WHEN NATIVE FIT MISLEADS: EVALUATING ATTENTION SURROGATES UNDER POSITIONAL INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention surrogates offer interpretable approximations of attention mechanisms, but strong reconstruction of native attention may not imply faithful prediction under positional interventions. We study this gap using controlled position-ID shifts that modify the positional information of a sequence suffix within a single attention head while keeping tokens, causal masks, and upstream representations fixed. We compare two parameter-matched surrogate families for RoPE attention in Qwen3-1.7B-Base: a smooth positional-field surrogate (Cubic Field) and a learned-frequency rotary surrogate (Rotary Field). Although native-validation KL consistently selects Rotary Field, we identify reproducible heads on OpenWebText and WikiText-103 for which it is less faithful to positional interventions than both Cubic Field and a no-change baseline. The misranking persists in model-output evaluations, although the output-level differences are small. An exploratory extension to T5 suggests that this phenomenon is not limited to RoPE. These results demonstrate that native reconstruction fidelity alone does not guarantee positional-intervention fidelity and highlight the need to evaluate attention surrogates directly on the mechanistic responses they aim to explain.

1 INTRODUCTION

Transformer-based language models (Vaswani et al., 2017) have achieved remarkable success (Yang et al., 2025). In self-attention, explicit positional encoding (PE) provides information about token order, and substantial effort has focused on improving PE design and model performance (Su et al., 2024). Understanding how PE shapes attention can help identify its limitations, analyze redundant attention components (Michel et al., 2019; Voita et al., 2019), and inform targeted modifications.

An attention surrogate is an interpretable model that approximates selected computations or behaviors of an original attention module. Recent work explores attention replacement models (He et al., 2026), programmatic attention surrogates (Hayes et al., 2026), sparse reconstructions of attention outputs (Kissane et al., 2024), and learned gates for analyzing head contributions (Nam et al., 2025). These approaches evaluate mechanistic hypotheses through reconstruction fidelity, controlled interventions, and task-performance preservation. This motivates using surrogates to study positional sensitivity and predict how attention responds to changes in relative position.

Native-validation Kullback-Leibler (KL) divergence measures how closely a surrogate matches the teacher's attention distribution under the unmodified position map, and we use it to select checkpoints and methods. For positional intervention analysis, however, native fit alone does not establish that the surrogate captures how the head responds to changes in position. This motivates our central question: does the surrogate preferred by native-validation KL also most faithfully predict the specified head-local positional response?

To test this question, we use rotary position embeddings (RoPE) (Su et al., 2024), whose explicit mapping from position IDs to query-key rotations enables controlled interventions on positional information. To construct such an intervention, we shift the position IDs of a sequence suffix within a single head while holding tokens, the causal mask, and upstream representations fixed. Each surrogate predicts the resulting change in attention scores, which we add to the teacher's native scores to isolate response error in the evaluation (Fig. 1). We fit two surrogate families, a smooth positional-field

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.

surrogate (Cubic Field) and a learned-frequency rotary surrogate (Rotary Field), using native attention KL plus centered-score mean squared error (MSE), with nearly matched parameter counts and equal optimization budgets. Checkpoints and methods are selected by native-validation attention KL, then evaluated on held-out documents for positional-intervention fidelity.

Our results show that native KL does not always identify the surrogate most faithful to positional interventions. Native selection consistently favors Rotary Field across a fixed 32-head Qwen3-1.7B-Base (Yang et al., 2025) panel in each corpus. We identify three heads on OpenWebText (Gokaslan & Cohen, 2019) and four on WikiText-103 (Merity et al., 2017) for which this choice is less faithful than both Cubic Field and a no-change baseline; the original three cases replicate across both datasets. The aggregate misranking also persists in model-output evaluations. The output-KL differences are small, of order 10^{-5} nats/token, and a Cubic Field improvement over the no-change baseline is not established. These reproducible counterexamples show that native fit alone does not guarantee fidelity to the specified positional intervention. In addition, an exploratory T5 extension (Raffel et al., 2020), using additive bucketed relative-position biases, further indicates that this mismatch is not confined to RoPE. Our main contributions can be summarized as:

- We develop a controlled RoPE intervention framework with a shared teacher-score baseline and an exact error identity relating positional-response fidelity to reconstruction errors at native and intervened positions.
- We identify reproducible counterexamples showing that native KL can select the less intervention-faithful surrogate: three Qwen heads qualify on OpenWebText and four on WikiText-103, with the original three cases replicating across both datasets.
- We trace this misranking from attention responses to aggregate model-output effects and apply the evaluation to an exploratory T5 extension with additive bucketed positional biases.

2 Related Work

Attention analysis and explanatory models. Faithfulness concerns how an explanation reflects model behavior (Jacovi & Goldberg, 2020). Attention-head analyses identify syntactic and positional patterns (Clark et al., 2019; Voita et al., 2019) and assess head importance for task performance (Michel et al., 2019). Attention flow and circuit studies examine information flow and causal behavior (Abnar & Zuidema, 2020; Ferrando et al., 2022; Elhage et al., 2021; Olsson et al., 2022). Interchange interventions test causal abstractions (Geiger et al., 2021), while MIB benchmarks task-grounded variables and pathways (Mueller et al., 2025). Other approaches automate circuit discovery (Conmy et al., 2023) or approximate component effects through attribution patching (Kramár et al., 2024). Patching conclusions depend on intervention and metric choices (Zhang & Nanda, 2024); circuit overlap alone can misrepresent functional faithfulness (Hanna et al., 2024). Surrogate and learned-replacement work likewise distinguishes reconstruction from intervention faithfulness (Dyer et al., 2024; Makelov et al., 2025; Dunefsky et al., 2024; Ameisen et al., 2025).

Positional structure and RoPE. Relative position representations, Transformer-XL, RoPE, and linear biases inject order through different attention operators (Shaw et al., 2018; Dai et al., 2019; Su et al., 2024; Press et al., 2022). TUPE separately parameterizes content and positional score terms (Ke et al., 2021); KERPLE adds nonlinear relative-distance kernels (Chi et al., 2022). RoPE frequency studies connect bands to positional patterns, measure frequency utilization, and causally manipulate frequency access (Barbero et al., 2025; Oka et al., 2026; Urrutia et al., 2025). Related analyses derive particular RoPE attention structures or interpret rotation as phase modulation (Cheng et al., 2026; Liu, 2026); AdaRoPE learns head-specific frequency schedules and scales (Wang et al., 2026). RoPE's explicit position map provides a reproducible intervention for testing whether a native-fit surrogate preserves the induced response.

3 RoPE Intervention and Fidelity Metrics

Testing whether native selection transfers to an intervention requires a reference response that can be evaluated separately from reconstruction. RoPE supplies this reference because position IDs can change within one head while its other inputs remain fixed. We compare the predicted and exact score changes, then follow their difference into the model's output distribution (Fig. 1).

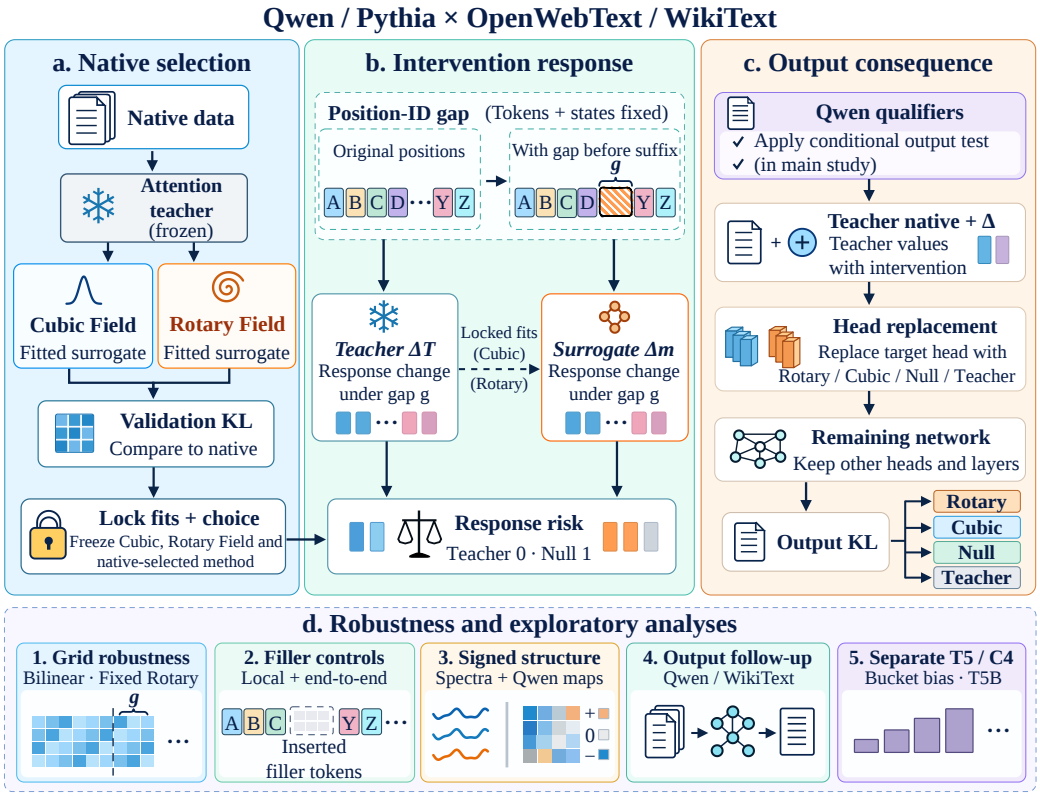

Figure 1: **Native selection and intervention fidelity.** (a) Native-validation KL selects checkpoints and method. (b) Centered responses use the teacher response and Null (no change). (c) Add predicted changes to native teacher scores before head replacement and output-KL evaluation. (d) Robustness and exploratory studies. Cubic/Rotary abbreviate the field names; Fixed Rotary fixes frequency locations. Hatching marks skipped positions.

3.1 An intervention with an exact response

For a fixed attention head in the original model (the teacher), let $q_i, k_j \in \mathbb{R}^{d_h}$ be the frozen pre-RoPE query and key vectors, where d_h is the dimension of one attention head, and let p_i be the position ID. Its score under our rotation convention is

$$S^T_{ij}(p) = \frac{q_i^\top R_\theta(p_j - p_i)k_j}{\sqrt{d_h}}. \tag{1}$$

We intervene on the selected head by changing its position map from p to p', while leaving the tokens, token order, upstream states, causal mask, and model parameters unchanged. Specifically, we shift the position IDs by a constant offset from a chosen cutoff onward. We parameterize this position-ID gap by (c, g), where c denotes the zero-based index of the first token in the shifted suffix (the gap position) and g denotes the added position-ID offset (the gap size). The resulting position map is $p'_i = p_i + g\mathbf{1}[i \geq c]$.

For a length-N sequence, we define the response of the teacher model ($\Delta_T \in \mathbb{R}^{N\times N}$) and surrogate model ($\Delta_m \in \mathbb{R}^{N\times N}$) to the intervention as the row-centered score difference on valid causal entries:

$$\Delta_T(p, p') = \mathcal{C}\big(S^T(p') - S^T(p)\big), \qquad \Delta_m(p, p') = \mathcal{C}(S^m(p') - S^m(p)). \tag{2}$$

Here, $\mathcal{C}(\cdot)$ is a row-centering operation defined as

$$\mathcal{C}(S)_{ij} = S_{ij} - \frac{1}{|\mathcal{V}_i|}\sum_{k\in\mathcal{V}_i} S_{ik}, \qquad j \in \mathcal{V}_i, \tag{3}$$

where $\mathcal{V}_i$ contains all valid causal keys for query i. Row-centering removes additive offsets that leave softmax attention unchanged (Blanchard et al., 2021).

3.2 Response Error and Normalized Risk

A response compares two score fields, so its error depends on reconstruction at both position maps. Let $e_m(p) = \mathcal{C}(S^m(p) - S^T(p))$ denote centered reconstruction error at position map p. By linearity:

$$\Delta_m(p, p') - \Delta_T(p, p') = e_m(p') - e_m(p). \tag{4}$$

For a length-N sequence with cutoff c, let $\mathcal{A} = \{(i, j) : c \leq i < N,\ 0 \leq j < c\}$ denote the affected pairs. Define $\langle X, Y \rangle_{\mathcal{A}} = \sum_{(i,j)\in\mathcal{A}} X_{ij} Y_{ij}$ and $\|X\|^2_{\mathcal{A}} = \langle X, X \rangle_{\mathcal{A}}$. All score errors are row-centered over valid causal keys before restriction to $\mathcal{A}$. Squaring Eq. 4 gives

$$\|\Delta_m(p, p') - \Delta_T(p, p')\|^2_{\mathcal{A}} = \|e_m(p')\|^2_{\mathcal{A}} + \|e_m(p)\|^2_{\mathcal{A}} - 2\langle e_m(p'), e_m(p) \rangle_{\mathcal{A}}. \tag{5}$$

We observe that the response error depends on the reconstruction errors of both position maps, as well as on their alignment. In contrast, native-validation attention KL measures agreement at unmodified positions rather than capturing these centered-score error components. Consequently, the identity alone is insufficient to determine the surrogate rankings.

To measure how accurately the surrogate predicts the effect of the intervention, we use the teacher's original score $S^T(p)$ as a fixed baseline. We then add the change predicted by the surrogate, $\Delta_m(p, p')$, to obtain the predicted intervened score $S^T(p) + \Delta_m(p, p')$. The teacher-response reference $S^T(p) + \Delta_T(p, p')$ differs from the teacher's true intervened score $S^T(p')$ only by a constant within each valid row, so they give the same attention distribution. The centered prediction error is therefore $\Delta_m(p, p') - \Delta_T(p, p')$, isolating error in predicting the intervention-induced change. Directly using the surrogate's intervened scores would also retain its native fitting error, since $e_m(p') = e_m(p) + \Delta_m(p, p') - \Delta_T(p, p')$.

Based on this isolated response error, we define a normalized response risk by dividing its squared magnitude by the teacher-response energy:

$$\mathcal{R}_m = \frac{\sum_{(i,j)\in\mathcal{A}} (\Delta_{m,ij}(p, p') - \Delta_{T,ij}(p, p'))^2}{\sum_{(i,j)\in\mathcal{A}} \Delta^2_{T,ij}(p, p')}. \tag{6}$$

This normalization provides a common baseline: when the teacher's response energy is positive, a perfect predictor has risk 0, while a Null predictor that always predicts no change has risk 1. A risk greater than 1 means that the surrogate's prediction is less accurate than simply predicting no change. For a fixed head and target, all surrogates share the same denominator, so response sum of squared errors (SSE) and normalized risk produce the same ranking. This risk measures how accurately the surrogate captures the teacher's response relative to the size of that response. It does not indicate the absolute magnitude of the teacher's response and is undefined when the teacher's response is zero.

3.3 Output-Level Fidelity Evaluation

Accurately predicting changes in attention scores does not necessarily lead to accurate predictions of the model's final output. This is because an attention head's output also depends on its value vectors, the output projection, and subsequent Transformer layers (Kobayashi et al., 2020; 2021). Previous work has studied how information propagates through these components using multilayer attribution, prediction-based decomposition, and value perturbations (Modarressi et al., 2022; 2023; Mohebbi et al., 2023). We therefore test whether the ranking of estimators at the score level also holds at the model-output level. For each estimator m, we construct the predicted intervened attention scores as $S^T(p) + \Delta_m(p, p')$. We convert these scores into attention weights and apply them to the teacher's unchanged value vectors. We then replace only the target head's output, before the output projection, while keeping all other heads unchanged at that site. The modified representation is subsequently passed through the rest of the network. We further define $p^{m,s}_{h,d,t}$ as the next-token distribution for head h, document d, and token position t, estimated by surrogate m fitted with seed s. Similarly, $p^T_{h,d,t}$ is the teacher's next-token probability distribution after the intervention. We define output-level error as the average KL divergence from the teacher model's distribution:

$$K_{m,s,h,d} = \frac{1}{|\mathcal{T}_d|} \sum_{t\in\mathcal{T}_d} D_{\mathrm{KL}}\left(p^T_{h,d,t} \,\|\, p^{m,s}_{h,d,t}\right). \tag{7}$$

Here, $\mathcal{T}_d$ is the set of evaluated prediction positions in document d. A smaller K indicates that the surrogate more closely reproduces the teacher's intervened output distribution.

3.4 Surrogates with different positional structure

Both surrogates predict the selected head's scores from frozen model hidden states $x_i \in \mathbb{R}^{d_{\text{model}}}$, where d_{model} is the full hidden-state dimension before projection into attention heads. We write their fitted score fields as $\widehat{S}^m(p) = S^m(p)$. Cubic Field projects x_i to $z_i = W_s x_i \in \mathbb{R}^{192}$ and combines the content-bilinear score $B_{ij} = z_i^\top A z_j$ with a query-conditioned positional field:

$$P_{ij} = a_i \bar{\delta}_{ij} + b_i \bar{\delta}_{ij}^2 + c_i \bar{\delta}_{ij}^3, \qquad \bar{\delta}_{ij} = \frac{p_j - p_i}{255}. \tag{8}$$

Three MLP heads map x_i to (a_i, b_i, c_i), with $b_i \le 0$ and $c_i \in (-1, 1)$. Writing Std_i for zero-mean row standardization over valid keys, with a 10^{-6} variance stabilizer, Cubic Field predicts

$$\widehat{S}_{ij}^{\text{Cubic Field}} = \alpha_c \,\mathrm{Std}_i(B)_{ij} + \alpha_p \,\mathrm{Std}_i(P)_{ij}. \tag{9}$$

We evaluate this score field at p and p' and center their difference to obtain the response in Equation 2.

Rotary Field uses asymmetric projections $u_i = W_Q x_i, v_j = W_K x_j \in \mathbb{R}^r$, with $r = 3180$, and 1,590 learned positive frequencies $\omega_k = \exp[-\,\mathrm{softplus}(\rho_k)]$:

$$\widehat{S}_{ij}^{\text{Rotary Field}} = \frac{\alpha_3}{\sqrt{r}} \left\langle \mathrm{RoPE}_\omega(u_i, p_i), \mathrm{RoPE}_\omega(v_j, p_j) \right\rangle. \tag{10}$$

The α coefficients are learned scalar gains in both models, and Rotary Field also learns its frequencies from an initialization independent of the teacher frequency bank. Both models apply the same causal mask and have nearly matched parameter counts: $13.025\mathrm{M}$ for Cubic Field and $13.027\mathrm{M}$ for Rotary Field, a difference of 0.012%.

Both methods fit native attention using KL and centered-score MSE; checkpoint selection uses native-validation attention KL. The response test then asks whether the selected positional structure transfers to the changed position map. Appendix A.2 gives the training objective and selector diagnostics.

4 Native Selection and Held-Out Intervention Tests

4.1 Experimental Design and Implementation

We first test whether native validation reliably selects the surrogate that best predicts a held-out positional intervention. The experiment crosses two models, Qwen3-1.7B-Base (Yang et al., 2025) and Pythia-1.4B at `step143000` (Biderman et al., 2023), with two corpora, OpenWebText (Gokaslan & Cohen, 2019) and WikiText-103 (Merity et al., 2017). Qwen is evaluated on 32 prespecified stress-test heads, while Pythia is evaluated on 12 heads formed by crossing layers $\{0, 11, 23\}$ with heads $\{0, 3, 7, 11\}$. The same head panels and experimental protocol are used across corpora. Qwen applies RoPE to the full head dimension; Pythia rotates 32 of 128 dimensions.

All examples are unpadded length-256 windows, with at most one window drawn from each source document. The held-out intervention inserts a position-ID gap after token 127 without changing the token sequence: $p'_i = i + 64\mathbf{1}[i \ge 128]$. This changes the relative positions of the 16,384 token pairs crossing the boundary. Their distances shift from 1–255 to 65–319, with 2,080 pairs (12.7%) extending beyond the native distance range. We also evaluate the identity operator and the common translation $p'_i = i + 256$ as operator-preserving controls.

Rotary Field and Cubic Field are fitted independently for each model–corpus setting using three seeds (42–44) and 30,000 optimization steps. The two methods share examples, batches, evaluation schedules, and checkpoint rules, but retain their method-specific learning-rate schedules. For each method and seed, we select the checkpoint with the lowest native-validation KL before examining intervention responses. Fitting, validation, response, and output documents are mutually disjoint, and each corpus replication uses fresh fits and document assignments.

A Qwen head is classified as a score-level counterexample only if, for every seed, (i) native-validation and held-out native KL favor Rotary Field, (ii) held-out response risk favors Cubic Field, and (iii)

$$\mathcal{R}_{\text{Rotary Field}} - 1 \ge 0.10, \qquad \mathcal{R}_{\text{Rotary Field}} - \mathcal{R}_{\text{Cubic Field}} \ge 0.10. \tag{11}$$

Table 1: Native selection and intervention fidelity across the four settings. Native validation selects Rotary Field on every head. Score counterexamples satisfy all seedwise and simultaneous-bound criteria. Output KL is conditional on those heads: K_N, K_A and K_0 are output KL for native-selected Rotary Field, alternative Cubic Field and Null, in 10^{-6} nats/token. Both Qwen comparisons support $K_N > K_A$ and $K_N > K_0$; neither establishes $K_A < K_0$.

