WHO GRADES THE EDIT? FROZEN AND RE-FIT READOUTS FOR POSITION INTERVENTIONS IN ROPE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary position embeddings (RoPE) encode relative token offsets, yet a linear readout of the residual stream tracks absolute position. Existing position-bias mitigations intervene on this coordinate. We ask whether an edit disrupts that readout or removes the information it reads. We evaluate upstream edits using a probe frozen before the edit and a probe re-fit on the edited activations. In Gemma, blocking the attention-sink value reduces position R^2 by 17.404 for the frozen probe but only 0.059 for the re-fit probe. The frozen score falls below zero even though most position information remains linearly recoverable. After removal of the fitted leading direction, re-fit probes retain 74–88% of baseline position R^2 across four base models. A separate behavioral audit asks whether editing a fitted position basis changes position-biased retrieval. On Llama-Instruct multi-document QA, sequence-wide basis removal flattens the lost-in-the-middle dip beyond a matched decoder-null control. The control also changes behavior, so the relevant quantity is the paired difference: -0.020 (95% CI: $[-0.031, -0.011]$). This finding does not establish information erasure and is limited to the tested model, task, layer, basis, and decoder; Qwen's behavioral audit stops because a control fails. Evaluating a position edit requires distinguishing readout disruption, remaining recoverability, and a separately controlled behavioral effect.

1 INTRODUCTION

An absolute-position coordinate is readily decodable in RoPE models, and prior work already uses it as an editing target. RoPE (Su et al., 2024) encodes relative position through query–key phase differences, without learned absolute-position embeddings. A linear probe nonetheless recovers absolute token index from the residual stream. In our four base models, its document-averaged predictions recover the ordering of token positions almost perfectly (Appendix Table 17). Long-context models retrieve information less reliably from the middle of their input than from either end (Liu et al., 2024), and position-bias mitigations target this coordinate to reduce that gap. Position therefore provides a useful test case: the editing target is established, and the positional answer profile measures the bias the edit aims to reduce. Yu et al. (2025) scale such a coordinate and alter position-biased retrieval. Ruscio et al. (2026) project out the direction that best predicts absolute position, with only a small reduction in attention-logit leakage.

An edit can break a readout without removing the information it is intended to suppress. In Gemma, blocking the value copied from the attention sink drives the frozen probe's position R^2 far below zero. A re-fit probe still retains most of the baseline position R^2 (Table 2). The comparison holds the prediction target, metric, and evaluation tokens fixed, and the edit passes the committed content check. The frozen probe's failure therefore greatly overstates the loss of linear recoverability.

Re-fitting tests what remains linearly recoverable without requiring the pre-edit decoder to remain calibrated. We begin with this frozen–re-fit comparison for upstream Formation edits (§4). Two complementary audits then test what survives direct removal of the fitted direction (Representation, §5) and which heads contribute beyond matched alternatives (Writing, §6). Finally, a separate Use audit measures behavior, not post-edit recoverability (§7). These levels address distinct claims rather than trace a circuit; Table 1 summarizes their controls and the conclusions they support.

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.

Our contributions are:

- **We compare frozen and re-fit readouts for upstream position edits.** With the target, metric, and evaluation tokens held fixed, sink-value blocking can disrupt the frozen readout while largely preserving recoverability, whereas visible-set edits cause larger re-fit losses.
- **We characterize redundancy and head contributions across models.** Position survives leading-direction removal, and selected head sets have larger effects than matched alternatives under both writer diagnostics in three of four models. A post-hoc surprisal control tests the scope of these two findings.
- **We separately establish a behavioral treatment–control difference.** On held-out Llama-Instruct QA, fixed-basis removal flattens the dip beyond an active decoder-null control matched on support and displacement. The result is conditioned on the tested basis and decoder (Table 4).

2 Related Work

Probing work established that decodability alone does not imply causal relevance (Hewitt & Liang, 2019; Belinkov, 2022; Geiger et al., 2021). Intervention studies address this limitation by editing activations and measuring behavioral consequences (Vig et al., 2020; Meng et al., 2022). Concept-erasure work shows that removing a single direction can leave a concept linearly decodable. Ravfogel et al. (2020) therefore iterate nullspace projection while new classifiers continue to recover the protected attribute. The behavioral audit is motivated by Elazar et al. (2021), who pair erasure with behavioral evaluation. Our contribution is an empirical audit of position interventions using this readout distinction. Motivated by coordinate-scaling edits, we test removal of the probe's *fitted* leading direction rather than optimal linear erasure.

