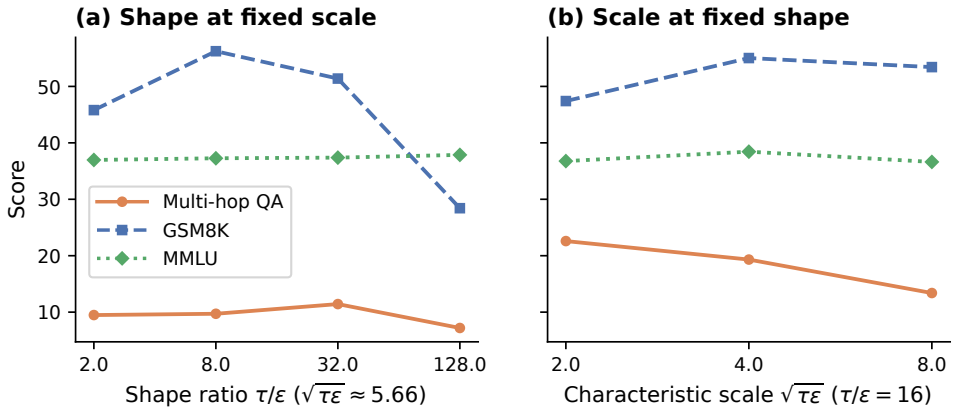

Figure 5: Pure-RESIBAND sensitivity on Qwen3-1.7B. Left: τ/ε varies while $\sqrt{\tau\varepsilon} \approx 5.66$. Right: $\sqrt{\tau\varepsilon}$ varies while $\tau/\varepsilon = 16$. Multi-hop QA is the mean of MuSiQue hop-macro F1, HotpotQA F1, and 2WikiMultiHopQA F1. The sweep exposes a task-dependent trade-off rather than a monotone preference for a wider or longer band.

B.2 TRAINING PROTOCOL

All trained arms use the same locked HotpotQA–GSM8K mixture after removal of held-out and policy-development identities: 7,998 HotpotQA examples and 7,361 GSM8K examples. Each run uses 960 optimizer steps, effective batch size 16, assistant-token loss, maximum sequence length 4096, and one of training seeds 42, 1, or 2. The shared RESIBAND operating point is $(\tau, \varepsilon) = (16, 0.5)$ on all three backbones. Hyperparameters are not selected on the formal test rows.

B.3 EVALUATION PROTOCOL

The evaluation uses fixed row identities and gold answers across arms: 200 examples for each MuSiQue hop count (2/3/4), 500 HotpotQA examples, 500 2Wiki examples, the fixed GSM8K difficulty buckets, and all 14,042 MMLU questions. We report normalized token F1 for MuSiQue, HotpotQA, and 2Wiki, macro-averaging the three MuSiQue hop means; GSM8K uses bucket-macro exact match, and MMLU uses accuracy.

Generative outputs are deterministically truncated at clear completion boundaries before scoring. GSM8K uses final-number extraction. MMLU is not generation-based: we score the complete text of each answer option with length-normalized conditional log-likelihood and select the highest-scoring option.

B.4 OPERATING-POINT SENSITIVITY

The main configuration is fixed at $(\tau, \varepsilon) = (16, 0.5)$. Figure 5 reports a separate Qwen3-1.7B sensitivity study that varies the characteristic scale $s = \sqrt{\tau\varepsilon}$ and shape ratio τ/ε. Every cell uses the same pure RESIBAND architecture and is independently trained under the main protocol; the sweep is used only for sensitivity diagnosis, not to select the headline configuration. At fixed $s \approx 5.66$, increasing the shape ratio from 2 to 32 improves the multi-hop aggregate from 9.48 to 11.43, but the extreme ratio 128 falls to 7.19 and sharply reduces GSM8K. At fixed ratio 16, increasing s from 2 to 8 reduces the multi-hop aggregate from 22.59 to 13.39, while GSM8K peaks at the intermediate scale $s = 4$ (55.03). MMLU remains within a narrower 36.62–38.43 range. Thus the response is neither monotone nor benchmark-independent. We consequently retain one shared operating point across backbones and never select (τ, ε) from formal test performance.

B.5 COMPUTE RESOURCES

Training and evaluation use single NVIDIA A800 80GB GPUs. Training jobs request one GPU, eight CPU cores, 80GB host memory, and a four-hour wall-time allocation. Dataset-level evaluations

are split into independent jobs but use identical checkpoints, prompts, row identities, and scoring code.