Setting	Native choice	Score counterexamples	K_N	K_A	K_0
Qwen / OpenWebText	Rotary Field, 32/32	3/32	20.197	3.106	3.619
Qwen / WikiText-103	Rotary Field, 32/32	4/32	17.302	7.382	9.046
Pythia / WikiText-103	Rotary Field, 12/12	0/12	—	—	—
Pythia / OpenWebText	Rotary Field, 12/12	0/12	—	—	—

Thus, the native-selected method must underperform both Cubic Field and the Null predictor, with positive familywise one-sided 95% lower confidence bounds for both contrasts. Inference uses 10,000 paired document-bootstrap replicates, averaging seed-level scores and ensuring simultaneous coverage within each fixed head panel. For Pythia, we also compare validation-KL, training-MSE, and Pareto selectors with fixed-method baselines and the response oracle.

We then test whether Qwen's score-level counterexamples persist after propagation through the rest of the model. Without refitting or reselection, each qualifying head is evaluated on 64 untouched output documents per corpus. Each document contributes its mean next-token KL over positions $\mathcal{T}_d = \{1, \ldots, 255\}$. Let K_N, K_A, and K_0 denote the aggregate output KL of the native-selected Rotary Field, the alternative Cubic Field, and the Null predictor, respectively. We bootstrap the paired contrasts: $K_N - K_0$, $K_N - K_A$, $K_0 - K_A$, and report the normalized ranking gain: $G = \frac{K_N - K_A}{K_0}$. The output test is conditional on passing the score-level criteria and therefore applies only to Qwen. Further implementation details are provided in Appendices A and L.

4.2 Results and Analysis

Native validation selects Rotary Field in all four settings. Table 1 contrasts this selection with score-response accuracy and conditional output fidelity.

On Qwen/OpenWebText, native validation selects Rotary Field for all 32 heads, but this ranking fails on three: L24H10, L25H12, and L21H13. Their Rotary Field response risks range from 2.10 to 3.26, making the selected method worse than both Cubic Field and the Null predictor. The two required response contrasts are consistent across all three seeds and have positive simultaneous lower confidence bounds. A fourth head, L18H02, shows the same ranking reversal but does not pass the prespecified Rotary-versus-Null harm threshold (see Fig. 2).

These reversals are concentrated in the prespecified stress-test subset: all four occur among six heads identified before the confirmation experiment, whereas none occurs among the other 26. On the three complete counterexamples, Cubic Field achieves response risks of 0.944–0.982 and therefore modestly improves on the Null predictor at the score level.

The main finding replicates on WikiText-103 under fresh fits and document assignments. The three OpenWebText counterexamples again satisfy all criteria, and L18H02 also becomes a complete counterexample. Native selection can therefore fail reproducibly for particular Qwen heads rather than only for a specific corpus split or optimization run.

Pythia provides a contrasting result. Rotary Field has lower native-validation KL, held-out native KL, and intervention-response risk for every tested head and seed on both corpora. On OpenWebText, its mean response risk is 0.365, compared with 1.295 for Cubic Field. All three selectors: validation KL, training MSE, and Pareto selection, agree with the response oracle on all 12 heads and incur zero selection regret. Thus, the positional intervention does not inherently break native selection; the failure is specific to the tested Qwen heads.

Qwen's aggregate misranking persists through the remaining network. Across three OpenWebText counterexamples, native-selected Rotary Field exceeds Cubic Field and Null in output KL by 1.709×10^{-5} and 1.658×10^{-5} nats per token, respectively, with positive simultaneous lower bounds. Normalized centered-logit error yields the same ordering (Fig. 2). WikiText-103 shows the same

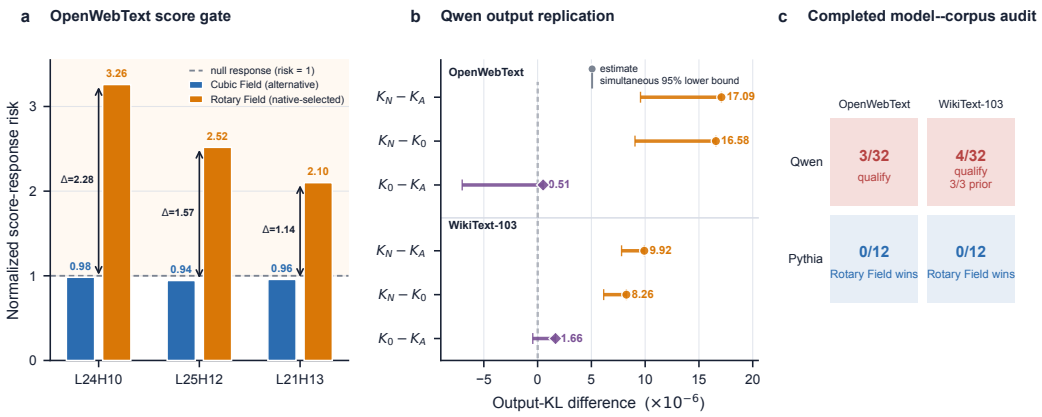

Figure 2: **The Qwen counterexample survives a corpus change.** (a) OpenWebText response risk on three qualifying heads; dashed line: Null risk 1; arrows: Rotary Field–Cubic Field gaps. (b) Output-KL contrasts with simultaneous one-sided 95% lower bounds on 64 untouched documents per corpus support $K_N > K_A$ and $K_N > K_0$, without establishing $K_A < K_0$. (c) Qualifying heads: Qwen 3/32 on OpenWebText and 4/32 on WikiText-103; Pythia 0/12 on each corpus.

aggregate pattern across four counterexamples, although L18H02 does not individually preserve the harmful direction. Appendices report headwise results; Table1 summarizes aggregate output KL.

Together, these results establish a reproducible but localized failure of native model selection. For several Qwen heads, native KL consistently selects a surrogate that is worse under the held-out intervention, and the aggregate error remains detectable at the model output. The Pythia results show that this failure is not an automatic consequence of the intervention. They do not, however, identify its cause because the models differ jointly in weights, training history, tokenizer, and rotary width.

The practical magnitude of the output effect is small: the excess KL is on the order of 10^{-5} nats per token. Moreover, although Cubic Field corrects the ranking failure, neither corpus shows a reliable output-level improvement over the Null predictor. The evidence therefore supports a failure of native selection, not a claim that Cubic Field is generally superior. Because the analysis uses fixed head panels, it also identifies counterexamples rather than estimating their prevalence across heads.

5 Sensitivity Analysis

We then examine whether anchor misranking persists across gap geometries and extends from position-ID gaps to visible insertions that alter upstream representations. The expanded grid crosses $c \in \{64, 128, 192\}$ with $g \in \{8, 16, 32, 64, 128\}$, yielding 15 targets for Qwen/WikiText and both Pythia corpora. Each panel uses a separate 512-document response cohort. Position-blind Bilinear Field and teacher-frequency Fixed-frequency Rotary Field serve as controls. A separate filler study compares masked, real, and collapsed donor-token insertions.

The three previously identified Qwen failure heads support both Rotary Field risk contrasts at all 15 grid targets (Table 2), so their misranking is not specific to $(c, g) = (128, 64)$. This result applies to the fixed three-head panel: among all 480 Qwen head–target cells, 51 have two positive point contrasts, 419 have two nonpositive contrasts, and ten have mixed signs. These counts are descriptive rather than simultaneous discoveries. Both Pythia panels remain adequately nonreversing throughout the grid (Appendix G.1).

Response magnitude and alignment help interpret these errors (see Fig. 3). At the anchor, Rotary Field predicts response magnitude on all three Qwen heads, with direction cosines of 0.371–0.475. Cubic Field produces smaller responses but has negative direction cosines on two heads, so lower response error does not imply better directional recovery. Fixed-frequency Rotary improves on Rotary Field but remains worse than Null, while signed spectra do not establish between-band cancellation as the mechanism (Appendices G.2 and I.2).

Table 2: **The three-head Qwen contrast persists across the intervention grid.** Entries are equal-head mean risk differences (simultaneous one-sided 95% lower bounds) on Qwen/WikiText. All 30 contrasts share the grid family; c is the cutoff and g the gap[†].

	$c = 64$		$c = 128$		$c = 192$	
g	Rotary-Cubic	Rotary-Null	Rotary-Cubic	Rotary-Null	Rotary-Cubic	Rotary-Null
8	1.419 (1.376)	1.432 (1.388)	1.250 (1.210)	1.255 (1.215)	1.148 (1.108)	1.151 (1.111)
16	0.902 (0.870)	0.926 (0.894)	0.794 (0.763)	0.805 (0.774)	0.703 (0.672)	0.711 (0.679)
32	0.670 (0.639)	0.723 (0.692)	0.594 (0.563)	0.624 (0.593)	0.482 (0.448)	0.502 (0.467)
64	0.442 (0.415)	0.515 (0.487)	0.384 (0.355)	0.433 (0.402)	0.248 (0.214)	0.280 (0.245)
128	0.567 (0.534)	0.656 (0.620)	0.545 (0.506)	0.623 (0.581)	0.416 (0.371)	0.468 (0.420)

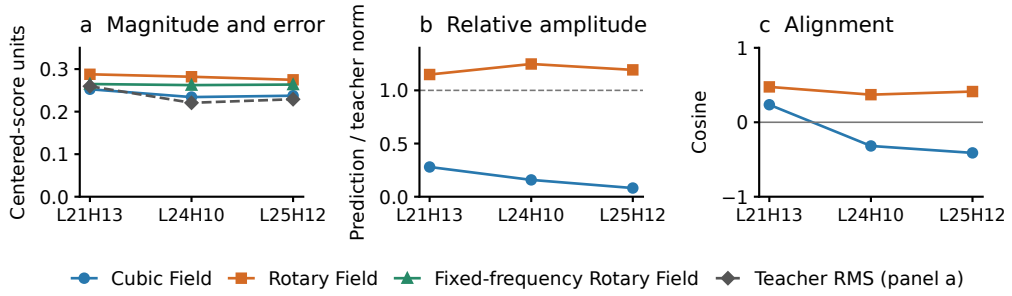

Figure 3: Response geometry of three Qwen failure heads at $(c, g) = (128, 64)$ on a separate 512-document cohort: (a) teacher-response RMS (dashed) and prediction RMSE; (b) prediction-to-teacher response norm ratio; and (c) cosine similarity between predicted and teacher responses. Seed responses are ensembled before scoring. See Appendix I.1 for full grids and definitions.

With visible insertion, Qwen meets the combined failure criterion at ten of 15 head-local targets, but not at the collapsed-state anchor or any end-to-end target. Thus, misranking persists across gap geometries and many local insertion responses but does not propagate end to end. Robustness to position-ID gaps therefore does not imply robustness when tokens and upstream representations also change. Complete designs and results appear in Appendices G–J.

6 Case Study: Diagnosing the Qwen Failures

The sensitivity analysis establishes that the Qwen misranking persists across gap geometries but does not explain its mechanism. We therefore test whether it reflects a mismatch between the frequency content of the teacher and the native-selected Rotary Field surrogate.

Using 64 native-validation documents, the prespecified diagnostic compares teacher and Rotary Field band power, weighted by $|H_{64}(\omega)|^2$ to emphasize frequencies affected by the anchor gap. The resulting discrepancy is summarized by the log-frequency Wasserstein distance W_h. We test whether W_h is positively associated with intervention-response risk and whether it separates the qualifying failure heads from the remaining Qwen heads. Signed spectra and response-geometry analyses supplement this unsigned scalar diagnostic in Appendix I.

The diagnostic satisfies neither prespecified criterion. Across the Qwen panel, W_h is not positively associated with intervention-response risk (Spearman $\rho = -0.107$, one-sided $p = 0.724$) and does not reliably separate the failure heads (AUC $= 0.759$, $p = 0.0819$) (See Appendix I, Fig. 5). Thus, scalar teacher–surrogate frequency discrepancy does not explain the Qwen misranking. Although W_h discards component signs and alignment, the supplementary signed analyses also do not establish between-band cancellation as the mechanism. This case study therefore rejects the proposed scalar

[†]All 15 Qwen targets pass with 3/3 supporting heads. In each Pythia corpus, every target has 12/12 material heads and both simultaneous upper bounds below zero across the 60-contrast family, exceeding the adequate-nonreversal criteria (bounds ≤ 0.10; at least 9/12 material heads). Results are for fixed panels.

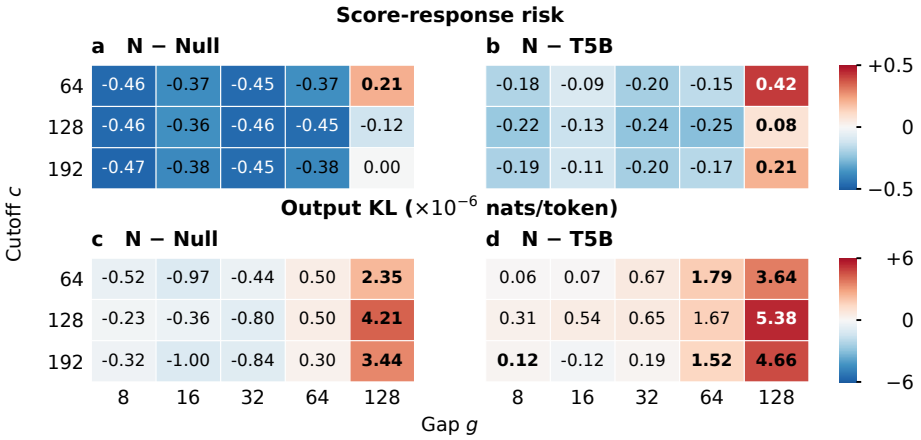

Figure 4: T5 native-policy contrasts across the 15-target grid. (a,b) Score-response risk over 32 heads; (c,d) output KL over 3 heads, in 10^{-6} nats/token. N: native policy; T5B: bucket-aware bilinear field. Entries are equal-head means. Positive values favor the comparator; bold marks positive simultaneous 95% lower bounds (two-sided response intervals; one-sided output bounds).

explanation without identifying an alternative mechanism. Full definitions, head-level results, and descriptive Pythia comparisons appear in Appendices F and I.

7 T5 Extension and Native-Data Predictions

We extend the analysis to T5-large/C4 English, which uses additive bucketed positional bias. At the same 15 (c, g) targets, we evaluate 32 response heads and three output heads, comparing the native-selection policy with fixed alternatives including Bucket-aware Bilinear Field (T5B). We also use three native-data prediction cards to test whether fitted surrogates anticipate intervention responses without teacher-response data.

At 13 targets, simultaneous upper bounds limit the native policy's equal-head mean response-risk excess to 0.10 relative to all five fixed alternatives (see Fig. 4); the exceptions are $(64, 128)$ and $(192, 128)$. No target meets the combined selection-failure rule, while T5B meets its recovery rule at $(192, 128)$. Although native-minus-Null output KL has positive lower bounds at all three $g = 128$ targets, teacher-to-Null KL is only 0.995–3.140×10^{-6} nats per valid decoder token, and none of the 60 comparisons meets the combined ranking or material-propagation rule.

The prediction cards achieve categorical agreement of $3/4$, $3/4$, and $0/4$, with all output-direction predictions abstaining. Thus, T5 exhibits no combined Qwen-like selection failure, but native-data diagnostics do not reliably predict intervention-response structure. Because model, corpus, task, and positional mechanism all differ, this comparison does not attribute the contrasting pattern to architecture. Complete results and audits appear in Appendix K.

8 Conclusion

In this paper, we show that native attention reconstruction and intervention fidelity can rank the same surrogates differently. The selected Qwen contrasts persist across corpora and gap configurations, and conditional output tests retain the aggregate misranking with small absolute effects. These findings motivate evaluating positional surrogates on intervention responses and downstream effects alongside reconstruction accuracy.

The end-to-end filler tests do not meet the combined failure criterion. Thus, the pure position-gap findings do not establish selection failure under token insertion, which also changes upstream representations. The contrast motivates further study of surrogate fidelity under interventions that modify both positional and contextual information.

Ethics statement This work analyzes frozen public language models on public text corpora and does not collect human-subject data or deploy a system. Public text can contain bias, offensive material, and personal information. Experiments process the text locally; tracked and released artifacts contain derived numerical records and do not redistribute raw corpus text.

Reproducibility statement The anonymous supplement provides numerical summaries, figure inputs, study-specific scientific source, and presentation scripts for the main four-setting study, expanded robustness and diagnostic studies, Qwen output follow-up, and separate T5 extension. Its CPU entrypoint regenerates tables and quantitative figures from the retained estimates and uncertainty fields. The supplementary README provides execution instructions, and Appendix L describes the included materials and additional inputs required for experimental or statistical reruns.

AI USE STATEMENT

Generative AI tools assisted with literature search and synthesis, code implementation, execution and debugging, figure generation, and manuscript polishing. The mathematical ideas and proofs, final scientific analyses, interpretations and conclusions are the authors' work. The authors take responsibility for the final content. Empirical evaluation texts were drawn from the specified corpora; no generated text was substituted for them.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, pp. 92408–92438, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 2021. doi: 10.1093/imanum/draa038. URL https://doi.org/10.1093/imanum/draa038.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Yuan Cheng, Fengzhuo Zhang, Yunlong Hou, Cunxiao Du, Chao Du, Tianyu Pang, Aixin Sun, and Zhuoran Yang. Demystifying the slash pattern in attention: The role of RoPE, 2026. URL https://arxiv.org/abs/2601.08297.

Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. KERPLE: Kernelized relative positional embedding for length extrapolation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8386–8399. Curran Associates, Inc., 2022. doi: 10.52202/068431-0610. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/37a413841a614b5414b333585e7613b8-Abstract-Conference.html.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24375–24410, 2024. doi: 10.52202/079017-0768. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2b8f4db0464cc5b6e9d5e6bea4b9f308-Abstract-Conference.html.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for complex simulation models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21814–21841, 2024. doi: 10.52202/079017-0686. URL https://papers.nips.cc/paper_files/paper/2024/hash/26b8e3dc3a21fcd660d80c63b767f324-Abstract-Conference.html.

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, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8698–8714, 2022. doi: 10.18653/v1/2022.emnlp-main.595. URL https://aclanthology.org/2022.emnlp-main.595.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

Amiri Hayes, Belinda Z. Li, and Jacob Andreas. Explaining attention with program synthesis, 2026. URL https://arxiv.org/abs/2606.19317.

Zhengfu He, Junxuan Wang, Rui Lin, Xuyang Ge, Wentao Shu, Qiong Tang, Junping Zhang, and Xipeng Qiu. Towards understanding the nature of attention with low-rank sparse decomposition. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 78269–78297, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7e57d4c3bb2bd2396a551e7aeea19aed-Paper-Conference.pdf.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=09-528y2Fgf.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.17759.

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

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4547–4568, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.373. URL https://aclanthology.org/2021.emnlp-main.373/.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Feilong Liu. Rotary positional embeddings as phase modulation: Theoretical bounds on the RoPE base for long-context transformers, 2026. URL https://arxiv.org/abs/2602.10959.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, pp. 33588–33636, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/53356aebeea8ffd40a8ac3bb66243162-Paper-Conference.pdf.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 258–271, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.19. URL https://aclanthology.org/2022.naacl-main.19/.

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

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3378–3400, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.245. URL https://aclanthology.org/2023.eacl-main.245/.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Andrew J. Nam, Henry C. Conklin, Yukang Yang, Thomas L. Griffiths, Jonathan D. Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 115609–115631. Curran Associates, Inc., 2025. doi: 10.52202/085713-3856. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a7f530e11fa19e9551b7a51dbd0f336f-Paper-Conference.pdf.

Yui Oka, Kentaro Hanafusa, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Probing rotary position embeddings through frequency entropy. In *The Fourteenth International Conference on Learning Representations*, pp. 6139–6167, 2026. URL https://openreview.net/forum?id=1JZuEDq62N.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468.

Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling positional and symbolic attention behavior in transformers, 2025. URL https://arxiv.org/abs/2511.11579.