Position-bias research provides the coordinate targeted by our interventions. Absolute position is recoverable in causal Transformers with no explicit positional encoding (Haviv et al., 2022; Kazemnejad et al., 2023; Zuo et al., 2025). Wu et al. (2025) trace part of that effect to causal masking, which biases attention toward earlier positions. Lost-in-the-middle work characterizes position bias (Liu et al., 2024) and modifies encodings and masks (Zhang et al., 2024). Attention-sink and spectral analyses identify further possible carriers (Xiao et al., 2024; Cancedda, 2024; Barbero et al., 2025). Our intervention target comes from Yu et al. (2025): a hidden-state coordinate varies monotonically with absolute position, and scaling that coordinate in last-token attention alters position-biased retrieval. Wang et al. (2025) eliminate position bias by a mechanistic edit. In attention logits, Ruscio et al. (2026) decompose position leakage into causal-mask and residual-stream components. Our Formation audit complements these behavioral and attention-logit endpoints by comparing frozen and re-fit residual probes on the same token-index task. This tests whether an upstream edit impairs the pre-edit decoder or reduces the position information a new linear readout can recover (Appendix A).

3 Experimental Framework

We distinguish an edit's effect on a frozen coordinate, the information accessible to a new linear probe, and behavior. Table 1 maps each audit to the controls its claim requires.

The frozen coordinate. We fit a ridge decoder from the post-block residual $h_l(t)$ at layer l to token index t on FineWeb-Edu splits (Lozhkov et al., 2024). We fix the peak-R^2 layer L^*, the decoder direction β, and the residual mean μ. These define the scalar readout:

$$C_\beta(t) = (h_{L^*}(t) - \mu)^\top \beta. \tag{1}$$

We call C_β the *frozen coordinate*. We implement interventions with TransformerLens hooks (Nanda & Bloom, 2022).

The two drop scores. In Formation, both scores measure the loss in position R^2 relative to each probe's unedited baseline. $\mathrm{drop}_{\mathrm{frozen}}$ uses the frozen probe, while $\mathrm{drop}_{\mathrm{refit}}$ uses a probe fitted afresh on the edited activations. A large frozen drop with a small re-fit drop indicates readout disruption with little loss of linear recoverability. Because out-of-sample R^2 can be negative, a drop can exceed

Level:	**Formation**
Question:	Which upstream computations change recoverable position information?
Intervention:	Mask, sink-value, and RoPE-band edits applied under one operator specification fixed before the runs.
Control or check:	A per-cell content check, phase-shuffle and high-frequency-band controls, and a count-matched scramble. The scramble fails the content check in three models; the high-frequency control fails in two. These cells cannot support causal attribution (Appendix E).
Supported conclusion:	Sensitivity of the re-fit readout to the tested operator; neither a pure counting effect nor a unique source of position information.
Level:	**Representation**
Question:	Does position information survive removal of the direction a probe selects?
Intervention:	Remove the fitted leading probe direction from the residual stream.
Control or check:	No matched control edit at this level. The edit removes the frozen probe's own direction, leaving the re-fit readout as the only score; the result is descriptive.
Supported conclusion:	Whether a re-fit probe still recovers position after the edit; not that any coordinate or subspace uniquely stores it.
Level:	**Writing**
Question:	Do the selected heads contribute more to position decoding than matched alternatives?
Intervention:	Ablate the preselected head set on documents disjoint from the sample that ranked it.
Control or check:	1,000 random sets matched on layer and head count. Next-token loss is recorded but not used as a validity gate.
Supported conclusion:	A larger effect than matched random sets; not a unique set of writers or specificity beyond the matched factors.
Level:	**Use**
Question:	Does the basis edit change behavior beyond a matched control?
Intervention:	Apply the edit at fixed support and dose on an untouched record block.
Control or check:	A no-op identity reference arm; a perturbation matched on support and displacement in a decoder's nullspace; a complement placebo; an α=0 potency anchor on a separate block.
Supported conclusion:	A difference conditional on the tested model, task, layer, basis, and decoder; not a causal channel that generalizes across models.

Table 1: Position-audit protocols by level. The four levels are audited separately and do not demonstrate a Formation–Writing–Representation–Use chain. Each conclusion is limited by the corresponding controls and checks.

one. A *cell* is a model–intervention or model–control pair. The writer audit instead compares frozen rank correlation with re-fit R^2; those scores differ in metric and aggregation as well as in the decoder (§6).

Vocabulary. Multidimensional edits use a greedily fitted 32-direction basis U. "Full U" means all 32 directions, and "sequence-wide" means every position. The *content check* requires perplexity to remain within a specified fraction of baseline, 0.15 in Formation. We call a rule *committed* when it was fixed before the runs it governs. We reserve *preregistered* for rules in the original specification; the amendment log in Appendix B lists those added later, including the content check.