C Complete Mechanism Diagnostics

This appendix reports the matched diagnostics behind Sections 3 and 5.5. Unless stated otherwise, all measurements use Qwen3-1.7B: frozen Base for the problem diagnosis and the fixed all-layer RESIBAND checkpoint for method-side interventions. Diagnostic examples, layers, and interventions were fixed before reading their outcomes and are disjoint from the formal behavioral test rows.

C.1 Matched-pair construction and statistical protocol

The confirmatory conditioning suite contains 256 independent sources and seven predeclared support positions per source, for 1,792 prompt variants. The support bundle and six equal-token controls exchange slots while the question, answer, total length, token and document multisets, and every unaffected token span remain fixed. All 28 raw residual layers are measured.

The filter and layerwise routing suites use the same 100 matched 8K-token sources as Figure 3, with one near/far exchange per source. The filter analysis uses eight fixed layers. For the routing analysis, at every layer we measure gold-support minus matched-control contrast in the residual stream, post-RoPE native QK score, native attention mass, and attention-weighted value norm. The gradient and Q/K/V interventions use these same sources. The content and deletion controls each use 24 paired passcode prompts at each of six distances (64–2048 tokens); the residual-write control uses 24 paired prompts at each of 512 and 1024 tokens. These controls use separate synthetic passcode cohorts.

Unless labeled as standard errors, intervals below are two-sided 95% paired bootstrap intervals from $10{,}000$ resamples at the source or matched-example level; the cross-backbone Base associations use $5{,}000$ source resamples. They quantify sample uncertainty for a fixed checkpoint.

C.2 Confirmatory nonlocal-conditioning sweep

Table 6 gives the source-level means behind Figure 1. Moving support from the query-local slot to the first nonlocal slot produces the dominant transition: c_L increases from 0.447 to 0.552, while $d_{S,C}$ decreases from 0.088 to 0.078 and remains nonzero. Consequently $R_{S,C}$ decreases from 0.201 to 0.141 and then largely plateaus. From the nearest to the farthest placement, the paired changes are $+0.107$ for c_L (95% CI $[0.100, 0.113]$), -0.0025 for $d_{S,C}$ ($[-0.0030, -0.0021]$), and -0.046 for $R_{S,C}$ ($[-0.050, -0.043]$).

Table 6: Matched relocation in frozen Qwen3-1.7B. Values are averaged over 256 sources and all 28 raw residual layers. Distance denotes the mean separation between the support and query in the predeclared slot.

Distance (K)	c_L	$d_{S,C}$	$R_{S,C}$
0.23	0.447	0.088	0.201
1.33	0.552	0.078	0.141
2.58	0.552	0.080	0.144
3.94	0.552	0.078	0.140
5.57	0.552	0.078	0.140
7.20	0.553	0.079	0.142
8.93	0.554	0.086	0.155

After averaging layers and centering each source across its seven placements, $R_{S,C}$ remains positively associated with query-relative hidden contrast ($r = 0.340$, 95% source-bootstrap CI $[0.285, 0.393]$).

C.3 DISTANCE-ONLY CONTEXT

For distance $d = q - i$, we measure query–token residual similarity and native attention as

$$C_h^{(\ell)}(d) = \mathbb{E}[\cos(h_q^{(\ell)}, h_i^{(\ell)})], \qquad A^{(\ell)}(d) = \mathbb{E}\left[\frac{1}{H}\sum_{r=1}^{H} a_{q,i,r}^{(\ell)}\right]. \tag{22}$$

Figure 6 shows this complementary distance profile in frozen Llama-3.1-8B. Query states are most similar to nearby tokens, and native attention is low at the distances occupied by annotated supports. The matched-support analysis above supplies the content-specific test used in the main diagnosis.