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*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

Shaowen Wang, Yuke Zheng, Tansheng Zhu, Shuang Chen, Shaofan Liu, Suncong Zheng, and Jian Li. AdaRoPE: Not all attention heads should rotate and scale equally, 2026. URL https://arxiv.org/abs/2607.19363. Accepted at ICML 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

APPENDIX

TABLE OF CONTENTS

SETUP

A Experimental Setup, Head Panels, and Study Scope **16**

A.1 Fitting and evaluation details 18

A.2 Operator response and native fitting 20

MAIN FOUR-SETTING RESULTS

B Qwen/OpenWebText Counterexample Details **21**

C Qwen/OpenWebText Output Details **21**

D Qwen/WikiText-103 Replication Details **22**

E Pythia Selected-Checkpoint and Per-Head Results **23**

F Gap-Weighted Spectral Diagnostic **25**

ROBUSTNESS AND DIAGNOSTIC STUDIES

G Intervention-Grid Robustness **26**

G.1 Fresh anchor, grid and full-panel results 27

G.2 Descriptive four-method response comparison 29

H Filler and Materiality Controls **31**

H.1 Materiality and insertion results 31

I Signed Spectral and Spatial Diagnostics **35**

I.1 Response magnitude and alignment 35

I.2 Signed bands and spatial structure 36

J Output-Propagation Case Study **46**

J.1 Descriptive output-grid results 46

EXPLORATORY EXTENSION AND MATERIALS

K Exploratory T5 Extension **48**

K.1 Design and inference 48

K.2 Score-response results 48

K.3 Output consequences 51

K.4 Native-data prediction audits 53

L Reproducibility and Evidence Scope **54**

A Experimental Setup, Head Panels, and Study Scope

The main four-setting study tests native selection in Qwen and Pythia on both corpora; Table 3 records its document roles and fitting budgets. The follow-up RoPE designs use separate 512-document response cohorts for Qwen/WikiText-103, Pythia/OpenWebText and Pythia/WikiText-103 (Appendices G–I). These cohorts are not pooled with the main study. The Qwen output follow-up has its own cohort (Appendix J), and T5/C4 is a separate exploratory extension (Appendix K).

Table 3: Data roles and budgets in the main four-setting study. All four settings give both methods a 30,000-step budget and evaluate the unique minimum-native-validation checkpoint selected before response access. Model-output evaluation opens only for qualifying heads.

Qwen

	Qwen / OpenWebText	Qwen / WikiText-103
Model	Qwen3-1.7B-Base	Qwen3-1.7B-Base
Rotary Field geometry	full head width	full head width
Panel heads	32	32
Fitting seeds	42, 43, 44	42, 43, 44
Score docs. (fit/val./response)	256/64/128	256/64/128
Model-output documents	64 untouched; 3 heads	64 untouched; 4 heads
Cubic Field maximum steps	30,000	30,000
Rotary Field maximum steps	30,000	30,000
Response access	after 192 qualified fits	after 192 qualified fits

Pythia

	Pythia / WikiText-103	Pythia / OpenWebText
Model	Pythia-1.4B	Pythia-1.4B
Rotary Field geometry	32 of 128 dimensions	32 of 128 dimensions
Panel heads	12	12
Fitting seeds	42, 43, 44	42, 43, 44
Score docs. (fit/val./response)	256/64/64	256/64/64
Model-output documents	0 qualifying heads	0 qualifying heads
Cubic Field maximum steps	30,000	30,000
Rotary Field maximum steps	30,000	30,000
Response access	after 72 qualified fits	after 72 qualified fits

Separation of fitting, selection and evaluation. Each Qwen comparison uses fresh document roles and fresh Cubic Field/Rotary Field fits. Fitting, qualification, and native selection precede score-response access, and the three OpenWebText heads and four WikiText-103 heads are fixed before their untouched output roles open. The WikiText implementation changes only the numerical common-translation gauge while preserving the operator and decision rules. In Pythia, the minimum-validation checkpoint rule is fixed before either response evaluation; final-step drift is reported as a non-gating diagnostic.

We write `LxxHyy` for zero-indexed layer `xx` and head `yy`; for example, `L00H01` denotes layer 0, head 1. The Qwen panel is

L00H01	L03H01	L06H04	L02H07	L04H08	L00H09	L03H12	L03H15
L13H03	L11H00	L09H05	L10H07	L09H11	L07H10	L07H14	L11H15
L14H01	L18H02	L16H05	L19H07	L19H09	L18H09	L18H15	L15H12
L21H02	L26H01	L27H07	L26H07	L25H09	L24H10	L25H12	L21H13

It contains two heads from each depth-bin–head-index-bin cell after excluding the discovery panel. The original six-head aggregate comprises L03H12, L13H03, L18H02, L21H13, L24H10 and L25H12, all within this 32-head panel. They were chosen from earlier repaired-Rotary Field reversals and fixed before the confirmation data were evaluated. They differ from the later six-head spectral panel described below. Each Qwen schedule contains 192 fresh qualified fits and 832 records. On OpenWebText, native validation selects Rotary Field on all 32 heads and retains the same sign on held-out native evaluation. On WikiText-103, all three frozen qualifiers replicate and L18H02 additionally passes the score gate.

The Pythia panel is

L00H00	L00H03	L00H07	L00H11	L11H00	L11H03
L11H07	L11H11	L23H00	L23H03	L23H07	L23H11

Each Pythia confirmation contains 72 fits, 12 frozen native choices, 72 held-out native evaluations, 72 translation controls, 72 internal-gap responses, and 12 teacher-response informativeness records, for 312 records total: 72 fit records and 240 selection/evaluation records. The OpenWebText confirmation reuses no WikiText fit, checkpoint, metric, choice, or record.

Panels for the later studies. Table 4 distinguishes the broad response panels from the smaller panels used to examine insertion effects and response structure. Qwen filler comparisons target the three original OpenWebText counterexample heads, L21H13, L24H10, and L25H12. Pythia uses the predeclared L00H00, L11H07, and L23H11 controls, spanning early, middle, and late layers. These comparisons examine insertion effects in the targeted cases; the broader response panels describe variation across the larger fixed head sets. Qwen spectra add the historical success controls L25H09, L26H01, and L27H07, whereas Pythia retains its three fixed controls in each corpus. Their different sizes reflect these diagnostic roles, and inference remains conditional on the specified panels.

Table 4: Head coverage of the supplementary autoregressive-model studies. Qwen denotes WikiText-103; Pythia uses a separate panel evaluation on each corpus. Spatial case counts include repeated targets for L21H13.

Study	Qwen	Pythia, per corpus
Broad response grid	32 heads	12 heads
Filler comparisons	3 failure heads	3 fixed controls
Spectral diagnostics	6 heads: 3 failure + 3 success	3 fixed controls
Spatial response fields	6 heads, 8 cases	Not prescribed
Downstream output cases	6 heads, 48 cases	Not prescribed

A.1 FITTING AND EVALUATION DETAILS

Optimization. All four settings fit the surrogates in float32 with Adam, batch size four, and a maximum of 30,000 steps. The learning rate increases linearly from 10^{-6} over 1,000 warmup steps to 10^{-3} for Cubic Field or 10^{-4} for Rotary Field, then follows cosine decay to 10^{-6}. Gradient norms are clipped at 1.0, and the centered-score MSE coefficient is 0.02. Adam's remaining options use the library defaults. Validation is performed at step 1 and every 500 steps; the selected checkpoint minimizes native-validation attention KL. Qwen additionally rejects a fit that exceeds 600 seconds; Pythia has no per-fit time cap.

Shared intervention and data separation All four score-response settings use length-256 causal sequences with native positions $p_i = i$. The intervention adds a position-ID gap after token 127 without inserting a token,

$$p'_i = i + 64\mathbf{1}[i \geq 128], \tag{12}$$

and evaluates the affected pairs $\mathcal{A} = \{(i, j) : 128 \leq i \leq 255,\ 0 \leq j < 128\}$. Identity and common translation $p'_i = i + 256$ are controls. Fitting, validation, score-response, and output documents are disjoint, with at most one no-padding window per source document. Table 3 summarizes the settings and schedules.

Qwen/OpenWebText supplies the prospective 32-head confirmation, and WikiText-103 repeats its panel, schedules, selector and decision rules. Pythia/WikiText-103 fixes a 12-head evaluation before response access; OpenWebText repeats it under the same within-model controls. These comparisons test corpus stability at a fixed intervention.

Qwen/OpenWebText fitting and counterexample criteria We use Qwen3-1.7B-Base at the revision recorded in the supplement (Yang et al., 2025) and OpenWebText (Gokaslan & Cohen, 2019). The 32-head panel is the fixed stress-test panel retained from development, not a prevalence sample. Fresh roles exclude every previously used Qwen document and contain 256 fitting, 64 native-validation, 128 score-response, and 64 output documents. Cubic Field and Rotary Field are fitted from scratch for 30,000 steps using the same examples, batches, evaluation cadence, checkpoint rule, and three seeds; each method uses its fixed learning-rate schedule. The unique minimum-native-validation-KL checkpoint is selected before response access; all 192 fits must meet the fitting criteria.

A head qualifies as a counterexample only when native validation and held-out native KL favor Rotary Field for every seed, response risk favors Cubic Field for every seed, $\mathcal{R}_{\text{Rotary Field}} - 1 \geq 0.10$, and $\mathcal{R}_{\text{Rotary Field}} - \mathcal{R}_{\text{Cubic Field}} \geq 0.10$. Familywise one-sided 95% lower bounds for both contrasts must also exceed zero. A joint 10,000-replicate paired document bootstrap covers both contrasts for all 32 heads and the two aggregate contrasts over the six previously frozen Qwen heads. Seeds are averaged before inference; heads are not treated as replicates.

Qwen/WikiText-103 replication The replication retains the Qwen revision, fixed 32-head panel, methods, three fitting and loader seeds, 30,000-step schedules with batch size four, checkpoint rule, gap, thresholds, and bootstrap. Fresh WikiText-103 roles contain 256 fitting, 64 native-validation, 128 score-response, and 64 conditional-output documents; they exclude all documents and token windows used by the earlier Qwen and Pythia WikiText workflows. The same counterexample criteria are applied without selecting new heads or changing the effect floors.

The replication represents Rotary Field rotary phase relative to the first position coordinate. This numerical gauge preserves every relative displacement, restores the common-translation control, and reproduces the original native and internal-gap maps that begin at zero. All 192 fits must qualify before score-response evaluation; the three OpenWebText qualifiers form a fixed replication stratum within the full panel.

Pythia/WikiText-103 evaluation This setting uses Pythia-1.4B at revision `step143000` (Biderman et al., 2023) and WikiText-103 (Merity et al., 2017). Its fixed panel crosses layers $\{0, 11, 23\}$ with heads $\{0, 3, 7, 11\}$, and Pythia rotates only 32 of each 128-dimensional head. Both surrogates receive 30,000 steps for three seeds. The evaluated estimator is the unique checkpoint with minimum native-validation attention KL, fixed before evaluating the 64 held-out response documents. Native KL is also evaluated on the held-out response documents. Response uncertainty uses 10,000 paired

document bootstrap replicates and Bonferroni-adjusted 95% simultaneous intervals across the 12 heads.

Pythia/OpenWebText cross-corpus replication The OpenWebText replication uses the same Pythia revision, 12-head panel, estimators, three seeds, 30,000-step schedules, minimum-native-validation checkpoint rule, intervention, controls, and simultaneous interval procedure. Disjoint 256/64/64 OpenWebText document roles replace WikiText-103 for fitting, native validation, and sealed score response. The response documents were accessed only after all 72 fresh fits qualified and their native choices were fixed.

Conditional Qwen output confirmations Each score protocol fixes the heads that pass every criterion before output access: three on OpenWebText and four on WikiText-103. Without refitting, all three selected checkpoints per method are evaluated on 64 untouched length-256 source documents for the corresponding corpus. The intervention replaces only the target head's attention-weighted teacher-value slice before the output projection, retaining tokens, upstream states, values and non-target heads at the replacement site while later activations propagate the change through the fixed model. Within each corpus, heads and documents receive equal weight. A joint 10,000-replicate paired document bootstrap tests $K_N - K_0$, $K_N - K_A$, and $K_0 - K_A$. In both corpora, the output-ranking rule requires $G = (K_N - K_A)/K_0 \geq 0.10$, a positive $K_N - K_A$ point estimate and a positive simultaneous lower bound. The stronger decision additionally requires positive contrasts and lower bounds for $K_N - K_0$ and $K_0 - K_A$. The stronger rule tests an alternative's benefit over Null as well as native-selection misranking. These confirmations test whether demonstrated score misrankings propagate downstream. Both Qwen settings pass the score gate. The Pythia cells contain no qualifying native-selection error, so the conditional output stage is inapplicable.

Exploratory gap-weighted spectral diagnostic Using 64 native-validation documents per setting, an outcome-blind extraction constructs teacher and selected-Rotary Field band-power measures weighted by $|H_{64}(\omega)|^2$ from Equation 13 for all 32 Qwen/OpenWebText and 12 Pythia/OpenWebText heads. Their exact one-dimensional Wasserstein distance in log-frequency, in octaves, is W_h. The Qwen association criterion requires Spearman $\rho \geq 0.35$ and one-sided permutation $p \leq 0.05$; qualifier separation requires AUC ≥ 0.80, exact $p \leq 0.05$ over all $\binom{32}{3} = 4{,}960$ assignments, and a positive seedwise harmful-minus-other difference. Pythia is descriptive and non-gating. Appendix F gives the spectral measure, exact detectable floor, and complete comparisons.

A.2 Operator response and native fitting

Frequency-wise gap response. The gap response can be examined frequency by frequency. Let $n_{\mathrm{rot}} = d_{\mathrm{rot}}/2$ and represent rotary plane ℓ by complex coordinates $\mathsf{q}_{i\ell} = q_{i\ell} + \mathrm{i}q_{i,\ell+n_{\mathrm{rot}}}$ and $\mathsf{k}_{j\ell} = k_{j\ell} + \mathrm{i}k_{j,\ell+n_{\mathrm{rot}}}$. Writing $c_{ij\ell} = \overline{\mathsf{q}}_{i\ell}\mathsf{k}_{j\ell}$ and $\delta_{ij} = p_j - p_i$, the plane contributes

$$\begin{aligned} S^T_{ij,\ell}(\delta_{ij}) &= \frac{1}{\sqrt{d_h}} \operatorname{Re}\left(c_{ij\ell} e^{\mathrm{i}\omega_\ell \delta_{ij}}\right), \\ S^T_{ij,\ell}(\delta_{ij} - g) - S^T_{ij,\ell}(\delta_{ij}) &= \frac{1}{\sqrt{d_h}} \operatorname{Re}\left(c_{ij\ell} e^{\mathrm{i}\omega_\ell \delta_{ij}} H_g(\omega_\ell)\right), \\ H_g(\omega) &= e^{-\mathrm{i}\omega g} - 1, \qquad |H_g(\omega)| = 2|\sin(\omega g/2)|. \end{aligned} \tag{13}$$

For an affected pair, the gap maps δ_{ij} to $\delta_{ij} - g$. The complex multiplier H_g changes each plane's contribution according to its frequency. Together with the phase of $c_{ij\ell}$, it determines how contributions reinforce or cancel. Unrotated coordinates cancel because they are position-independent, and a common translation leaves every relative offset unchanged. The frequency dependence motivates the original magnitude-based diagnostic (Appendix F). The retained phase motivates the signed-response design (Appendix I).

Native fitting objective. Section 3.4 defines the Cubic Field and Rotary Field score parameterizations. Both are fitted to the same native-position task.

We fit each method $m \in \{\mathrm{Cubic\,Field}, \mathrm{Rotary\,Field}\}$ only on native-position scores. Let $\mathcal{V}_{n,i}$ denote the valid causal keys for query i in training sequence n, $\mathcal{Q}_{\mathrm{tr}}$ the valid training query rows, $\Omega_{\mathrm{tr}} = \{(n,i,j) : (n,i) \in \mathcal{Q}_{\mathrm{tr}}, j \in \mathcal{V}_{n,i}\}$ the valid causal score entries, and $A^T_{n,i} = \mathrm{softmax}_{\mathcal{V}_{n,i}}(S^T_{n,i})$ and $\widehat{A}^m_{n,i} = \mathrm{softmax}_{\mathcal{V}_{n,i}}(\widehat{S}^m_{n,i})$ the teacher and surrogate attention rows. The training objective is

$$\begin{aligned} \mathcal{L}^m_{\mathrm{train}} = \frac{1}{|\mathcal{Q}_{\mathrm{tr}}|} \sum_{(n,i)\in\mathcal{Q}_{\mathrm{tr}}} D_{\mathrm{KL}}\Big(A^T_{n,i} \,\Big\|\, \widehat{A}^m_{n,i}\Big) \\ + \frac{\lambda_{\mathrm{score}}}{|\Omega_{\mathrm{tr}}|} \sum_{(n,i,j)\in\Omega_{\mathrm{tr}}} \Big(\mathcal{C}(S^T)_{n,i,j} - \mathcal{C}(\widehat{S}^m)_{n,i,j}\Big)^2, \qquad \lambda_{\mathrm{score}} = 0.02. \end{aligned} \tag{14}$$

The KL term matches the teacher attention distribution, and the centered-score term preserves pre-softmax structure without penalizing row-constant shifts. Checkpoint selection uses native-validation attention KL rather than the composite training loss. Selector diagnostics additionally compare training MSE, a strict two-metric Pareto rule, fixed method choices and the response oracle.

Attention and value aggregation. For a fixed value matrix V, different admissible attention rows can produce the same aggregated output when their difference lies in the left null space of V (Brunner et al., 2020). This distinction motivates separate comparisons in score space and output space.

Centered-logit response error. For OpenWebText, centered-logit response error provides a complementary output measure. Given logits z and vocabulary-centering operator C_v, it is

$$E_{m,h} = \frac{\sum_{s,d,t} \|C_v[(z^{m,s} - z^0) - (z^\star - z^0)]_{h,d,t}\|_2^2}{3\sum_{d,t} \|C_v(z^\star - z^0)_{h,d,t}\|_2^2}, \qquad E_m = \frac{1}{3}\sum_h E_{m,h}, \tag{15}$$

where $z^\star$ and z^0 are the teacher's intervened and native logits.

B QWEN/OPENWEBTEXT COUNTEREXAMPLE DETAILS

Table 5 gives the three Qwen/OpenWebText heads that satisfy all counterexample criteria. Each retains its native preference for Rotary Field while the intervention response favors Cubic Field and Null.

Table 5: The three prospective Qwen counterexamples. Values are three-seed means; lower is better. Native KL selects Rotary Field, while response risk and both simultaneous one-sided lower bounds establish that Rotary Field is worse than Cubic Field and Null. Null has risk 1.

Head	Native KL		Response risk		Simultaneous lower bound	
	Cubic Field	Rotary Field	Cubic Field	Rotary Field	$R_{\text{Rotary Field}} - 1$	$R_{\text{Rotary Field}} - R_{\text{Cubic Field}}$
L24H10	0.0820	0.0770	0.9823	3.2591	2.1064	2.1241
L25H12	0.0831	0.0651	0.9440	2.5172	1.3645	1.4205
L21H13	0.0944	0.0822	0.9572	2.0998	0.9471	0.9900

L18H02 is a fourth seed-stable descriptive reversal, with Cubic Field and Rotary Field risks 0.9382 and 1.1495. It does not qualify because the simultaneous lower bound for $R_{\text{Rotary Field}} - 1$ is -0.0032. Across the six heads fixed before this prospective run (listed in Appendix A), aggregate native harm is 0.8230 with lower bound 0.6703, and aggregate Rotary Field-versus-Cubic Field disadvantage is 0.7919 with lower bound 0.6392. Native validation selects Rotary Field on all 32 heads and remains seed-stable on held-out native data; only four heads favor Cubic Field for every seed, all within the six-head stratum. None of the other 26 heads has a seed-stable reversal.

The three Rotary Field-minus-Cubic Field risk differences are 2.277, 1.573 and 1.143, with simultaneous lower bounds 2.124, 1.421 and 0.990. The corresponding Rotary Field-minus-null effects are 2.259, 1.517 and 1.100, with lower bounds 2.106, 1.365 and 0.947.

C QWEN/OPENWEBTEXT OUTPUT DETAILS

Table 6: Output comparison on the three prospective Qwen counterexamples. KL values are in 10^{-6} nats/token; E is normalized centered-logit response error (Equation 15). Lower is better.

Head	K_0	K_A Cubic Field	K_N Rotary	$E_{\text{Cubic Field}}$	$E_{\text{Rotary Field}}$
L24H10	2.8038	2.5856	16.9197	0.9264	4.5169
L25H12	2.7804	2.1962	31.7200	0.8169	6.3117
L21H13	5.2735	4.5367	11.9511	0.8842	2.3089

Rotary Field has $K_N > K_0 > K_A$ pointwise on all three heads. The paired aggregate contrasts and simultaneous one-sided lower bounds are

$$K_N - K_A = 1.709 \times 10^{-5}, \qquad \text{LB}_{95\%} = 9.567 \times 10^{-6}, \tag{16}$$

$$K_N - K_0 = 1.658 \times 10^{-5}, \qquad \text{LB}_{95\%} = 9.054 \times 10^{-6}, \tag{17}$$

$$K_0 - K_A = 0.513 \times 10^{-6}, \qquad \text{LB}_{95\%} = -7.010 \times 10^{-6}. \tag{18}$$

The KL of the native-selected path is 6.50 times Cubic Field's and 5.58 times the no-response baseline, with normalized ranking gain 4.722. Cubic Field has the lower point estimate on each head, but the simultaneous $K_0 - K_A$ bound crosses zero. The result supports native-selection misranking without establishing that Cubic Field improves on Null.