Models and transfer boundary. The Representation, Writing, and Formation levels use Llama-3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B (Qwen et al., 2025), Gemma-2-9B (Gemma Team et al., 2024), and Pythia-2.8B (Biderman et al., 2023). For the behavioral experiments, position constructs are fitted to instruct variants at a variant-specific L^* before intervention. Across the three base–instruct pairs, the leading position directions are aligned, while the full 32-direction bases overlap weakly (Appendix Table 16). SmolLM3-3B (Bakouch et al., 2025) tests readout maintenance through its no-position-encoding (NoPE) layers.

Intervention validity and uncertainty. Interpretation requires the matched control edit—the *placebo*—to pass its own validity checks. Intervals are percentile bootstraps at the stated sampling unit; behavioral contrasts resample paired questions. We use Holm correction (Holm, 1979) and equivalence tests (Schuirmann, 1987; Lakens, 2017) for the specified families. Appendix C defines the checks, dose conventions, resample counts, and Monte Carlo resolution. Appendix Table 9 records the status and evidential boundary of each claim family.

Intervention	Readout	Pythia	Qwen	Gemma	Llama
sink-value block	frozen drop	0.034	−0.005	17.404	0.409
sink-value block	re-fit drop	0.004	−0.046	0.059	0.070
visible-set clamp	re-fit drop	0.524	0.453	0.618†	0.557
lowest-frequency derotation	re-fit drop	0.001	−0.001	−0.007	−0.000

Table 2: Readout disruption and remaining recoverability under Formation edits. All entries are losses in position R^2 relative to each probe's unedited baseline. The paired sink-value rows show that a large frozen drop need not imply a comparable re-fit loss; the remaining rows contrast two other upstream edits. Both scores use eight evaluation documents from the 16-document grid; entries are point estimates without cell-level intervals. †Gemma's clamp raises perplexity by about 20% and is content-confounded. Full matrix in Appendix Table 14.

4 Formation: Readout Disruption versus Recoverability

The clearest frozen–re-fit contrast comes from blocking the value copied from the attention sink. In Gemma, this edit reduces position R^2 by 17.404 for the frozen probe but only 0.059 for the re-fit probe (Table 2). Both probes predict token index and use R^2, and the edit passes the content check. The frozen readout fails while most position information remains linearly recoverable.

Evaluation. We split 16 FineWeb-Edu windows of 2,048 tokens in stream order, using eight to fit condition-specific ridge probes and eight for evaluation. Frozen and re-fit R^2 are pooled over the same evaluation tokens, with the partition and ridge penalty fixed across baseline and edits.

4.1 Sink-value blocking largely preserves recoverability

Llama shows the same qualitative pattern as Gemma, while Qwen and Pythia change little on either readout. In Gemma, most of the position information a re-fit probe recovers remains available without the sink value. Prior analyses describe the sink as a key bias that holds attention mass without contributing through the value path (Gu et al., 2025), and the position-0 activation as a near-input-independent bias term (Sun et al., 2024). These accounts are consistent with the large frozen drop and small re-fit loss; our preregistered prediction instead required re-fit degradation (Appendix B).

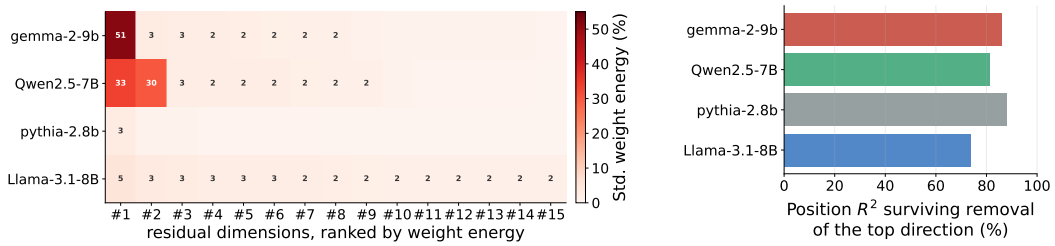
4.2 Visible-set and position-0 edits reduce recoverability

Among the tested interventions, mask and cardinality edits produce the largest re-fit losses. Clamping the visible set reduces re-fit R^2 by 0.45–0.62 across the four models. Qwen remains below the committed 0.5 threshold. The content check passes in Pythia, Qwen, and Llama, but fails in Gemma.

The clamp narrows each query's trailing window and fixes position-0 attention at a baseline weight. Excluding position 0 also damages recoverability in Pythia, Qwen, and Llama; Appendix E reports it with the count-matched scramble control. These edits change cardinality, recency, and visible content together, so they do not isolate pure counting. Qwen's clamp and position-0 cells motivated the readout amendment: the frozen score alone does not capture the losses measured by a newly fitted probe (Appendix Table 9).