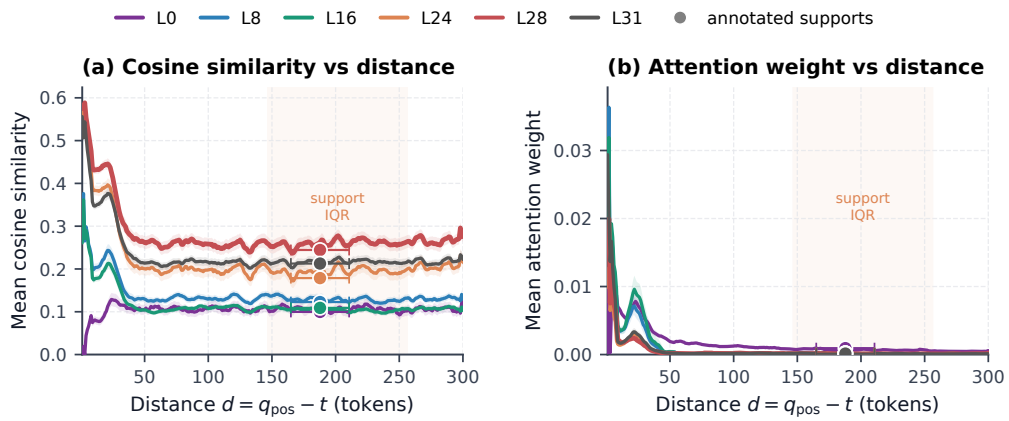

Figure 6: Distance-only residual similarity and native attention on 128 HotpotQA validation examples in frozen Llama-3.1-8B. Dots mark annotated supports; shading marks their interquartile distance range.

C.4 CROSS-BACKBONE BASE ROUTING ASSOCIATIONS

Table 7 traces the association between successive evidence-routing measurements in three frozen Base models. Each model uses 100 matched 8K sources. We average measurements within predeclared depth thirds, then calculate Spearman correlations across sources. The hidden-to-QK and attention-to-value associations are positive across all three models and depth thirds.

Table 7: Source-level Spearman ρ with 95% source-bootstrap intervals (5,000 resamples). H, S, A, and V denote query-relative hidden, post-RoPE QK-score, attention-mass, and attention-weighted-value evidence contrasts, respectively.

Model	Depth	$H \to S$	$S \to A$	$A \to V$
Qwen3-1.7B	Early	.406 [.217, .568]	.335 [.143, .506]	.659 [.516, .770]
	Middle	.604 [.481, .698]	.648 [.514, .752]	.956 [.919, .974]
	Late	.590 [.445, .704]	.627 [.474, .742]	.878 [.813, .918]
Qwen3-8B	Early	.628 [.476, .748]	.048 [−.149, .236]	.630 [.484, .745]
	Middle	.602 [.471, .699]	.288 [.082, .485]	.862 [.768, .925]
	Late	.560 [.418, .672]	.125 [−.090, .340]	.633 [.452, .784]
Llama-3.1-8B	Early	.484 [.303, .636]	.322 [.128, .489]	.343 [.143, .513]
	Middle	.667 [.522, .775]	.478 [.293, .631]	.977 [.957, .985]
	Late	.329 [.133, .496]	.392 [.195, .564]	.958 [.929, .972]

C.5 THE FIXED BAND IMPROVES THE CONDITIONING OF THE WRITABLE SIGNAL

On the same 100 matched sources as Figure 3, the support–control contrast to local common-mode ratio rises from 0.135 in the raw residual stream to 0.193 after band-pass filtering (paired gain

$+0.058$, 95% CI $[0.029, 0.090]$). Thus the fixed band makes evidence contrast more prominent relative to the locally shared component, chiefly in later layers (Figure 7).

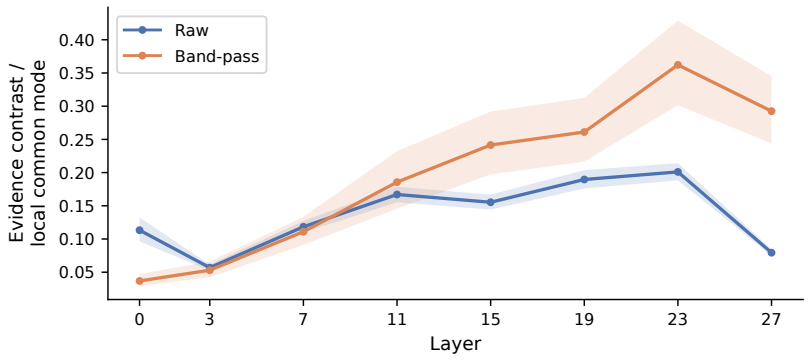

Figure 7: Evidence-contrast to local-common-mode ratio in the raw and band-pass filtered residual stream on the same 100 matched sources used in Figure 3. Points show source means at eight fixed layers; shading shows 95% source-bootstrap intervals.