D QWEN/WIKITEXT-103 REPLICATION DETAILS

The three OpenWebText counterexamples, L24H10, L25H12 and L21H13, replicate on WikiText-103; L18H02 additionally qualifies. All four are within the original six-head stratum. Across the frozen three-head stratum, native harm is 1.6411 with simultaneous lower bound 1.5118; the Rotary Field-versus-Cubic Field disadvantage is 1.6779 with lower bound 1.5485. All 192 fits meet the fitting criteria, and their unique checkpoints and native selections are fixed before score-response evaluation. The score analysis uses 10,000 paired document-bootstrap replicates with seed 20260819 and a 66-member simultaneous family; the output analysis uses the same replicate count and seed over its three contrasts. All 192 translation controls pass the frozen centered-score, attention, energy-ratio, and prediction limits. Each output record preserves predictions, teacher values, and non-target heads.

Table 7: Qwen/WikiText-103 output comparison on the four score-passing heads. KL values are in 10^{-6} nats/token; lower is better.

Head	K_0	K_A Cubic Field	K_N Rotary
L18H02	17.7955	13.3751	13.4882
L24H10	2.1773	2.3182	12.9674
L25H12	5.4340	4.4656	24.2967
L21H13	10.7785	9.3709	18.4540
Mean	9.0463	7.3824	17.3016

The simultaneous lower bounds are 6.143×10^{-6} for $K_N - K_0$, 7.806×10^{-6} for $K_N - K_A$, and -0.449×10^{-6} for $K_0 - K_A$. Rotary Field has 1.91 times the no-response KL and 2.34 times Cubic Field's KL, with normalized ranking gain 1.096. These values use the four-head aggregate in Table 7 and support $K_N > K_A$ and $K_N > K_0$, but not $K_A < K_0$. The three OpenWebText counterexamples retain the harmful output direction; L18H02 does not. Thus the output decision is ranking-only, and the score-level sign need not propagate on every qualifying head.

E PYTHIA SELECTED-CHECKPOINT AND PER-HEAD RESULTS

Table 8: Pythia/WikiText-103 three-seed mean native-validation KL and score-response risk. Rotary Field is lower in every comparison.

Head	Cubic Field native KL	Rotary Field native KL	Cubic Field risk	Rotary Field risk
L00H00	0.1408	0.0203	0.9988	0.1276
L00H03	0.2113	0.0546	1.2615	0.2301
L00H07	0.1549	0.0467	1.8603	0.6529
L00H11	0.3366	0.0799	2.3480	0.2216
L11H00	0.0695	0.0249	0.9951	0.4947
L11H03	0.1319	0.0475	1.5653	0.3774
L11H07	0.1363	0.0173	1.5268	0.2043
L11H11	0.1761	0.0658	1.5559	0.4000
L23H00	0.4235	0.2032	1.0890	0.1522
L23H03	0.3896	0.2084	1.0224	0.3078
L23H07	0.2816	0.2654	0.9938	0.7400
L23H11	0.2577	0.1095	0.9877	0.3161

The WikiText-103 values in Table 8 give mean risks of 1.3503863 for Cubic Field and 0.3520628 for Rotary Field. Always-Cubic Field has mean oracle regret 0.9983235. Fifteen of its 72 Cubic Field/Rotary Field fits trigger the non-gating endpoint-drift diagnostic: three seeds for L11H07/Cubic Field and all Cubic Field fits in layer 23. The maximum final-minus-best validation KL is 0.3074908 at L23H00/Cubic Field/seed 42. Evaluation uses the unique minimum-native-validation checkpoint, rather than the final step.

Table 9: Pythia/OpenWebText three-seed mean native-validation KL and score-response risk, with the supplied simultaneous 95% intervals for $R_{\text{Cubic Field}} - R_{\text{Rotary Field}}$. The paired document-bootstrap intervals retain the original Bonferroni adjustment across 12 heads. Rotary Field is lower in every comparison. All values are rounded to four decimal places.

Head	Native KL		Response risk		Simultaneous interval
	Cubic Field	Rotary Field	Cubic Field	Rotary Field	$R_{\text{Cubic Field}} - R_{\text{Rotary Field}}$
L00H00	0.2034	0.0304	1.0124	0.1090	[0.8960, 0.9099]
L00H03	0.1514	0.0522	1.2701	0.2578	[0.9889, 1.0359]
L00H07	0.1569	0.0574	1.8390	0.7560	[1.0350, 1.1351]
L00H11	0.3125	0.0890	2.2569	0.2344	[1.9561, 2.0985]
L11H00	0.0637	0.0272	1.1500	0.5352	[0.5830, 0.6455]
L11H03	0.1290	0.0453	1.4021	0.4385	[0.8935, 1.0382]
L11H07	0.1760	0.0222	1.2240	0.2034	[1.0026, 1.0386]
L11H11	0.1803	0.0611	1.3630	0.3764	[0.9190, 1.0620]
L23H00	0.3907	0.1873	1.0310	0.1419	[0.8767, 0.9000]
L23H03	0.3463	0.1940	1.0058	0.2650	[0.7208, 0.7581]
L23H07	0.1806	0.1123	0.9940	0.7275	[0.1752, 0.3376]
L23H11	0.2215	0.0951	0.9890	0.3365	[0.6292, 0.6725]

On OpenWebText, the corresponding mean risks are 1.2947810 and 0.3651276, and always-Cubic Field oracle regret is 0.9296534 (Table 9). Fourteen of 72 fits trigger the endpoint-drift diagnostic, with maximum final-minus-best KL 0.3384795 at L23H00/Cubic Field/seed 42. All 72 fits remain qualified in each corpus. The maximum common-translation residual MSE is 9.00×10^{-11}, attained by L23H00/Rotary Field. This statistic is the mean squared row-centered surrogate score change, averaged over valid causal pairs within each document, then over documents and the three fitting seeds, before taking the maximum over head–method pairs. It is an unnormalized translation-invariance residual in squared score units; the exact teacher response is zero.

In each corpus, all 36 seed-level comparisons favor Rotary Field for native validation, held-out native testing and score response. Native-validation KL, training MSE, strict Pareto, always-Rotary Field and the response oracle select Rotary Field on all 12 heads with zero regret; always-Cubic Field selects the response winner on none. Cubic Field and Rotary Field contain 13,025,285 and 13,026,871 trainable parameters, respectively. Appendix A accounts for all 312 records per corpus, including the 72 fits. The accompanying analyses retain the per-head simultaneous intervals for both corpora.

F GAP-WEIGHTED SPECTRAL DIAGNOSTIC

The original scalar diagnostic tests whether a mismatch in nonnegative, gap-weighted rotary band power explains the selected Rotary Field response errors. It uses the fixed selected Rotary Field checkpoints and native-validation documents.

For rotary plane ℓ, projected two-dimensional coordinates $q_{xi\ell}$ and $k_{xj\ell}$, angular frequency ω_ℓ, score scale a, and the cross-cutoff support $i \in [128, 255], j \in [0, 127]$, define

$$P_\ell = a^2 |H_{64}(\omega_\ell)|^2 \sum_x \left(\sum_i \|q_{xi\ell}\|_2^2 \right) \left(\sum_j \|k_{xj\ell}\|_2^2 \right). \tag{19}$$

Normalizing the positive P_ℓ values gives a measure over log-frequency,

$$\mu = \sum_\ell \frac{P_\ell}{\sum_s P_s} \delta_{\log_2 \omega_\ell}. \tag{20}$$

For each head, W_h is the exact one-dimensional 1-Wasserstein distance, in octaves, between the teacher measure and the equal mixture of the three seed-specific Rotary Field measures. The comparison covers all 32 Qwen/OpenWebText heads and the fixed 12-head Pythia/OpenWebText panel. Direct-score, identity, translation, selected-checkpoint, and Pythia unrotated-coordinate controls all pass.

Table 10: Prespecified Qwen tests and descriptive Pythia comparisons. One-sided permutation tests use 100,000 draws for correlations; the Qwen qualifier test enumerates all $\binom{32}{3} = 4{,}960$ label assignments. Pythia and cross-model rows are descriptive and non-gating.

Comparison	Statistic	Estimate	Effect floor	p	Decision
Qwen W_h vs. Rotary Field response risk	Spearman ρ	-0.107	≥ 0.35	0.724	Fail
Qwen harmful vs. other heads	Rank AUC	0.759	≥ 0.80	0.0819	Fail
Pythia W_h vs. Rotary Field response risk	Spearman ρ	0.378	—	0.114	Descriptive
Harmful Qwen vs. Pythia	Rank AUC	0.250	—	0.910	Descriptive

With three qualifying and 29 other heads, the rank comparison contains 87 positive–negative pairwise orderings and AUC advances in steps of $1/87$. An exact one-sided $p \leq 0.05$ is first attained at $U = 70$, corresponding to AUC 0.8046 and $p = 0.0478$; this is the design's discrete detectable-separation floor. The observed statistic is $U = 66$, four pairwise orderings below that floor.

The qualifying heads have mean $W_h = 0.889$ octaves, compared with 0.667 for the other 29 heads, and the harmful-minus-other difference is positive for each fitting seed. Their W_h values rank eighth, ninth, and tenth among the 32 heads. The continuous association has the opposite sign from the hypothesis.

G INTERVENTION-GRID ROBUSTNESS

The main study establishes response misranking at one cutoff and gap. The grid asks how native-selection behavior varies with intervention location and size. The cutoff determines which content pairs change relative offset; the gap determines their frequency-dependent phase shifts (Equation 13). It covers Qwen3-1.7B-Base on WikiText-103 and Pythia-1.4B on OpenWebText and WikiText-103. Each setting has 256 fitting documents, 64 validation documents, and a distinct 512-document response role. The fixed panels contain 32 Qwen heads and 12 Pythia heads; their composition is given in Appendix A.

Position Interpolation examines RoPE score behavior beyond the fitted position range (Chen et al., 2023). Our grid keeps sequence length and upstream states fixed.

For a 256-token sequence, the position map is

$$p_i^{(c,g)} = i + g\mathbf{1}[i \geq c], \qquad c \in \{64, 128, 192\}, \quad g \in \{8, 16, 32, 64, 128\}. \tag{21}$$

All 15 cells retain tokens, the target head's input state, values, and causal mask. The measured pairs have suffix queries and prefix keys, $i \geq c$ and $j < c$. Responses are row-centered over all valid keys before this pair mask is applied. The previously selected anchor $(c, g) = (128, 64)$ and the other 14 target cells have different selection histories. The analysis therefore distinguishes replication at the anchor from sensitivity beyond it; the fixed head panel supports neither a random model-wide sample nor a prevalence estimate.

The main comparison uses Cubic Field and Rotary Field. The expanded RoPE studies add Bilinear Field and Fixed-frequency Rotary Field to examine explicit position dependence and frequency support, respectively. Table 11 summarizes the four fitted surrogates.

Table 11: Methods in the expanded RoPE studies. Cubic Field and Rotary Field form the main-text comparison; Bilinear Field and Fixed-frequency Rotary Field are additional controls. All use the prescribed nearly matched parameter budgets; recorded selections and historical fixed snapshots are distinguished below.

Method	Score structure	Role in the comparison
Cubic Field	Content-dependent bilinear term plus query-conditioned smooth relative-position terms through cubic order	Smooth positional reconstruction prior
Bilinear Field	Position-blind content-dependent bilinear score	Control for explicit position dependence
Rotary Field	Content-dependent rotary bilinear score with learned frequencies	Learned-frequency reconstruction prior
Fixed-frequency Rotary Field	Rotary Field bilinear score with fixed teacher-supported frequencies and learned pair gains	Control for frequency support

Bilinear Field has zero pure-gap response when the input representations are fixed. Fixed-frequency Rotary Field fixes teacher-supported frequency locations, learns pair gains, and retains an explicit zero-frequency content path for Pythia's nonrotary component. Its other weights are fitted separately. Better fidelity would support this parameterization's usefulness; poor fidelity would show that fixed support alone is insufficient in the fitted model. This comparison does not causally isolate learned frequencies.

Fits with recorded selection use native-validation attention KL before response evaluation. Historical fixed Rotary Field snapshots retain unavailable selection histories (Appendix G.2). Three fitted members, initialized with seeds 42–44, contribute equally to each response field before its error is measured. This ensemble differs from the main study's averaging of seed-specific errors and output KL values.

For centered teacher response Δ_d and prediction X_d, risk per head and target is $\sum_d \|X_d - \Delta_d\|^2 / \sum_d \|\Delta_d\|^2$ over affected pairs. Null and the teacher response give 1 and 0 when teacher-

response energy is positive. Heads receive equal weight; corpus estimates remain separate. The 10,000-draw source-document bootstrap recomputes ratios of sums and constructs simultaneous 95% bounds within the specified contrast families, quantifying document uncertainty for each fixed head panel.

G.1 Fresh anchor, intervention grid and full-panel results

The fresh Qwen/WikiText anchor has equal-head mean risk contrasts of 0.3841 for Rotary Field minus Cubic Field and 0.4328 for Rotary Field minus Null. Their simultaneous one-sided 95% lower bounds are 0.3642 and 0.4121 in the two-contrast anchor family, exceeding its 0.10 threshold. All three historical failure heads are material and support both positive point contrasts. The 30-contrast grid family uses separate bounds and requires both lower bounds above zero with at least two supporting material heads. All 15 targets pass with three supporting heads; Table 2 reports that family's bounds, not the anchor-family bounds.

Table 13 shows all 480 head–target cells in the 32-head panel. Both point contrasts are positive at 51 cells, both are nonpositive at 419, and ten have mixed signs. All cells pass their materiality checks. These categories describe point estimates; the accompanying 960 two-sided simultaneous intervals form a separate family. Fixed-panel counts do not estimate model-wide prevalence.

The expanded Qwen/OpenWebText branch was not opened: its conditional power calibration lower bound, 0.8985, missed the prescribed 0.90 threshold. The corresponding WikiText bound was 0.9994. These Monte Carlo bounds are conditional on the historical empirical cohorts, not guarantees of population power. Expanded Qwen cross-corpus breadth is therefore unavailable. This disposition does not alter the completed historical OpenWebText/WikiText replication in Appendices B–D.

For Pythia, both risk-contrast upper bounds are negative at every target in both corpora, with all 12 heads material. Every target meets the prescribed adequate-nonreversal rule: at least nine material heads and both simultaneous upper bounds at most 0.10. No target meets the reversal rule, which requires both positive lower bounds and at least seven supporting material heads. The reversal rule also requires at least nine material heads. The two-sided family contains 60 contrasts across both corpora. These are bounds for the fixed panels, not a universal absence of misranking.

Table 12: Qwen/WikiText response-risk contrasts across all 15 targets. Estimates average L21H13, L24H10 and L25H12 equally. LB is the simultaneous one-sided 95% lower bound over 30 contrasts.

c	g	$\widehat{\Delta}_{R-C}$	LB_{R-C}	$\widehat{\Delta}_{R-0}$	LB_{R-0}	Support
64	8	1.4191	1.3760	1.4317	1.3884	3/3
64	16	0.9021	0.8698	0.9263	0.8938	3/3
64	32	0.6701	0.6392	0.7232	0.6918	3/3
64	64	0.4421	0.4148	0.5153	0.4868	3/3
64	128	0.5670	0.5337	0.6555	0.6202	3/3
128	8	1.2498	1.2099	1.2552	1.2153	3/3
128	16	0.7937	0.7630	0.8051	0.7744	3/3
128	32	0.5944	0.5633	0.6239	0.5925	3/3
128	64	0.3841	0.3546	0.4328	0.4022	3/3
128	128	0.5445	0.5057	0.6227	0.5814	3/3
192	8	1.1475	1.1077	1.1509	1.1110	3/3
192	16	0.7034	0.6718	0.7107	0.6790	3/3
192	32	0.4822	0.4483	0.5016	0.4675	3/3
192	64	0.2476	0.2138	0.2800	0.2451	3/3
192	128	0.4159	0.3706	0.4680	0.4197	3/3

Note. R, C, and 0 denote Rotary Field, Cubic Field, and Null. The fresh anchor-only family has lower bounds 0.3642 and 0.4121 at $(128, 64)$; the grid-family bounds differ because they cover all 15 targets.

Table 13: Point-sign patterns across all 32 Qwen/WikiText heads and 15 targets. Columns separate cutoff c and gap g. Symbols and shading encode the two risk contrasts, Rotary-Cubic and Rotary-Null; they do not encode statistical significance.

	$c = 64$					$c = 128$					$c = 192$				
Head / g	8	16	32	64	128	8	16	32	64	128	8	16	32	64	128
L00H01	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L03H01	≤ 0	≤ 0	≤ 0	≤ 0	$+$	≤ 0	≤ 0	≤ 0	$\pm$	$+$	≤ 0	≤ 0	≤ 0	$+$	$+$
L06H04	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L02H07	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$
L04H08	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L00H09	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L03H12	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$+$
L03H15	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L13H03	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$
L11H00	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L09H05	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L10H07	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L09H11	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L07H10	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L07H14	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L11H15	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L14H01	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L18H02	≤ 0	≤ 0	≤ 0	≤ 0	$\pm$	≤ 0	≤ 0	≤ 0	≤ 0	$+$	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L16H05	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L19H07	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L19H09	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L18H09	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L18H15	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L15H12	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L21H02	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L26H01	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L27H07	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L26H07	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L25H09	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
L24H10	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
L25H12	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
L21H13	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$

Note. $+$ (orange): both contrasts positive; ≤ 0 (gray): both nonpositive; $\pm$ (blue): one positive and one nonpositive. All 480 cells are material: 51 have both contrasts positive, 419 both nonpositive, and ten mixed signs. The supporting CSV retains both estimates and simultaneous two-sided 95% bounds (960-contrast family), plus response sufficient statistics and scalars.

Table 14: Pythia/OpenWebText: all 15 target contrasts for the fixed 12-head panel. LB and UB are simultaneous two-sided 95% bounds over the joint 60-contrast, two-corpus family.

c	g	$\widehat{\Delta}_{R-C}$	LB	UB	$\widehat{\Delta}_{R-0}$	LB	UB
64	8	-0.8653	-0.8697	-0.8608	-0.8571	-0.8615	-0.8527
64	16	-0.8900	-0.8945	-0.8855	-0.8655	-0.8699	-0.8612
64	32	-0.9242	-0.9315	-0.9170	-0.8063	-0.8123	-0.8002
64	64	-1.1020	-1.1143	-1.0898	-0.6202	-0.6284	-0.6119
64	128	-1.3761	-1.3966	-1.3556	-0.2462	-0.2612	-0.2311
128	8	-0.8673	-0.8724	-0.8621	-0.8606	-0.8657	-0.8554
128	16	-0.8902	-0.8954	-0.8850	-0.8716	-0.8768	-0.8665
128	32	-0.9097	-0.9173	-0.9020	-0.8256	-0.8327	-0.8184
128	64	-1.0221	-1.0327	-1.0114	-0.6628	-0.6707	-0.6548
128	128	-1.1677	-1.1850	-1.1505	-0.3102	-0.3233	-0.2970
192	8	-0.8652	-0.8715	-0.8589	-0.8594	-0.8656	-0.8531
192	16	-0.8890	-0.8955	-0.8825	-0.8693	-0.8756	-0.8629
192	32	-0.9181	-0.9283	-0.9078	-0.8124	-0.8220	-0.8028
192	64	-1.0714	-1.0855	-1.0574	-0.6067	-0.6172	-0.5962
192	128	-0.9028	-0.9212	-0.8844	-0.2306	-0.2457	-0.2156

Note. R, C, and 0 denote Rotary Field, Cubic Field, and Null. Every target has 12/12 material heads, no confirmed reversal, and adequate nonreversal under the 0.10 margin. Headwise point contrasts and response scalars are retained in the anonymous full-panel CSV; the displayed intervals concern equal-head means.

Table 15: Pythia/WikiText-103: all 15 target contrasts for the fixed 12-head panel. LB and UB are simultaneous two-sided 95% bounds over the joint 60-contrast, two-corpus family.

c	g	$\widehat{\Delta}_{R-C}$	LB	UB	$\widehat{\Delta}_{R-0}$	LB	UB
64	8	-0.9025	-0.9044	-0.9005	-0.8900	-0.8920	-0.8881
64	16	-0.9303	-0.9324	-0.9281	-0.8946	-0.8965	-0.8927
64	32	-1.0109	-1.0155	-1.0063	-0.8409	-0.8438	-0.8380
64	64	-1.3321	-1.3421	-1.3221	-0.6559	-0.6612	-0.6505
64	128	-1.6942	-1.7086	-1.6797	-0.2798	-0.2886	-0.2709
128	8	-0.9026	-0.9046	-0.9007	-0.8925	-0.8944	-0.8907
128	16	-0.9262	-0.9282	-0.9242	-0.8997	-0.9015	-0.8979
128	32	-0.9752	-0.9788	-0.9716	-0.8604	-0.8629	-0.8579
128	64	-1.1752	-1.1833	-1.1670	-0.6998	-0.7040	-0.6955
128	128	-1.3656	-1.3778	-1.3534	-0.3325	-0.3397	-0.3254
192	8	-0.8998	-0.9019	-0.8977	-0.8917	-0.8938	-0.8896
192	16	-0.9236	-0.9258	-0.9214	-0.8978	-0.8997	-0.8959
192	32	-0.9842	-0.9886	-0.9798	-0.8499	-0.8527	-0.8471
192	64	-1.2285	-1.2396	-1.2175	-0.6493	-0.6543	-0.6443
192	128	-1.0305	-1.0446	-1.0164	-0.2538	-0.2617	-0.2459

Note. R, C, and 0 denote Rotary Field, Cubic Field, and Null. Every target has 12/12 material heads, no confirmed reversal, and adequate nonreversal under the 0.10 margin. Headwise point contrasts and response scalars are retained in the anonymous full-panel CSV; the displayed intervals concern equal-head means.

G.2 Descriptive four-method response comparison

The separately assembled response comparison covers all 32 Qwen/WikiText heads, four methods and 15 targets, yielding 1,920 head–method–target entries, including 480 for Fixed-frequency Rotary Field. Table 16 shows the complete anchor panel; the accompanying data retain every target at full precision. These are descriptive response comparisons, without a new confidence family or

method-selection rule. The earlier three-head illustration retains its original scope and numerical projection.

Native-fit evidence requires a separate distinction. The 384 checkpoint receipts retain recorded selection histories where available and identify 96 historical Rotary Field snapshots whose selection histories are unavailable. A snapshot step does not establish a selected step. The 192 Cubic/Rotary common-panel native remeasurements are separate from selected-validation KL; corresponding Bilinear/Fixed-frequency remeasurements are absent. Consequently, the response table does not establish a common-cohort four-method native-fit ranking.

Fixed-frequency Rotary Field changes frequency support within a separately fitted parameterization, with learned gains, projections and checkpoint differences. Its response comparisons test that parameterization without uniquely attributing an improvement or remaining error to learned frequencies. Teacher energy and pair counts are shared across methods and must not be pooled over repeated method rows. Native-score energy uses the full causal mask, whereas response RMS and error RMSE use affected pairs.

Table 16: Descriptive response risk at $(128, 64)$ for all 32 Qwen/WikiText heads. The four separately fitted methods use the same 512-document response cohort. Null risk is 1; smaller values indicate lower response error. The complete 15-target values are supplied in the supporting data.

Head	Cubic Field	Bilinear Field	Rotary Field	Fixed-frequency Rotary Field
L00H01	1.348	1.000	0.245	0.222
L03H01	1.062	1.000	1.005	0.965
L06H04	1.506	1.000	0.240	0.253
L02H07	1.473	1.000	0.641	0.601
L04H08	1.489	1.000	0.268	0.294
L00H09	1.245	1.000	0.177	0.113
L03H12	1.325	1.000	0.900	0.810
L03H15	2.106	1.000	0.444	0.456
L13H03	1.557	1.000	0.683	0.598
L11H00	1.800	1.000	0.331	0.320
L09H05	2.014	1.000	0.413	0.437
L10H07	1.044	1.000	0.406	0.385
L09H11	1.312	1.000	0.104	0.105
L07H10	1.515	1.000	0.233	0.272
L07H14	1.808	1.000	0.395	0.400
L11H15	1.292	1.000	0.245	0.224
L14H01	1.767	1.000	0.405	0.393
L18H02	0.991	1.000	0.747	0.603
L16H05	0.921	1.000	0.389	0.357
L19H07	1.410	1.000	0.139	0.119
L19H09	0.987	1.000	0.449	0.369
L18H09	1.132	1.000	0.060	0.055
L18H15	1.262	1.000	0.148	0.121
L15H12	1.310	1.000	0.077	0.070
L21H02	0.972	1.000	0.145	0.142
L26H01	0.993	1.000	0.255	0.244
L27H07	0.962	1.000	0.222	0.213
L26H07	0.988	1.000	0.259	0.224
L25H09	0.988	1.000	0.379	0.327
L24H10	1.126	1.000	1.632	1.412
L25H12	1.074	1.000	1.438	1.324
L21H13	0.946	1.000	1.228	1.041

H FILLER AND MATERIALITY CONTROLS

The pure gap holds the head's input representations fixed. Visible insertion also changes the representations produced by preceding layers, so its response can combine a direct position effect with upstream changes. The filler design separates these effects using 32 prespecified response sources per setting, each assigned four fixed donors from a disjoint 128-document donor role. The Qwen heads are L21H13, L24H10, and L25H12; the Pythia heads are L00H00, L11H07, and L23H11. The schedule records four entries: masked, real, collapsed, and paired position effect. The last is the real–collapsed contrast, with separate local and end-to-end estimands, rather than a fourth token construction. Three heads, four methods, four entries and 15 targets give 720 records per setting, plus 15 separate equivalence gates. These counts enumerate comparisons, not independent observations.

Insert the first g donor tokens before x_c. In the real condition, original suffix positions increase by g; in the collapsed condition, original tokens retain position i. Donor positions are $c, \ldots, c+g-1$ in both conditions, including intentional duplicate position identifiers in the collapsed map. Token order, content, and the causal mask are identical. A masked-donor condition removes donor attention participation and compares the aligned original-token computation with the pure gap. The initial equivalence panel comprises the first 16 sources; its score and attention checks apply to every donor and scheduled method. Real and collapsed conditions separately measure native fidelity. Their paired contrast defines the position-effect estimands.

The head-local estimand freezes the upstream state from the collapsed condition and changes the target head's position map. It therefore uses the collapsed-insertion state, whereas the pure gap uses the native state. The end-to-end estimand subtracts two complete forwards and includes upstream changes. Bilinear Field supplies the head-local null control but need not be null end to end. Both contrasts are centered over all valid expanded-sequence keys and scored on aligned original suffix-query/prefix-key pairs. Full expanded-sequence fidelity has a separate denominator. Teacher filler forwards and accumulated statistics use float64. Surrogates use float64 for Qwen and float32 for Pythia before float64 ensemble aggregation. The primary source statistic averages the four donor sufficient statistics; averaging fields before scoring is a separate descriptive summary. Resampling keeps the four donors nested within their source.

Materiality is evaluated separately for each estimand. Let T be summed teacher-effect energy on the target mask, V summed centered native-score energy over all valid expanded pairs, and n the target-pair count. The gates use $A = \sqrt{T/V}$ and $B = \sqrt{T/n} - 100\sqrt{T_{\mathrm{noop}}/n}$, where the no-op translates every collapsed position by 256 and is measured with the same mask. Simultaneous lower bounds must satisfy $\mathrm{LB}(A - 0.01) \geq 0$ and $\mathrm{LB}(B) > 0$. These controls separate a meaningful teacher effect from a large normalized error on a negligible response. Head-local and end-to-end contrasts retain separate materiality decisions.

H.1 MATERIAL TEACHER EFFECTS DO NOT IMPLY END-TO-END SELECTION FAILURE

All 15 masked-insertion equivalence gates pass in each setting. All 45 head–target materiality decisions pass separately for each of the two estimands in each setting, giving 270 passing decisions. Thus the tested teacher effects exceed both materiality thresholds, independently of which surrogate predicts them better.

On Qwen/WikiText, the head-local criterion is supported at ten targets: all five gaps for $c = 64$, gaps 8, 16 and 32 for $c = 128$, and gaps 8 and 16 for $c = 192$. Both simultaneous risk-contrast lower bounds are positive there, with the required material-head support. No end-to-end target meets the combined criterion. Rotary-minus-Cubic has a positive end-to-end lower bound at five targets, but Rotary-minus-Null never does. The two comparisons therefore support different conclusions.

At the filler anchor, the local Rotary-minus-Cubic contrast is -0.0267 with lower bound -0.0991, and Rotary-minus-Null is -0.0115 with lower bound -0.0850. The native-state pure gap and collapsed-state filler contrast are not matched responses: their upstream states and their 512-document versus 32-source cohorts differ. The local filler result does not invalidate the pure-gap counterexample or establish that visible insertion reproduces it.

All Pythia contrast point estimates favor Rotary Field in both estimands and corpora. The stored intervals are one-sided lower bounds for detecting harm; negative lower bounds alone do not certify

superiority. Tables below retain every target, including unsupported contrasts. Head risks are formed from source-summed sufficient statistics and then averaged equally across the three fixed heads. Four donors remain nested within each source.

Table 17: Qwen/WikiText filler effects on the fixed three-head panel. $D_{\mathrm{C}} = R_{\mathrm{Rotary}} - R_{\mathrm{Cubic}}$ and $D_0 = R_{\mathrm{Rotary}} - 1$ average per-head risks equally. LB denotes the accepted simultaneous one-sided 95% lower bound (60 contrasts; 10,000 source bootstrap draws). Positive lower bounds support greater Rotary Field error; negative lower bounds alone do not establish superiority. Mat. counts heads passing both materiality gates.

c	g	D_{C}	LB	D_0	LB	Mat.
Head-local						
64	8	1.425	1.220	1.437	1.231	3/3
64	16	0.870	0.748	0.888	0.766	3/3
64	32	0.608	0.490	0.633	0.514	3/3
64	64	0.331	0.234	0.349	0.249	3/3
64	128	0.308	0.200	0.360	0.248	3/3
128	8	1.227	1.047	1.233	1.053	3/3
128	16	0.703	0.591	0.714	0.603	3/3
128	32	0.369	0.267	0.391	0.290	3/3
128	64	-0.027	-0.099	-0.011	-0.085	3/3
128	128	0.063	-0.027	0.063	-0.026	3/3
192	8	1.051	0.879	1.055	0.882	3/3
192	16	0.519	0.407	0.527	0.414	3/3
192	32	0.084	-0.017	0.097	-0.006	3/3
192	64	-0.303	-0.373	-0.299	-0.370	3/3
192	128	-0.156	-0.244	-0.166	-0.256	3/3
End-to-end						
64	8	0.187	0.141	-0.284	-0.345	3/3
64	16	0.080	0.051	-0.406	-0.453	3/3
64	32	-0.014	-0.039	-0.551	-0.590	3/3
64	64	-0.033	-0.055	-0.629	-0.661	3/3
64	128	-0.041	-0.063	-0.637	-0.667	3/3
128	8	0.132	0.099	-0.374	-0.433	3/3
128	16	0.032	0.007	-0.485	-0.537	3/3
128	32	-0.033	-0.054	-0.587	-0.632	3/3
128	64	-0.045	-0.068	-0.641	-0.677	3/3
128	128	-0.042	-0.062	-0.652	-0.679	3/3
192	8	0.036	0.012	-0.525	-0.576	3/3
192	16	-0.005	-0.029	-0.560	-0.613	3/3
192	32	-0.048	-0.070	-0.630	-0.677	3/3
192	64	-0.052	-0.073	-0.660	-0.700	3/3
192	128	-0.046	-0.066	-0.649	-0.680	3/3

Table 18: Pythia/OpenWebText filler effects on the fixed three-head panel. $D_{\mathrm{C}} = R_{\mathrm{Rotary}} - R_{\mathrm{Cubic}}$ and $D_0 = R_{\mathrm{Rotary}} - 1$ average per-head risks equally. LB denotes the accepted simultaneous one-sided 95% lower bound (120 contrasts; 10,000 source bootstrap draws). Positive lower bounds support greater Rotary Field error; negative lower bounds alone do not establish superiority. Mat. counts heads passing both materiality gates.

c	g	D_{C}	LB	D_0	LB	Mat.
Head-local						
64	8	-0.931	-0.941	-0.929	-0.938	3/3
64	16	-0.943	-0.952	-0.932	-0.941	3/3
64	32	-0.937	-0.955	-0.892	-0.910	3/3
64	64	-0.928	-0.967	-0.764	-0.801	3/3
64	128	-0.838	-0.921	-0.531	-0.617	3/3
128	8	-0.936	-0.943	-0.933	-0.940	3/3
128	16	-0.949	-0.956	-0.939	-0.946	3/3
128	32	-0.944	-0.953	-0.914	-0.927	3/3
128	64	-0.881	-0.908	-0.806	-0.840	3/3
128	128	-0.813	-0.879	-0.561	-0.644	3/3
192	8	-0.933	-0.940	-0.933	-0.940	3/3
192	16	-0.940	-0.946	-0.939	-0.945	3/3
192	32	-0.924	-0.936	-0.910	-0.923	3/3
192	64	-0.856	-0.897	-0.778	-0.822	3/3
192	128	-0.680	-0.760	-0.584	-0.678	3/3
End-to-end						
64	8	-0.910	-0.925	-0.914	-0.927	3/3
64	16	-0.902	-0.921	-0.910	-0.923	3/3
64	32	-0.843	-0.866	-0.857	-0.877	3/3
64	64	-0.774	-0.797	-0.736	-0.759	3/3
64	128	-0.788	-0.828	-0.564	-0.595	3/3
128	8	-0.909	-0.922	-0.912	-0.925	3/3
128	16	-0.908	-0.919	-0.913	-0.924	3/3
128	32	-0.868	-0.881	-0.874	-0.888	3/3
128	64	-0.793	-0.816	-0.768	-0.791	3/3
128	128	-0.791	-0.821	-0.589	-0.619	3/3
192	8	-0.906	-0.917	-0.907	-0.918	3/3
192	16	-0.897	-0.908	-0.900	-0.911	3/3
192	32	-0.855	-0.868	-0.856	-0.871	3/3
192	64	-0.790	-0.816	-0.760	-0.778	3/3
192	128	-0.692	-0.730	-0.618	-0.653	3/3

Table 19: Pythia/WikiText filler effects on the fixed three-head panel. $D_{\mathrm{C}} = R_{\mathrm{Rotary}} - R_{\mathrm{Cubic}}$ and $D_0 = R_{\mathrm{Rotary}} - 1$ average per-head risks equally. LB denotes the accepted simultaneous one-sided 95% lower bound (120 contrasts; 10,000 source bootstrap draws). Positive lower bounds support greater Rotary Field error; negative lower bounds alone do not establish superiority. Mat. counts heads passing both materiality gates.

c	g	D_{C}	LB	D_0	LB	Mat.
Head-local						
64	8	-0.948	-0.953	-0.941	-0.946	3/3
64	16	-0.964	-0.969	-0.943	-0.947	3/3
64	32	-0.986	-0.994	-0.913	-0.920	3/3
64	64	-1.057	-1.082	-0.783	-0.797	3/3
64	128	-1.096	-1.145	-0.517	-0.550	3/3
128	8	-0.951	-0.956	-0.943	-0.947	3/3
128	16	-0.967	-0.971	-0.947	-0.951	3/3
128	32	-0.978	-0.983	-0.929	-0.935	3/3
128	64	-0.945	-0.962	-0.824	-0.833	3/3
128	128	-0.978	-1.017	-0.541	-0.567	3/3
192	8	-0.943	-0.947	-0.943	-0.947	3/3
192	16	-0.948	-0.953	-0.947	-0.950	3/3
192	32	-0.942	-0.954	-0.923	-0.928	3/3
192	64	-0.898	-0.926	-0.798	-0.809	3/3
192	128	-0.769	-0.793	-0.575	-0.603	3/3
End-to-end						
64	8	-0.928	-0.936	-0.928	-0.935	3/3
64	16	-0.926	-0.935	-0.924	-0.931	3/3
64	32	-0.887	-0.903	-0.878	-0.889	3/3
64	64	-0.844	-0.865	-0.745	-0.758	3/3
64	128	-0.943	-0.977	-0.504	-0.528	3/3
128	8	-0.931	-0.938	-0.928	-0.934	3/3
128	16	-0.931	-0.939	-0.927	-0.933	3/3
128	32	-0.896	-0.909	-0.892	-0.902	3/3
128	64	-0.825	-0.839	-0.785	-0.796	3/3
128	128	-0.888	-0.916	-0.521	-0.543	3/3
192	8	-0.928	-0.936	-0.922	-0.930	3/3
192	16	-0.927	-0.933	-0.920	-0.924	3/3
192	32	-0.880	-0.891	-0.875	-0.883	3/3
192	64	-0.810	-0.827	-0.763	-0.773	3/3
192	128	-0.742	-0.761	-0.568	-0.588	3/3

I SIGNED SPECTRAL AND SPATIAL DIAGNOSTICS

The scalar spectral test in Appendix F does not explain the Qwen failures. The signed diagnostic asks which response components align with the full teacher response, which oppose it, and how strongly they cancel. Spatial summaries locate the corresponding errors within the affected query–key block. It covers Cubic Field, Rotary Field, and Fixed-frequency Rotary Field on six Qwen/WikiText heads and three heads in each Pythia setting across the 15 targets: respectively 270, 135, and 135 cases. The six Qwen heads combine the three filler heads with L25H09, L26H01, and L27H07; Pythia uses the filler panel. Teacher frequencies define eight contiguous bands ordered by absolute frequency, with adjacent pairs merged into four bands. Geometric-midpoint cutpoints assign learned Rotary Field frequencies before ensemble averaging; edge bands retain frequencies outside the teacher range.

Rotary Field and Fixed-frequency Rotary Field contribute their signed rotary response components. Cubic Field uses a document-wise float64 minimum-norm projection onto the teacher-band response basis, retaining an explicit residual. For band component $X_{b,d}$, signed alignment with the complete teacher response is

$$q_b = \frac{\sum_d \langle X_{b,d}, \Delta_d \rangle}{\sum_d \|\Delta_d\|^2}. \tag{22}$$

Constructive and destructive contributions, cancellation $\sum_b |q_b| - |\sum_b q_b|$, bandwise error, and residual energy distinguish amplitude from phase alignment. The decomposition is checked against the original response in float64. Ratios with zero response energy remain undefined. The projection describes the fitted response in a teacher-band basis; it does not identify rotary components inside Cubic Field.

The spatial design contains eight Qwen/WikiText comparisons: all six heads at $(128, 64)$ and L21H13 additionally at gaps 8 and 128. Teacher, Cubic Field, and Rotary Field fields are summarized across the 512 response documents by signed means and RMS values. Error panels separate same-sign amplification excess from wrong-sign energy and show the affected block within the valid causal geometry. Comparable signed fields share a symmetric color scale; errors and RMS fields use separate scales. One illustrative document is selected by proximity to median teacher-response energy, with source order breaking ties. Scalar norm, direction, projection, and residual summaries quantify the patterns shown in the images. Downstream consequences are assessed separately in Appendix J.

I.1 RESPONSE MAGNITUDE, AMPLITUDE AND DIRECTIONAL ALIGNMENT

Let $T = \sum_d \|\Delta_d\|_{\mathcal{A}}^2$, $P = \sum_d \|X_d\|_{\mathcal{A}}^2$, $D = \sum_d \langle X_d, \Delta_d \rangle_{\mathcal{A}}$ and let n count all affected valid pairs across the 512 documents. Teacher-response RMS is $\sqrt{T/n}$ and prediction RMSE is $\sqrt{\sum_d \|X_d - \Delta_d\|_{\mathcal{A}}^2 / n}$. They are in centered-score units. Relative prediction amplitude is $\sqrt{P/T}$, alignment is $D/\sqrt{PT}$, and the projection coefficient is D/T. Direction is undefined when prediction energy is zero. These measures use the affected-pair mask and do not measure response magnitude relative to native-score energy on a different mask.

At the Qwen/WikiText anchor, teacher RMS is 0.2597, 0.2207 and 0.2291 on L21H13, L24H10 and L25H12, respectively. Rotary Field norm ratios are 1.149, 1.248 and 1.193, with cosines 0.475, 0.371 and 0.413. Cubic Field's smaller ratios, 0.280, 0.158 and 0.082, coexist with cosines 0.236, -0.318 and -0.411. The last two Cubic Field risks are 1.126 and 1.074: these new-cohort values do not reproduce every historical comparison with Null. Fig. 3 therefore identifies the heads by their historical selection, rather than claiming a fresh individual qualification for every method comparison.

Fixed-frequency Rotary Field RMSE at this anchor is 0.2649, 0.2623 and 0.2636, compared with learned-frequency Rotary Field's 0.2878, 0.2820 and 0.2748. These descriptive differences concern the separately fitted parameterizations and do not isolate the causal effect of frequency support. The supporting data retain all 15 targets: magnitude/error for three designated heads in each setting, Qwen Cubic/Rotary amplitude and alignment for all 32 heads, and the three-head fixed-frequency comparison. No additional confidence family is inferred from these summaries. The separate full-panel response comparison is reported in Appendix G.2; it does not change this three-head illustration.