Across these 100 sources, we average the raw ratio and query-relative support–control hidden contrast over the same eight prespecified layers within each source. Their Spearman correlation is $\rho = 0.285$ (95% source-bootstrap CI $[0.098, 0.451]$). Sources with stronger relative contrast tend to show clearer support–control separation at the answer-query state.

A local-gradient diagnostic differentiates the gold-over-counter mean full-answer margin with respect to the residual write on the same 100 source pairs. It finds positive alignment between the band-pass candidate and the answer-improving direction at the decision token (mean cosine 0.0295, 95% CI $[0.0231, 0.0356]$) and across local token positions (0.0117, 95% CI $[0.0093, 0.0139]$). The learned rank-8 write subspace captures 2.17% of decision-gradient energy (95% CI $[2.01\%, 2.33\%]$), compared with 0.38% for dimension-matched random subspaces. This probes task alignment at the all-writes-Off operating point.

C.6 Layerwise hidden-to-routing trajectory

Figure 8 gives trajectories across all 28 layers. Relative to Off, ResiBand increases the mean far-evidence residual-stream contrast in 22/28 layers, post-RoPE support–control score in 20/28, support–control attention mass in 20/28, and weighted-value contrast in 20/28. These paired trajectories trace the gold-versus-control difference from residual state to native routing and value delivery on the same 100 sources.

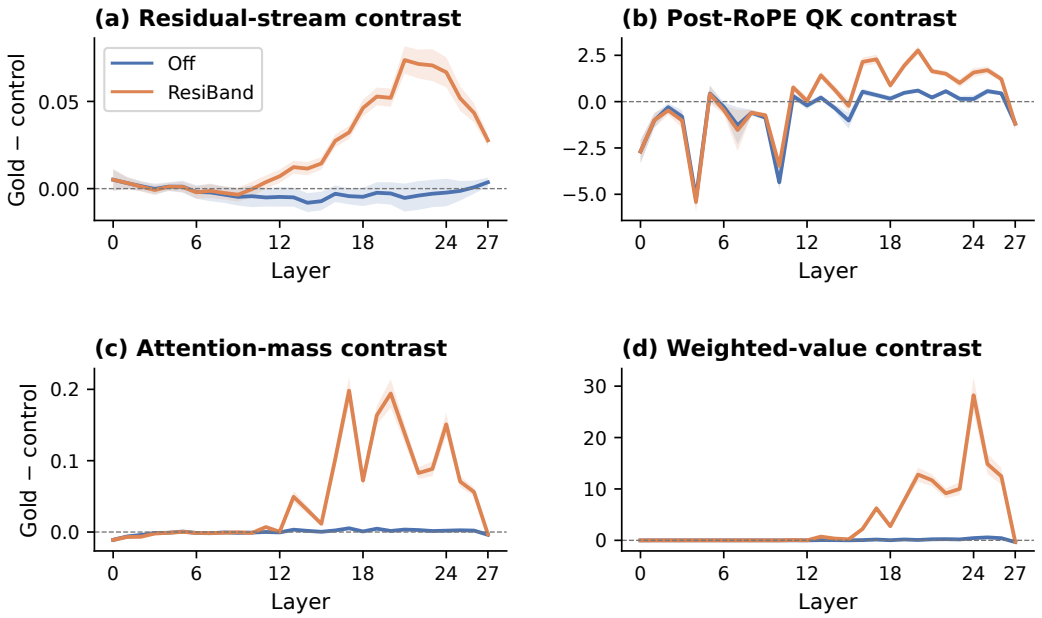

Figure 8: Complete 28-layer routing curves on the 100 matched 8K sources in Figure 3. Curves show source means and shaded 95% bootstrap intervals. All four panels compare gold support with equal-token matched control; higher contrast is better.

Figure 9 displays the paired On-minus-Off changes from the same 100 sources, making their signs visible across depth.