I.2 SIGNED BANDS AND SPATIAL STRUCTURE

The recorded signed decompositions reconstruct all 540 response cases within the prescribed tolerances. This is an algebraic check on the decomposition, not evidence of mechanistic identification. Cubic Field's teacher-band projection retains a residual; its band labels do not identify internal rotary components.

Four-band cancellation is zero in all 90 Qwen Rotary Field head–target cases. Eight-band cancellation is positive in 22 cases, so the result depends on partition resolution. At the anchor, the three historical failure heads have no four-band cancellation for either rotary parameterization. These diagnostics do not establish destructive between-band cancellation as the explanation for their errors. The original scalar frequency-distance test remains negative (Appendix F).

At the six-head spatial anchor, Rotary Field direction cosines are 0.371–0.475 on the historical failure heads and 0.789–0.883 on the three success controls. This selected-head contrast describes directional fidelity; it does not identify a causal frequency mechanism. Simultaneous geometry intervals are available only for L21H13 at the three scheduled gaps and must not be extended to other heads or spectral bands.

Spatial risk, norm and cosine use affected-pair teacher energy. The saved amplification-excess, wrong-sign and total-error ratios instead use full valid-field teacher energy. Region fractions describe predicted response energy, not error energy. Wrong-sign energy counts squared prediction energy where signs differ; together with amplification excess it is not an exact partition of response risk. Supporting data preserve these definitions and the full spatial arrays.

Table 20: Complete signed-spectrum anchor summary at $(c, g) = (128, 64)$: all twelve prescribed setting–head combinations and all three methods. C_4, C_8 are the accepted four-/eight-band cancellation $\sum_b |q_b| - |\sum_b q_b|$, with $q_b = \langle X_b, \Delta\rangle / \|\Delta\|^2$ using full affected teacher-response energy. Entries are descriptive and partition-dependent; zero four-band cancellation does not imply zero eight-band cancellation or accurate response reconstruction. Cubic Field components are a teacher-band projection with a retained residual. Full-precision cancellation, band-specific alignment and geometry, teacher energies, and all 540 cases are supplied with the source data.

	Cubic Field		Rotary Field		Fixed-frequency	
Head	C_4	C_8	C_4	C_8	C_4	C_8
Qwen/WikiText						
L21H13	0.0045	0.00516	0	2.61e-06	0	4.13e-06
L24H10	0.00539	0.00577	0	0	0	1.99e-05
L25H12	0	0	0	0	0	4.44e-06
L25H09	0.00475	0.00475	0	0	0	2.93e-06
L26H01	0.00119	0.00145	0	0	0	2.19e-06
L27H07	0.00636	0.0168	0	7.66e-06	2.46e-06	2.46e-06
Pythia/OpenWebText						
L00H00	0.00465	0.015	0.000233	0.000233	0	0
L11H07	0.0267	0.0349	0.0176	0.0206	0.0285	0.0291
L23H11	0.00518	0.00822	0	0	0	0
Pythia/WikiText						
L00H00	0.00208	0.00846	0.00118	0.00118	0	0
L11H07	0.0696	0.0696	0.0132	0.0161	0.0296	0.0302
L23H11	0.0185	0.021	0	0	0	0

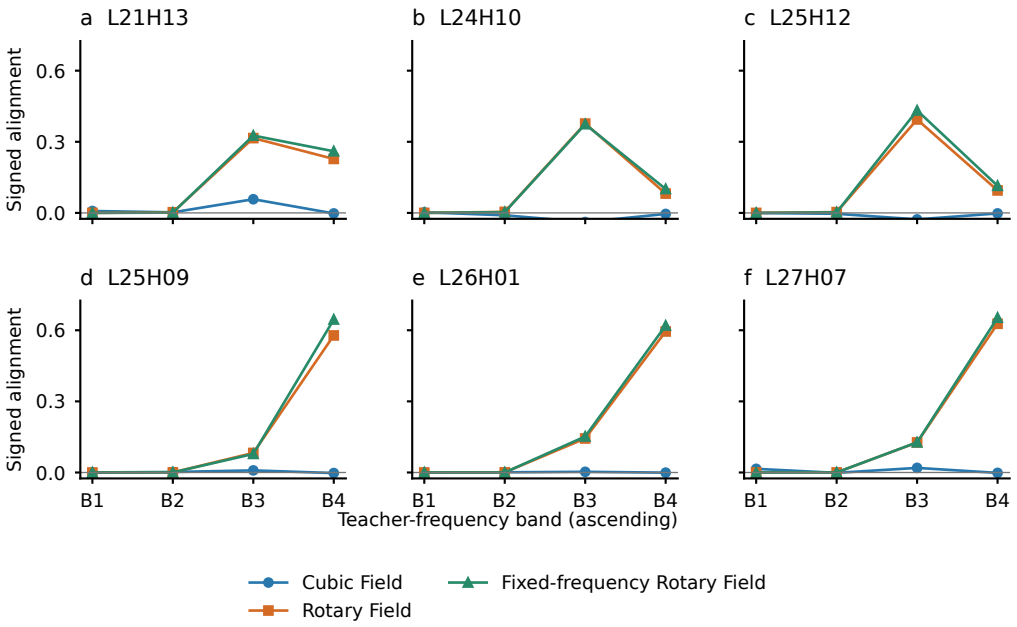

Figure 5: Signed alignment q_b at $(128, 64)$ on all six prescribed Qwen/WikiText spectral heads. Bands increase in absolute teacher frequency; all panels share the same vertical scale. Values use the full affected teacher-response energy as denominator and are descriptive. Cubic Field denotes its teacher-band projection, not internal rotary components.

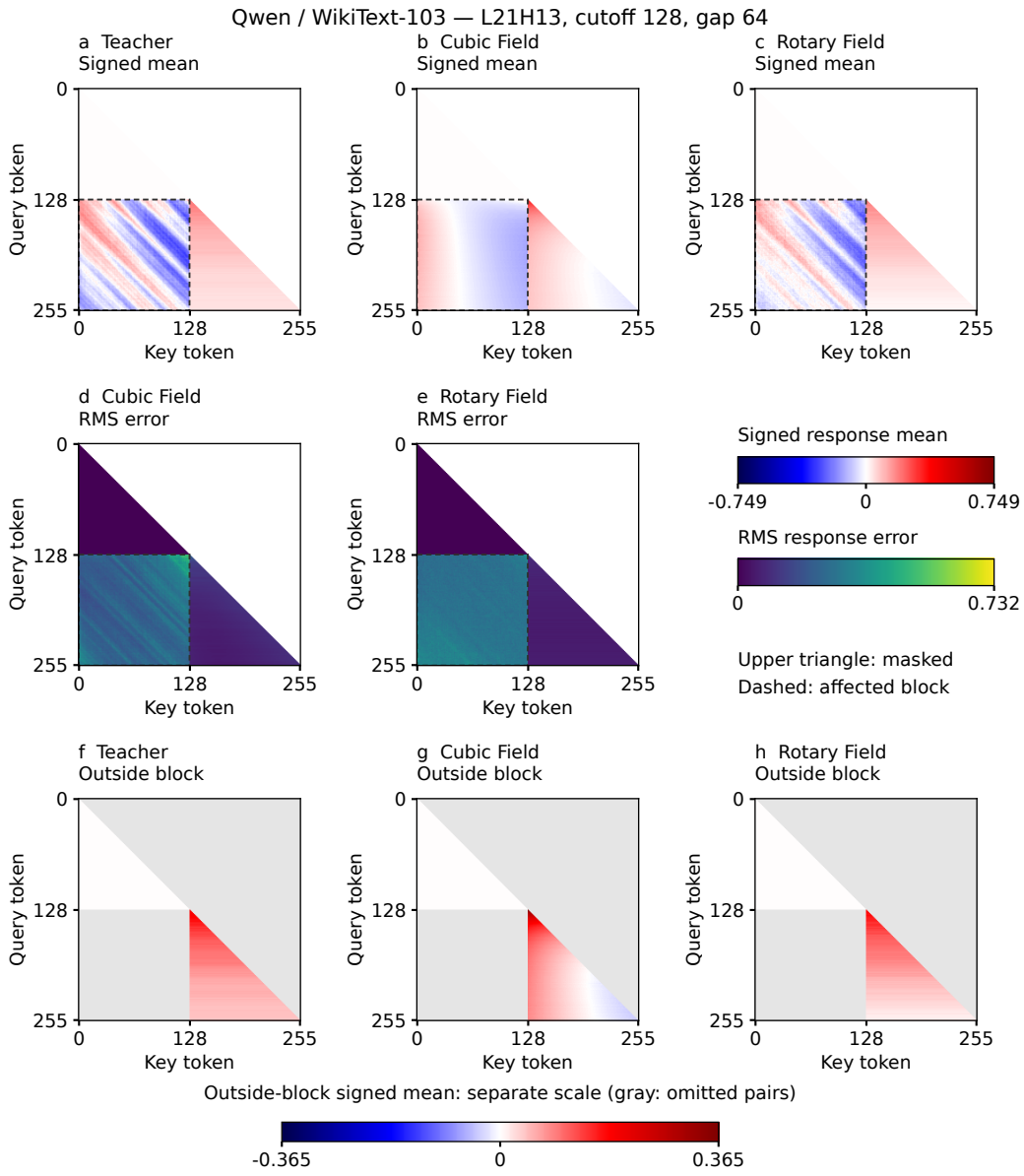

Figure 6: **Spatial response summaries: L21H13,** $(c, g) = (128, 64)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across L21H13 gaps 8, 64, and 128; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

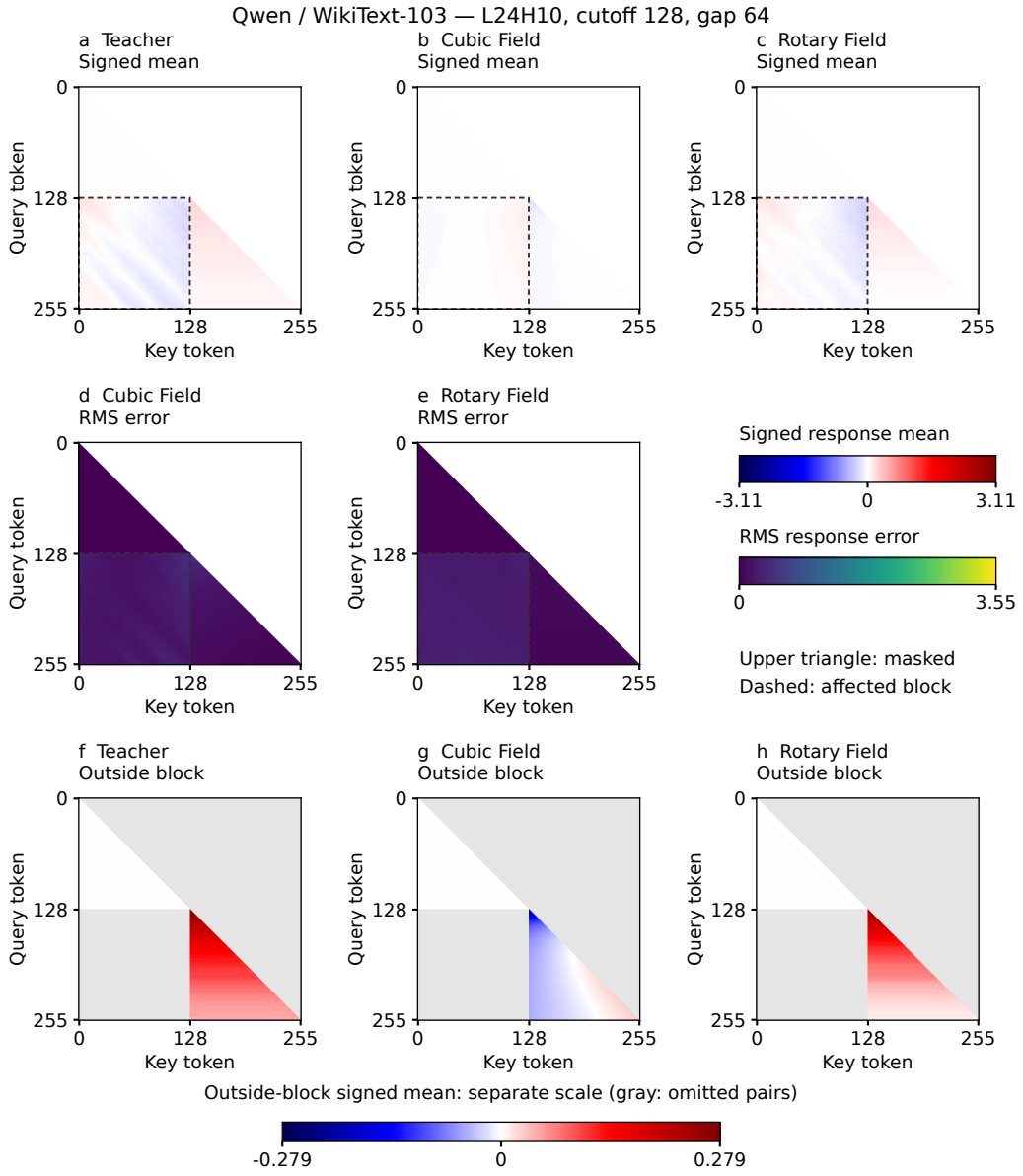

Figure 7: **Spatial response summaries: L24H10,** $(c, g) = (128, 64)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across the five non-L21H13 anchor heads; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

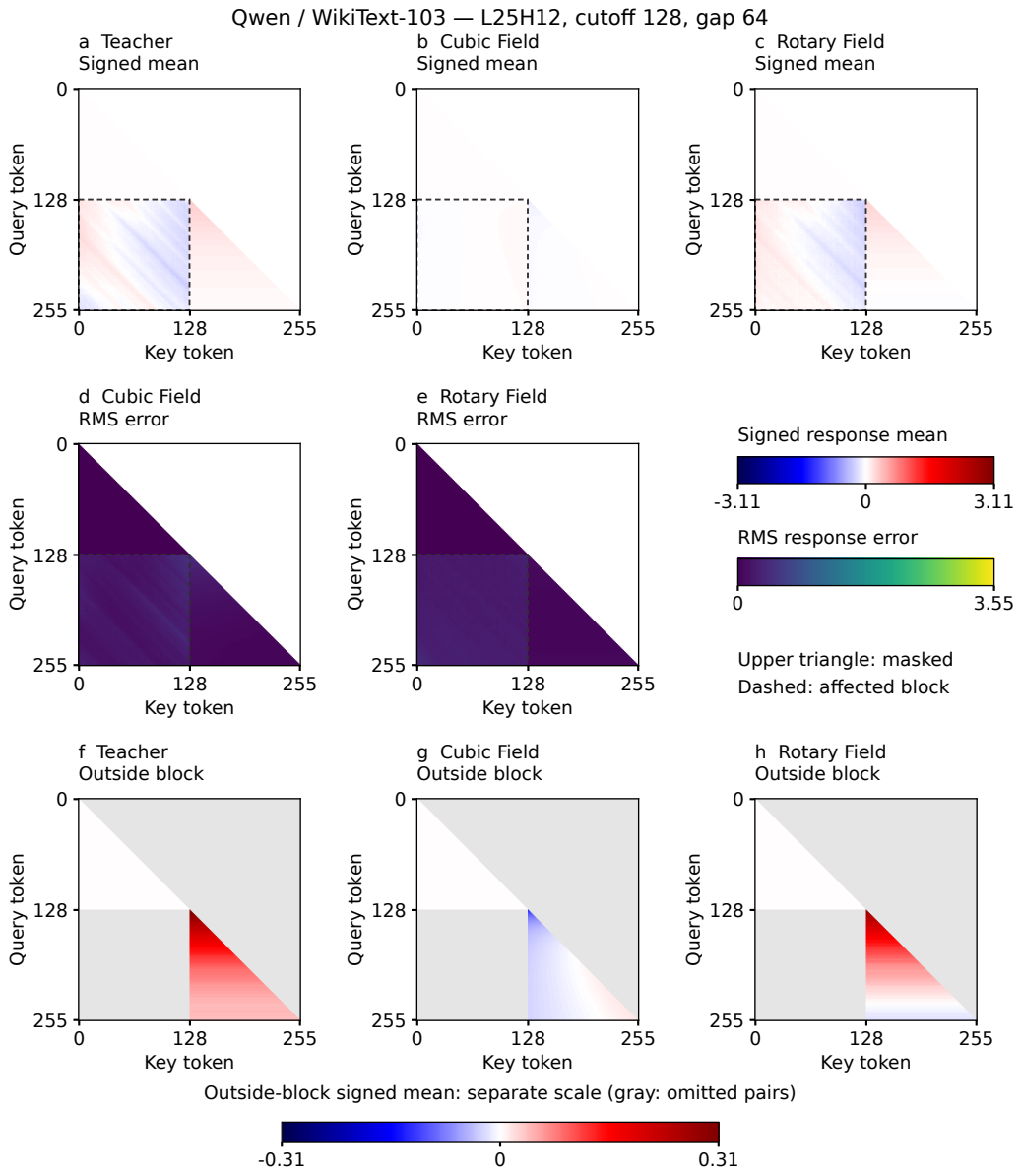

Figure 8: **Spatial response summaries: L25H12,** $(c, g) = (128, 64)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across the five non-L21H13 anchor heads; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

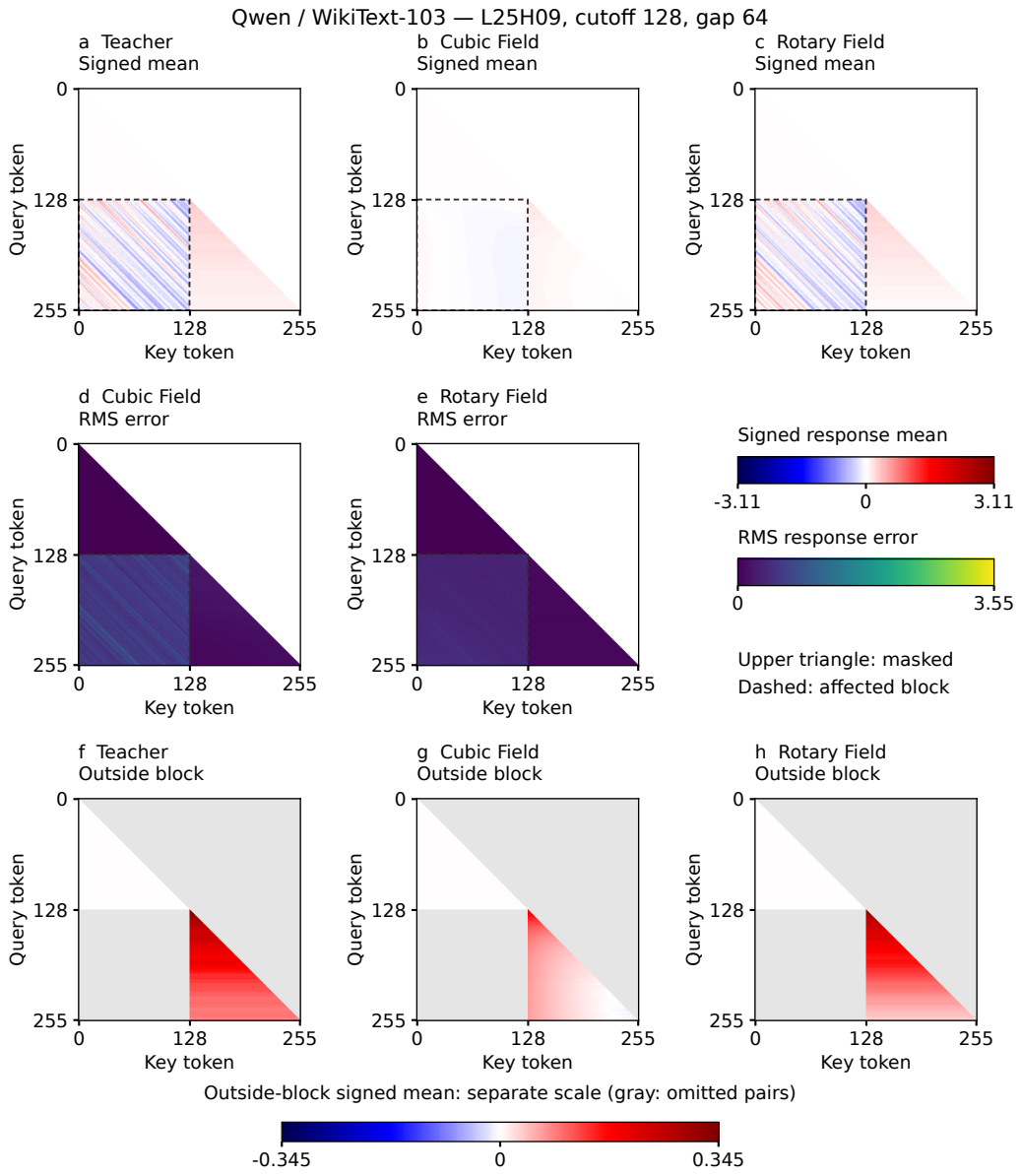

Figure 9: **Spatial response summaries: L25H09,** $(c, g) = (128, 64)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across the five non-L21H13 anchor heads; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

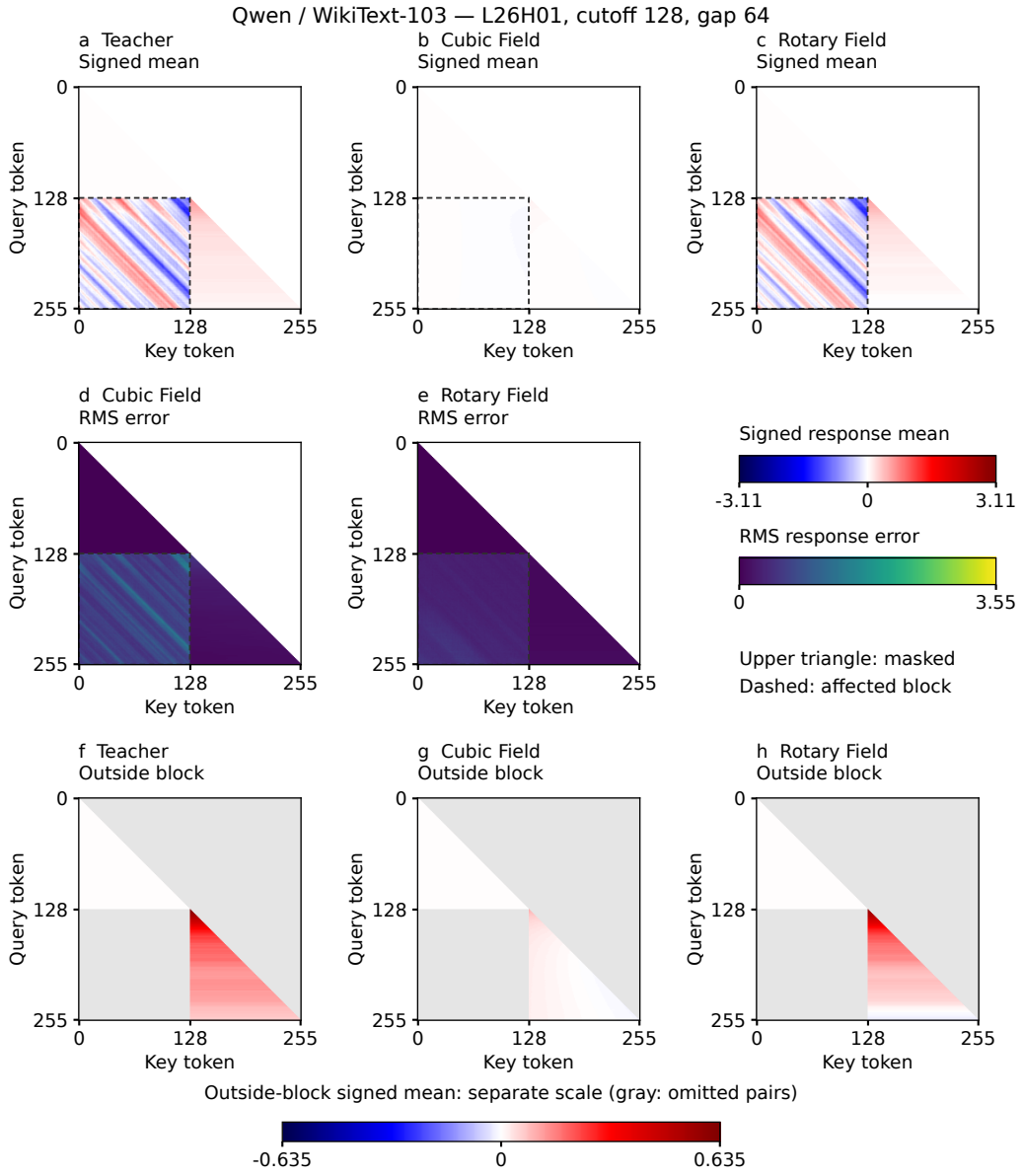

Figure 10: **Spatial response summaries: L26H01,** $(c, g) = (128, 64)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across the five non-L21H13 anchor heads; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

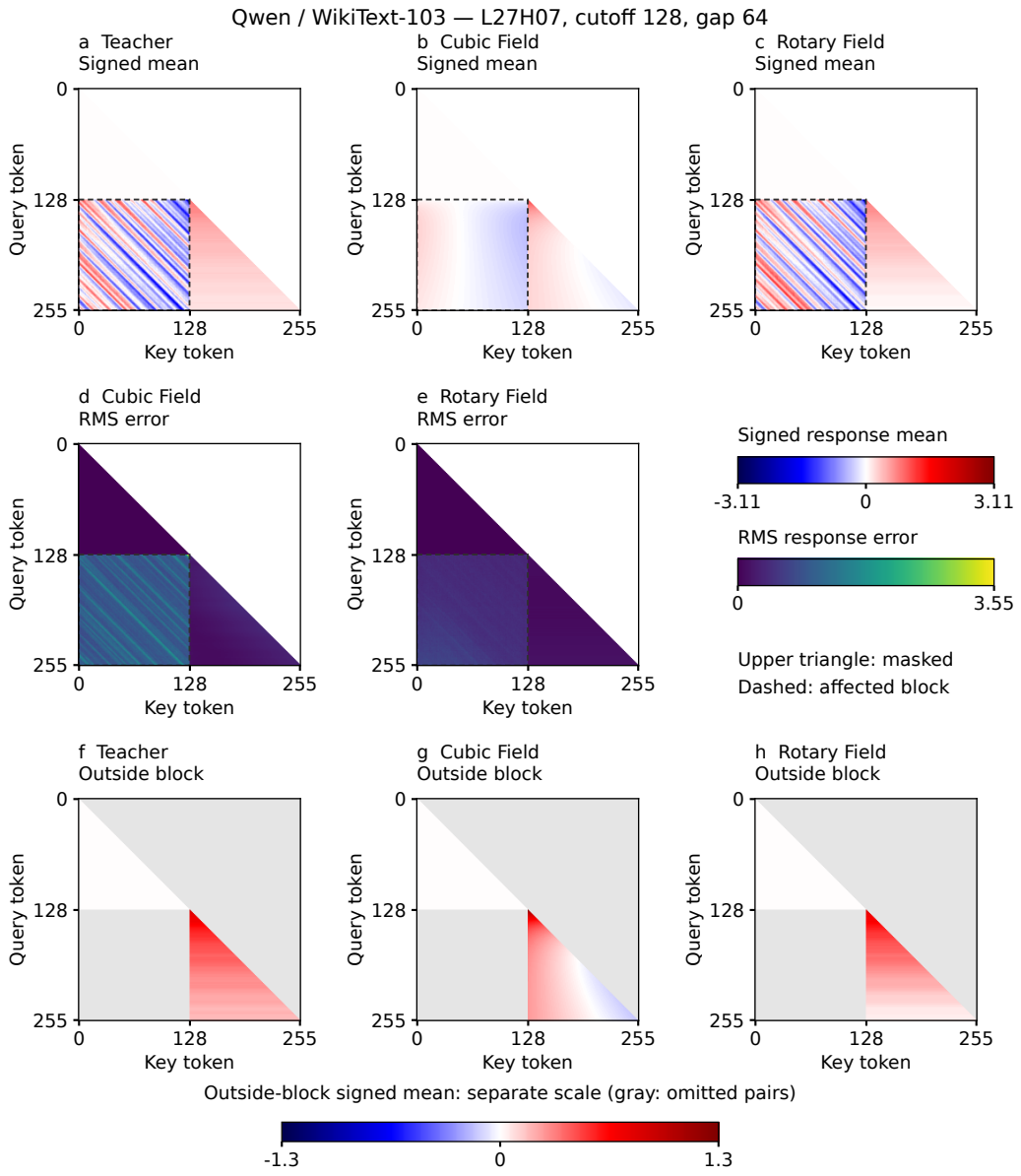

Figure 11: **Spatial response summaries: L27H07,** $(c, g) = (128, 64)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across the five non-L21H13 anchor heads; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

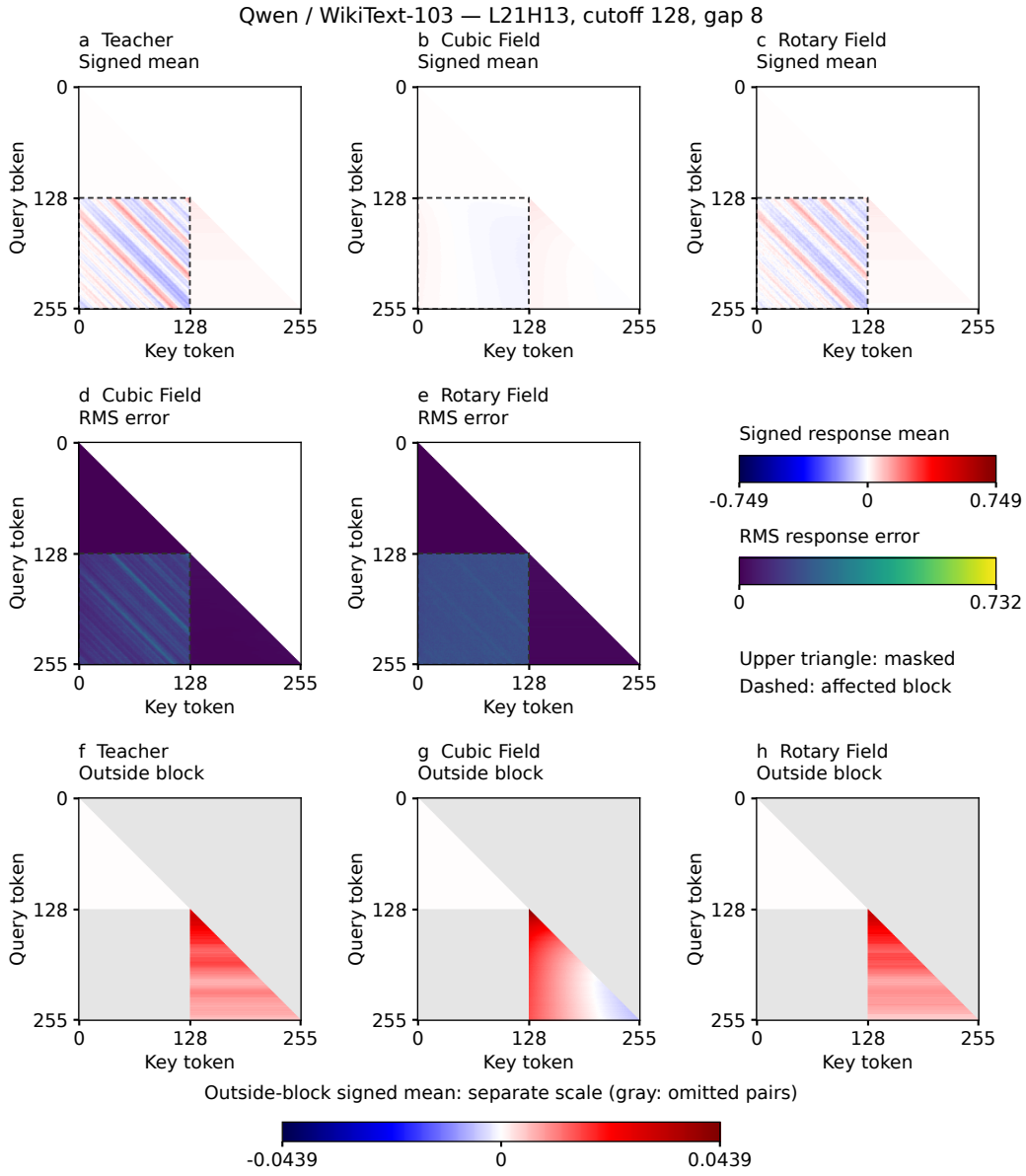

Figure 12: **Spatial response summaries: L21H13,** $(c, g) = (128, 8)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across L21H13 gaps 8, 64, and 128; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

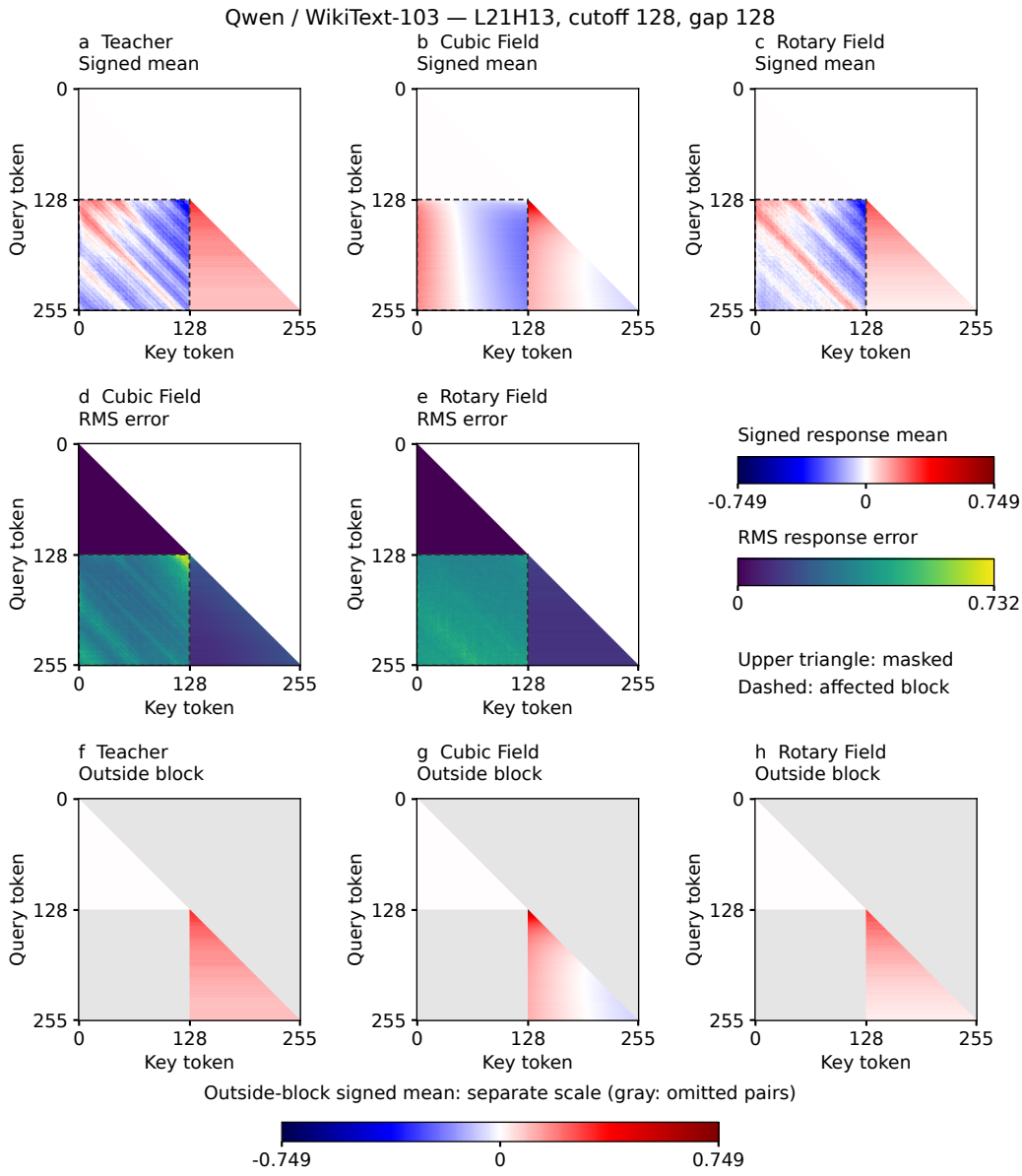

Figure 13: **Spatial response summaries: L21H13,** $(c, g) = (128, 128)$**.** Top: signed means of Teacher, Cubic Field and Rotary Field responses. Middle: Cubic Field and Rotary Field root-mean-square response errors relative to Teacher. Each accepted summary uses the fixed 512 response documents. Responses are row-centered over valid causal keys; full 256×256 original query/key coordinates are shown. White marks invalid causal pairs; dashed rectangles mark the affected suffix-query/prefix-key block. The original shared scales are retained across L21H13 gaps 8, 64, and 128; signed means and RMS errors use distinct scales, both in attention-score units. Bottom: the same signed means outside the dashed block, displayed with a separate scale shared by the three models in this figure; gray omits the affected block and invalid pairs. No values are clipped. These descriptive fields do not establish a mechanism.

J Output-Propagation Case Study

Score-response error concerns the intervened attention head. Its downstream consequence depends on how the remaining network propagates that error. The output case study evaluates this connection by inserting the predicted response into the target head and comparing the resulting token distribution with the teacher's intervened output. It uses a separate cohort from the conditional Qwen confirmations in Appendices C and D.

The Qwen/WikiText design uses 128 held-out output documents. It crosses the three historical failure heads with all 15 cutoff–gap targets and evaluates the three historical success controls at $(128, 64)$, giving 48 head–target cases. The failure panel tests whether score-level discrepancies propagate across interventions; the success controls provide a comparison at the anchor. Each case compares Cubic Field, Bilinear Field, Rotary Field, Fixed-frequency Rotary Field, and no response with the teacher's intervened output. The selected checkpoint responses are averaged before replacement, unlike the main study's average of seed-specific output KL. Adding a predicted response to the teacher's native score field separates response error from native reconstruction error. Values and non-target head outputs are retained at the replacement site, and the remaining network computes the downstream distributions.

Metrics capture different consequences of replacement: teacher-to-estimator KL in nats per valid token, signed NLL difference from the intervened teacher, centered-logit squared error, and top-1 agreement. Top-1 agreement is also reported on tokens whose intervened-teacher prediction differs from the native prediction, with its own denominator. Comparisons retain paired documents and the same valid-token masks. Normalized logit-response risk is unavailable for this cohort because the centered teacher logit-response energy was not retained. The other metrics provide a separate descriptive output comparison, with absolute differences reported alongside agreement rates.

T5 has a separate 45-case output design using encoder interventions and teacher-forced decoding (Appendix K). Its output panel, native-selected policy and token denominators are interpreted within that model and task. For both designs, the endpoint is fidelity to the teacher's intervened output, evaluated within the model-specific output cohort.

J.1 Descriptive consequences across the output grid

The separate output companion retains all 48 head–target cases and five policies. Each head–policy summary uses 32,640 valid prediction tokens from 128 documents. Table 21 reports the existing equal-head KL summaries for the three historical failure heads across all 15 targets; Table 22 reports the individual success-control anchor cases. These are descriptive estimates without new confidence intervals or significance tests.

At $(128, 64)$, output KL is 6.348×10^{-6} for Cubic Field, 14.573×10^{-6} for Rotary Field, 13.413×10^{-6} for Fixed-frequency Rotary Field and 5.538×10^{-6} for Null, in nats per valid token. Bilinear Field matches Null here. All three fitted positional surrogates have larger descriptive output KL than Null at this anchor. The comparison preserves the small absolute scale alongside the ranking.

Metrics capture different consequences. At the anchor, Rotary Field's equal-head agreement on teacher-changed tokens is 0.589, compared with 0.198 for Cubic Field, despite its higher KL. These fractions have only 14, ten and 15 teacher-changed tokens on the three heads. Overall agreement is near one and must not obscure those restricted denominators. Signed NLL differences can be negative relative to the intervened teacher; lower NLL alone is not greater fidelity to the teacher distribution.

Supporting data retain headwise KL and NLL quantities, centered-logit SSE, KL-normalized gain, and both top-1 numerators and denominators for every case. KL-normalized gain averages the headwise quantity $(K_{0,h} - K_{m,h})/K_{0,h}$ and is distinct from normalized logit-response risk. The latter remains unavailable because its original teacher-response energy was not retained. The 1,440 display rows are repeated metric views of the 48 cases, not independent observations. This descriptive companion does not supply results for the closed combined output-analysis family.