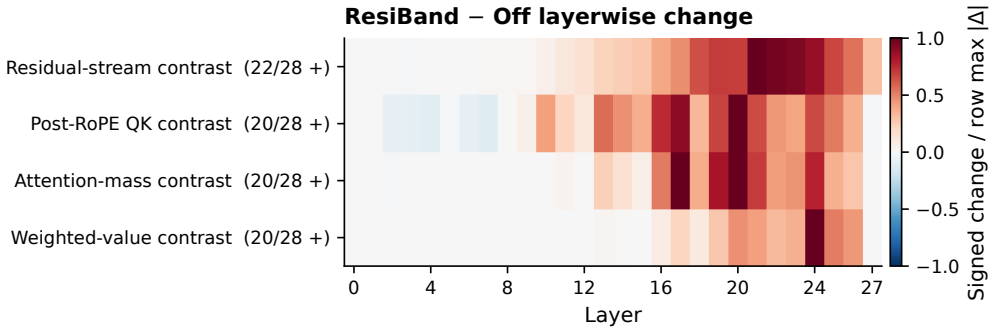

Figure 9: Layerwise changes in hidden, post-RoPE QK, attention, and weighted-value evidence contrasts. Each row is scaled by its own maximum absolute change for display; counts in parentheses use unscaled values.

C.7 Q/K/V FACTORIAL: A DISTRIBUTED NATIVE ANSWER PATH

To test how native channel states affect the mean full-answer margin, we use the same 100 far prompts as Figure 3. We run an otherwise-Off forward pass and replace one predeclared layer's native channels with the corresponding states from the natural RESIBAND trajectory. Q is patched at the final prompt position; K and V are patched over all prompt positions. The endpoint is the gold-versus-counter mean full-answer log-probability margin. Figure 10 shows all channel combinations at layers 11, 15, 19, and 23.

Table 8: Far-prompt mean full-answer margin change relative to Off for the complete Q/K/V factorial ($n = 100$ paired sources). Natural RESIBAND minus Off is $+3.917$ on these same sources.

Layer	Q	K	V	QK	QV	KV	QKV
11	-0.126	-0.176	-0.098	-0.154	-0.227	-0.336	-0.154
15	-0.018	$+0.128$	-0.191	$+0.095$	-0.086	-0.016	$+0.002$
19	$+0.284$	-0.030	$+0.086$	$+0.272$	$+0.410$	$+0.071$	$+0.453$
23	-0.001	$+0.054$	$+0.264$	-0.019	$+0.109$	$+0.156$	$+0.212$

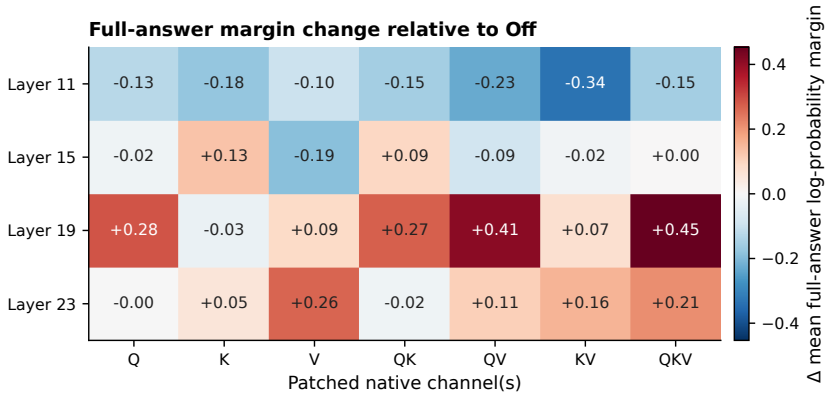

Figure 10: Q/K/V factorial on 100 far matched rows. Cells show changes in the mean full-answer margin relative to Off.

The largest local intervention is QKV at layer 19: $+0.453$ (95% CI $[0.294, 0.620]$), equivalent to 11.6% of the $+3.917$ natural Full-minus-Off gain on these same sources in Figure 3c. At that layer, Q alone gives $+0.284$ and QV gives $+0.410$; the largest V-only effect is $+0.264$ at layer 23. The variation across layers and channel combinations points to a distributed answer path.

C.8 CONTENT SELECTIVITY AND LOAD-BEARING EVIDENCE

On separate synthetic passcode prompts, the content-selectivity endpoint uses the mean full-answer log-probability margin for gold versus an irrelevant passcode at the same distance. Deletion specificity compares the margin loss from deleting gold support with the loss from deleting its matched control; a positive value means the gold support is more load-bearing. Table 9 and Figure 11 report both endpoints across six distances.

Table 9: Content and deletion specificity by target support distance ($n = 24$ paired prompts per distance for each endpoint).