Table 21: Qwen/WikiText output KL (10^{-6} nats/token) across the full target grid, averaged equally over the three historical failure heads L21H13, L24H10, and L25H12.

c	g	Cubic Field	Bilinear Field	Rotary Field	Fixed-frequency Rotary Field	Null
64	8	0.590	0.546	5.293	4.519	0.546
64	16	1.407	1.259	7.567	6.187	1.259
64	32	3.100	2.628	8.836	7.985	2.628
64	64	6.917	5.485	16.166	15.667	5.485
64	128	14.778	11.552	36.616	33.016	11.552
128	8	0.611	0.599	4.621	3.992	0.599
128	16	1.477	1.410	6.589	4.947	1.410
128	32	2.893	2.683	8.312	7.017	2.683
128	64	6.348	5.538	14.573	13.413	5.538
128	128	14.391	11.789	31.957	26.819	11.789
192	8	0.385	0.392	2.664	2.131	0.392
192	16	0.894	0.903	3.792	2.922	0.903
192	32	1.695	1.681	5.357	4.446	1.681
192	64	3.555	3.419	8.844	8.733	3.419
192	128	8.161	7.207	17.662	15.076	7.207

Note. Descriptive points from 128 held-out documents; each head–policy summary has 32,640 valid predicted tokens. Values are teacher-to-estimator KL in nats per valid token, displayed in units of 10^{-6}. The accompanying data retain NLL differences, KL-normalized gain, centered-logit squared-error sums, and both overall and changed-token top-1 counts. Changed-token denominators can be sparse. Normalized logit-response risk remains unavailable because the original teacher logit-response energy was not retained. No confidence intervals or inferential decisions are asserted.

Table 22: Qwen/WikiText output KL (10^{-6} nats/token) for the three historical success controls at $(c, g) = (128, 64)$. These headwise descriptive values are separate from the failure-head averages.

Head	Cubic Field	Bilinear Field	Rotary Field	Fixed-frequency Rotary Field	Null
L25H09	3.887	4.121	11.904	11.290	4.121
L26H01	10.818	11.534	140.745	102.972	11.534
L27H07	19.178	18.609	101.632	101.312	18.609

Note. Descriptive points from 128 held-out documents; each head–policy summary has 32,640 valid predicted tokens. Values are teacher-to-estimator KL in nats per valid token, displayed in units of 10^{-6}. The accompanying data retain NLL differences, KL-normalized gain, centered-logit squared-error sums, and both overall and changed-token top-1 counts. Changed-token denominators can be sparse. Normalized logit-response risk remains unavailable because the original teacher logit-response energy was not retained. No confidence intervals or inferential decisions are asserted.

K EXPLORATORY T5 EXTENSION

K.1 DESIGN AND INFERENCE

The same position maps define a response when position enters attention through an additive bias. T5-large encoder attention supplies this separate test of native selection with bucketed relative-position bias on C4 English (Raffel et al., 2020). The validation-corpus design allocates 256 fitting, 64 validation, 512 response, and 128 output documents to mutually disjoint roles. Document selection excludes the documented prior uses; completeness of earlier project exposure is unknown. The extension therefore retains its exploratory interpretation. It is not a matched architecture-isolation experiment against the autoregressive models.

The panel contains four heads in each of eight layers: H0–H3 at L0 and L12, H4–H7 at L3 and L15, H8–H11 at L6 and L18, and H12–H15 at L9 and L23. Cubic Field, Bilinear Field, Rotary Field, and Bucket-aware Bilinear Field (T5B) use three fitted seeds, giving 384 fits and $32 \times 4 \times 15 = 1{,}920$ response cases. Native-validation selection fixes both checkpoints and the method policy before response evaluation. T5B uses the teacher's 32-bucket bidirectional geometry with maximum distance 128; Rotary Field remains a rotary surrogate, while the teacher itself uses additive bias.

For the position map in Eq. 21, the primary pair mask is $b(p_j - p_i) \neq b(j - i)$, intersected with token validity, where b is the teacher bucket function. Row-centering precedes mask selection. Cross-boundary pairs whose buckets are unchanged provide a separate leakage diagnostic. Breadth counts structurally distinct bucket-transition targets; a target with no changing pairs supplies no response evidence.

The output design crosses L0H0, L12H0, and L23H15 with all 15 targets, giving 45 cases. Deterministic span corruption defines 256-token encoder inputs and stored decoder targets, with noise density 0.15 and mean span length 3.0. Teacher-forced decoding measures the consequences of native selection, fixed methods, no response, and the teacher response. Score-response fidelity and output effects remain separate estimands. Output evaluation uses float32 forwards with fixed batches of four documents and the teacher's full-head attention contraction shapes. Replacement changes the selected head's valid score field while retaining its value vectors, native masking, and the other heads' outputs at that layer.

Native-validation attention KL selects Rotary Field on 20 heads, T5B on seven and Cubic Field on five, using the arithmetic mean over the three fitting seeds. These head-specific choices are fixed across targets. The output heads L0H0, L12H0 and L23H15 use Rotary Field, Rotary Field and Cubic Field, respectively.

For each target, score risk is computed as a ratio of summed document-level squared errors to summed teacher-response energy, then averaged equally over the 32 heads. The native-policy family contains 75 contrasts against Cubic Field, Bilinear Field, Rotary Field, T5B and no response. Its simultaneous two-sided 95% intervals use 10,000 paired source-document bootstrap draws. The 45 T5B recovery contrasts use a separate family of simultaneous one-sided lower bounds, and the 960 score-materiality statistics retain their own family. Uncertainty is over source documents within the fixed panel.

A selection-failure target requires at least 24 material heads, positive simultaneous lower bounds for native-policy risk against both null and a fixed competitor, and at least 16 material heads supporting both signs. A target satisfies the bounded-excess criterion when all five native-policy contrast upper bounds are at most 0.10; remaining eligible targets are classified as head-dependent. The separate T5B recovery criterion requires positive lower bounds against Cubic Field, Rotary Field and null and at least 16 material heads supporting all three signs.

K.2 NATIVE-POLICY EXCESS RISK IS BOUNDED ON 13 TARGETS

On 13 of the 15 targets, the simultaneous upper bounds place native-policy equal-head mean response risk no more than 0.10 above each of the five tested fixed alternatives. This one-direction bound establishes neither equivalence nor optimality and does not apply to every individual head. The two head-dependent targets are $(64, 128)$ and $(192, 128)$ (Table 24). Every target has 30 or 31 material heads, and all 15 satisfy the Bilinear Field null-response identity. No target meets the combined selection-failure criterion.

Table 23: T5 contrasts on the 15-cell grid. c: zero-based cutoff index; g: position-ID gap size; N: native policy. Equal-head means use 32 response heads and three output heads; KL is in 10^{-6} nats per valid decoder token. Positive values favor the comparator; bold marks positive simultaneous 95% lower bounds (two-sided response intervals; one-sided output bounds). Full bounds: Appendix K.

		Score-response risk		Output KL	
c	g	$\overline{R}_N - \overline{R}_0$	$\overline{R}_N - \overline{R}_{\mathrm{T5B}}$	$\overline{K}_N - \overline{K}_0$	$\overline{K}_N - \overline{K}_{\mathrm{T5B}}$
64	8	-0.464	-0.177	-0.517	0.062
64	16	-0.370	-0.090	-0.969	0.070
64	32	-0.454	-0.197	-0.439	0.672
64	64	-0.366	-0.151	0.502	$\mathbf{1.786}$
64	128	$\mathbf{0.211}$	$\mathbf{0.418}$	$\mathbf{2.353}$	$\mathbf{3.639}$
128	8	-0.460	-0.221	-0.233	0.312
128	16	-0.360	-0.129	-0.356	0.542
128	32	-0.463	-0.236	-0.796	0.651
128	64	-0.453	-0.249	0.499	1.669
128	128	-0.119	$\mathbf{0.083}$	$\mathbf{4.208}$	$\mathbf{5.382}$
192	8	-0.472	-0.191	-0.317	$\mathbf{0.124}$
192	16	-0.378	-0.107	-1.000	-0.117
192	32	-0.450	-0.200	-0.839	0.194
192	64	-0.381	-0.168	0.302	$\mathbf{1.516}$
192	128	-0.002	$\mathbf{0.206}$	$\mathbf{3.445}$	$\mathbf{4.659}$

The head-dependent cases reveal variation that the overall decision alone would conceal. At $(64, 128)$, native-policy mean risk exceeds null by 0.211 and T5B by 0.418, with simultaneous intervals $[0.195, 0.226]$ and $[0.403, 0.433]$. Only seven material heads support both signs against T5B, below the required 16 despite the supported aggregate differences. At $(192, 128)$, the native-minus-null contrast includes zero, while native risk exceeds T5B by 0.206. T5B meets its separate recovery criterion at this target: its risk improvements over Cubic Field, Rotary Field and null are 0.930, 2.613 and 0.208, respectively, with positive simultaneous lower bounds and the required head support.

The prescribed excess-risk bound therefore holds at most targets, alongside localized benefits from the bucket-aware control. These results concern the fixed T5/C4 panel and do not isolate architecture from corpus or task.

Table 24: T5 score-response contrasts over 32 fixed heads: equal-head means above simultaneous two-sided 95% intervals. N denotes the native-selected policy and 0 no response. Material heads meet score-energy/no-op criteria. Bounded excess requires all five native-minus-policy upper bounds ≤ 0.10; other eligible targets are head-dependent (Appendix K.1).

c	g	Material heads	$\overline{R}_N - \overline{R}_0$	$\overline{R}_N - \overline{R}_{\mathrm{T5B}}$	Target state
64	8	30	-0.464	-0.177	Bounded excess
			$[-0.468, -0.461]$	$[-0.181, -0.174]$	
64	16	30	-0.370	-0.090	Bounded excess
			$[-0.375, -0.365]$	$[-0.095, -0.085]$	
64	32	30	-0.454	-0.197	Bounded excess
			$[-0.458, -0.450]$	$[-0.201, -0.193]$	
64	64	30	-0.366	-0.151	Bounded excess
			$[-0.371, -0.360]$	$[-0.156, -0.146]$	
64	128	30	0.211	0.418	Head-dependent
			$[0.195, 0.226]$	$[0.403, 0.433]$	
128	8	30	-0.460	-0.221	Bounded excess
			$[-0.464, -0.457]$	$[-0.224, -0.217]$	
128	16	30	-0.360	-0.129	Bounded excess
			$[-0.365, -0.355]$	$[-0.134, -0.123]$	
128	32	31	-0.463	-0.236	Bounded excess
			$[-0.467, -0.458]$	$[-0.240, -0.231]$	
128	64	30	-0.453	-0.249	Bounded excess
			$[-0.457, -0.449]$	$[-0.253, -0.245]$	
128	128	30	-0.119	0.083	Bounded excess
			$[-0.125, -0.112]$	$[0.076, 0.089]$	
192	8	30	-0.472	-0.191	Bounded excess
			$[-0.476, -0.469]$	$[-0.195, -0.188]$	
192	16	30	-0.378	-0.107	Bounded excess
			$[-0.383, -0.373]$	$[-0.112, -0.102]$	
192	32	30	-0.450	-0.200	Bounded excess
			$[-0.455, -0.446]$	$[-0.204, -0.196]$	
192	64	30	-0.381	-0.168	Bounded excess
			$[-0.386, -0.377]$	$[-0.173, -0.163]$	
192	128	30	-0.002	0.206	Head-dependent
			$[-0.013, 0.009]$	$[0.195, 0.217]$	

K.3 Output rankings differ without meeting the material-propagation criterion

The 45 output cases compare the three fixed heads across all 15 targets on 128 documents. KL and NLL are normalized by valid teacher-forced decoder tokens; each head contributes 7,424 such tokens. Head-level token-weighted estimates are then averaged equally. Here N denotes the headwise native-selected policy and 0 the no-response baseline. The native-policy, null/gain and output-materiality comparisons retain separate families of 75, 120 and 15 statistics, with simultaneous one-sided 95% lower bounds from 10,000 source-document bootstrap draws.

T5B has lower descriptive output KL than no response at every target (Table 25). Six targets have a positive lower bound for native-policy KL minus T5B KL. Native-policy KL also exceeds no response at all three gap-128 targets, with positive simultaneous lower bounds. The absolute null-minus-T5B contrast has no positive lower bound in its family. The head-normalized T5B gain has a positive lower bound at 11 targets (Table 26). It averages $(K_{0,h} - K_{\mathrm{T5B},h})/K_{0,h}$ over heads, rather than taking a ratio of the displayed mean KL values. This head normalization weights effects differently from the absolute KL contrast.

The KL from the teacher's intervened distribution to its native distribution, K_0, ranges from 0.995×10^{-6} to 3.140×10^{-6} nats/token. No target meets the 10^{-5} nats/token output-materiality criterion, and none of the 60 method–target comparisons meets the combined ranking or material-propagation rule. Figure 4 shows the response and output contrasts across the grid. Individual contrast differences coexist with small absolute output effects. The rule requires positive lower bounds against both the native policy and null; material propagation additionally requires positive gain and output-materiality bounds and joint support from at least two score-material heads.

Top-1 comparisons have similarly limited changed-token support. There are only zero to five teacher prediction changes under intervention per head and target. All 75 comparisons restricted to these changed tokens lack sufficient support for their prescribed intervals; ten also have undefined equal-head point estimates because a head has no changed tokens. All-token agreement remains estimable. The accompanying numerical data retain every policy contrast, headwise contribution, logit-risk and NLL summary, and top-1 numerator, denominator and interval status.

Table 25: T5 output KL from the intervened teacher, in 10^{-6} nats per valid decoder token. Policy columns are equal-head means of three head-specific token-weighted estimates on the same 128 output documents. Bilinear Field equals no response (K_0). The final column gives simultaneous one-sided 95% lower bounds for output materiality, in the same units.

c	g	Cubic Field	Bilinear Field	Rotary Field	T5B	K_0	LB($K_0 - 10^{-5}$)
64	8	1.001	0.995	0.716	0.415	0.995	-9.538
64	16	1.812	1.377	1.057	0.338	1.377	-9.347
64	32	2.542	1.664	2.797	0.554	1.664	-9.091
64	64	4.023	1.914	6.652	0.630	1.914	-8.862
64	128	8.105	2.047	24.532	0.760	2.047	-8.738
128	8	1.467	1.057	1.111	0.512	1.057	-9.713
128	16	1.765	1.183	1.610	0.285	1.183	-9.630
128	32	2.802	2.395	3.586	0.948	2.395	-8.926
128	64	3.906	2.544	7.479	1.374	2.544	-8.846
128	128	6.764	2.760	27.337	1.585	2.760	-8.706
192	8	1.398	1.044	1.054	0.603	1.044	-9.391
192	16	2.089	1.884	1.933	1.001	1.884	-8.978
192	32	3.063	2.313	4.353	1.279	2.313	-8.726
192	64	3.941	2.743	9.074	1.529	2.743	-8.373
192	128	7.040	3.140	27.650	1.925	3.140	-8.139

Table 26: T5 output-KL contrasts and normalized gains. Equal-head estimates appear above simultaneous one-sided 95% lower bounds in parentheses. KL contrasts use 10^{-6} nats per valid decoder token; $\overline{G}_{\mathrm{T5B}}$ is the dimensionless mean of per-head gains. Families, gain definition and combined decision rules follow Appendix K.3.

c	g	$\overline{K}_N - \overline{K}_0$	$\overline{K}_N - \overline{K}_{\mathrm{T5B}}$	$\overline{K}_0 - \overline{K}_{\mathrm{T5B}}$	$\overline{G}_{\mathrm{T5B}}$
64	8	-0.517	0.062	0.579	0.483
		(-1.009)	(-0.103)	(-0.075)	(0.257)
64	16	-0.969	0.070	1.039	0.596
		(-2.171)	(-0.153)	(-0.335)	(0.475)
64	32	-0.439	0.672	1.110	0.570
		(-1.959)	(-0.332)	(-0.342)	(0.208)
64	64	0.502	1.786	1.284	0.645
		(-0.526)	(0.818)	(-0.329)	(0.256)
64	128	2.353	3.639	1.287	0.613
		(1.165)	(2.239)	(-0.341)	(0.226)
128	8	-0.233	0.312	0.545	0.453
		(-2.244)	(-1.446)	(-0.508)	(0.116)
128	16	-0.356	0.542	0.897	0.597
		(-1.964)	(-0.350)	(-0.793)	(0.494)
128	32	-0.796	0.651	1.447	0.493
		(-2.798)	(-1.062)	(-0.890)	(0.105)
128	64	0.499	1.669	1.170	0.476
		(-1.988)	(-0.180)	(-1.205)	(-0.088)
128	128	4.208	5.382	1.174	0.449
		(1.578)	(3.205)	(-1.264)	(-0.096)
192	8	-0.317	0.124	0.441	0.349
		(-1.266)	(0.008)	(-0.769)	(-0.665)
192	16	-1.000	-0.117	0.883	0.394
		(-2.184)	(-0.778)	(-0.534)	(0.067)
192	32	-0.839	0.194	1.033	0.374
		(-2.132)	(-0.580)	(-0.405)	(-0.010)
192	64	0.302	1.516	1.214	0.432
		(-1.094)	(0.602)	(-0.247)	(0.062)
192	128	3.445	4.659	1.215	0.379
		(1.390)	(2.780)	(-0.270)	(0.002)

K.4 NATIVE-DATA PREDICTIONS HAVE MIXED ACCURACY

The three anchor heads also test what a fitted surrogate predicts before observing the teacher's intervention response. For each head, a deterministic prediction card uses native-validation evidence and the selected surrogate's shifted scores to specify a fidelity category, a response-best method set, a structural pattern and a dominant query/key region. The native KL selector supplies the predicted best method. The fidelity rule predicts better-than-null risk when the surrogate's response RMS is at least 0.01 times the native centered-score RMS on affected pairs and exceeds 100 times its no-op RMS. Teacher counterfactual scores and the response/output cohorts are withheld during prediction.

At $(c, g) = (128, 64)$, the L0H0 and L12H0 cards match three of the four audited features: fidelity category, overlap with the response-best method set, and query/key region (Table 27). The L23H15 card matches none. All three structural labels disagree with the audited outcome. Their predictions inherit the selected surrogate's form: learned-frequency non-bucket structure for Rotary Field and smooth cubic relative-position structure for Cubic Field. In these cases, agreement on method ranking and response location does not ensure a correct structural description.

All three cards abstain on output direction, so that dimension is unscored. The fractions below summarize four categorical checks within each fixed case, not independent trials or population prediction accuracy. The accompanying data retain the original predictions, native selections and audit outcomes.

Table 27: Native-data prediction audits at the fixed T5 target $(c, g) = (128, 64)$. Yes/No indicates agreement with the held-out outcome for each categorical check. Best-set overlap permits tied response-best methods. Correct counts summarize these four checks within each case. All three cards abstain on output direction, which is unscored.

Head	Native choice	Fidelity category	Best-set overlap	Structure	Region	Correct
L0H0	Rotary Field	Yes	Yes	No	Yes	3/4
L12H0	Rotary Field	Yes	Yes	No	Yes	3/4
L23H15	Cubic Field	No	No	No	No	0/4

L Reproducibility and Evidence Scope

The anonymous supplement contains retained numerical summaries, spatial arrays, figure inputs, study-specific scientific source, and presentation scripts for the main four-setting study, expanded response and insertion studies, and separate T5 extension. Table 28 summarizes the main-study inventory. The supplementary `README.md` provides the quick-start commands; `REPRODUCIBILITY_README.md` describes the files, environment and computational resources.

The CPU entrypoint regenerates numerical tables and quantitative figures from the retained summaries into a new output directory. Pre-rendered tables and figures are not bundled. Table reproduction uses the Python standard library; figure reproduction additionally uses NumPy and Matplotlib. The commands preserve the supplied estimates, intervals and decisions without repeating model evaluation or statistical inference.

The expanded materials retain full target grids, headwise values, stored intervals, spatial arrays and masks. The four-method comparison includes 1,920 response entries and 384 native-fit summaries, with historical selection records distinguished from native remeasurements. The T5 extension retains its separate response, output and native-prediction audits. Normalized logit-response risk remains unavailable for the descriptive Qwen output companion because its original teacher logit-response energy was not retained.

The scientific source documents the implemented methods and estimators. Full experimental or statistical reruns additionally require the matching source dependencies, original document-level records, model and dataset assets, and experiment environment described in `INPUTS.md`. Model weights, raw corpora, fitted checkpoints and complete document-level record streams are not included in this ZIP. The main Qwen/Pythia experiments used an NVIDIA RTX 3090; recorded software versions and computational resources are provided in the supplementary README.

Table 28: Inventory of the main Qwen/Pythia experiments summarized in the anonymous supplement. Counts describe the underlying studies, not the number of files distributed. Output evaluation opens only for heads meeting the score-counterexample criteria. Pythia record counts include fitted-model records.

Setting	Fits	Score records	Score decision	Output decision
Qwen / OpenWebText	192	832	Prior stratum passes; 3 heads qualify	Ranking only; 3 heads
Qwen / WikiText-103	192	832	3/3 replicate; 4 heads qualify	Ranking only; 4 heads
Pythia / WikiText-103	72	312	0 reversals	Not applicable
Pythia / OpenWebText	72	312	Agreement replicates; 0 reversals	Not applicable
Spectral diagnostic	—	44 heads	Both criteria fail	Not applicable