Endpoint	Arm	64	128	256	512	1024	2048
Content margin	Off	0.254	0.285	0.337	0.277	0.224	0.244
	RESIBAND	1.064	1.143	1.159	1.151	1.136	1.155
Deletion specificity	Off	2.392	2.376	2.267	2.096	2.006	1.952
	RESIBAND	2.795	3.004	2.823	2.736	2.735	2.676

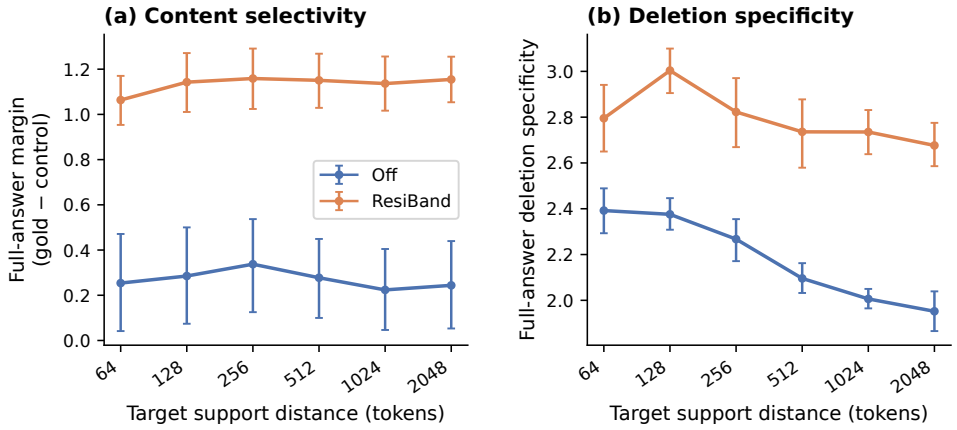

Figure 11: Matched-distance content selectivity and deletion specificity ($n = 24$ paired prompts per distance for each control). Error bars are 95% bootstrap intervals.

Pooled over distances, the mean full-answer content margin rises from 0.270 to 1.134, a paired gain of $+0.864$ (95% CI $[0.820, 0.908]$; 144 rows). Deletion specificity rises from 2.181 to 2.795, a paired gain of $+0.613$ ($[0.561, 0.663]$; 144 rows). Both controls indicate stronger use of task-relevant evidence than distance-matched irrelevant content.

C.9 Residual scale and direction controls

We intervene directly on the learned write using another synthetic passcode cohort: 24 source seeds, each tested at evidence distances 512 and 1024. We measure the change in the per-token mean full-answer log-probability margin for gold versus counter passcodes relative to the learned scale 1. Table 10 reports the paired means and intervals.

Table 10: Change in the per-token mean gold-versus-counter full-answer margin under residual-write interventions, relative to learned scale 1. Each distance uses $n = 24$ paired prompts; brackets show 95% paired-bootstrap intervals.

Intervention	Distance 512	Distance 1024
Reverse (-1)	-1.026 $[-1.212, -0.846]$	-1.030 $[-1.237, -0.823]$
Zero	-0.909 $[-1.032, -0.781]$	-0.955 $[-1.112, -0.796]$
Scale 0.25	-0.702 $[-0.789, -0.615]$	-0.738 $[-0.844, -0.633]$
Scale 0.5	-0.331 $[-0.398, -0.274]$	-0.303 $[-0.363, -0.251]$
Learned scale 1	0	0
Scale 1.5	$+0.012$ $[-0.016, +0.039]$	$+0.002$ $[-0.024, +0.027]$
Random, norm matched	-0.900 $[-1.021, -0.772]$	-0.961 $[-1.120, -0.797]$

Zeroing the write lowers the margin by 0.909 and 0.955 at the two distances. Attenuating or reversing it also lowers the margin. These results show that both write strength and learned direction matter.

C.10 Interpretation

Together, the diagnostics connect signal conditioning to evidence use. In the frozen Base model, nonlocal support remains useful as its contrast against matched controls weakens. The fixed band makes evidence contrast more prominent relative to the local common mode, and the learned write changes residual contrast, native QK scores, attention, and weighted-value delivery across layers. Q/K/V patches contribute to the answer margin, while content, deletion, and residual-write interventions show stronger use of task-relevant evidence and the importance of the learned write direction.