4.3 The tested low-frequency band leaves the readout intact

Derotating the lowest-frequency RoPE planes leaves the frozen coordinate intact in every model, and the phase-shuffle control is null as well. Derotating every plane moves the readout in all four models. Llama's tested band fails the manipulation check because its realized rotation is too small, leaving that null uninterpretable. The null results in the other models apply only to the tested band; their realized rotations are reported in Appendix C. Cross-band interactions and attention-mediated routes remain untested. The separate SmolLM3 maintenance test is reported in Appendix Figure 5.



(a) Standardized-weight concentration.

(b) Recoverability after removal.

Figure 1: Concentration and redundancy are different properties. Panel (a) ranks the top 15 residual dimensions by squared standardized decoder weight, $(\beta_j\sigma_j)^2/\sum_i(\beta_i\sigma_i)^2$, for weight β_j and dimension standard deviation σ_j. This coefficient-energy summary excludes cross-dimension covariances: it is not a decomposition of total decoder variance, $\beta^\top\Sigma\beta$ for residual covariance Σ. Single-dimension R^2 relative to the full probe's score is a separate measure and ranks Qwen differently (Appendix D). Panel (b) shows evaluation-set position R^2 after leading-direction removal and re-fitting, as a percentage of the frozen probe's pre-removal score on the same documents.

4.4 Threshold-based verdicts depend on the readout

The two readouts agree in 23 of 28 Formation cells at the committed threshold. In all five disagreements, only the frozen drop exceeds the threshold (Appendix Table 13). Three of these cells pass the content check; two are content-confounded. Among the passing cells, Gemma's sink edit largely preserves recoverability, while the position-0 edits still cause substantial re-fit loss. A nearby threshold produces disagreements in both directions (Appendix E). The central conclusion therefore rests on the Gemma comparison, not a particular disagreement count: a failed frozen readout need not mean that position is no longer linearly recoverable.

5 Representation: Position Survives Removal of Its Leading Direction

The sink result leaves open what happens when an edit targets the readout direction itself. Removing that fitted direction makes the frozen decoder constant by construction. Yet a decoder re-fit on the edited activations retains 74–88% of baseline position R^2 on the evaluation set across the four base models (Figure 1b). This holds even for Gemma, whose decoded position is concentrated in one residual dimension.

Gemma's best single-dimension readout attains 54% of the full frozen probe's pooled position R^2 (95% CI: 52–55%). The corresponding fractions are much smaller in the other models (Appendix D). Both concentration summaries describe the unedited representation through the frozen probe; redundancy instead describes what remains recoverable after an edit.

A post-hoc control repeats the removal for current-token surprisal, $-\log p(x_t \mid x_{<t})$. Most of that probe's evaluation-set R^2 also survives in every model, although single-coordinate concentration changes with the target. The surprisal probe reaches its highest score at a deeper layer in every model, so these comparisons confound target with depth. Appendix D reports the per-model values, sample split, and denominators.

6 Writing: Selected Heads Exceed Matched Controls in Three Models

Recoverability after leading-direction removal does not show whether some heads contribute more than others. We therefore rank heads by how strongly their contribution to the frozen coordinate covaries with token index. We then fix the top 15 head identities and ablate them on documents disjoint from the ranking sample. Their ablation causes larger losses than matched random-head ablations under both diagnostics in three of four models (Table 3).

	Frozen-ρ readout			Re-fit R^2 readout		
Model	Damage	Null 99th	Holm-p	Drop	Null 99th	Separation
Llama	0.965	0.004	0.004	0.538	0.039	both readouts
Pythia	0.193	0.018	0.004	0.478	0.008	both readouts
Gemma	0.022	0.002	0.004	0.139	0.013	both readouts
Qwen	0.095	1.153	0.315	0.449	0.241	re-fit only†

Table 3: Held-out position writer-head audit against 1,000 random sets matched on layer and head count. Damage is the loss in rank correlation with token index; drop is the loss in re-fit R^2. The metrics and aggregation differ, so the comparison does not isolate re-fitting. Null columns give the exclusion null's 99th percentiles. Holm-p covers the four preregistered frozen-ρ tests; re-fit separation is assessed only after the frozen test passes. †Qwen's re-fit result comes from a separate exploratory Holm family and does not change its frozen-readout result. The nulls do not match output norm or displacement. Protocol and inclusion-null results appear in Appendices C and E.

Two different diagnostics. Unlike Formation, this audit compares frozen Spearman ρ with re-fit R^2. The frozen score correlates the document-averaged curve with token index; the re-fit score measures token-level prediction error on a held-out subset of those documents. Both the metric and the evaluation aggregation change, so their disagreement does not isolate the effect of re-fitting.

In Qwen, the diagnostics diverge. Among random sets matched on layer composition and head count, 31% cause at least as much frozen-ρ damage as the selected set. None of the 1,000 sets causes as much re-fit damage. The re-fit comparison is an exploratory reanalysis of logged arrays with its own Holm correction; it does not change the preregistered frozen-readout result.

These comparisons establish relative contribution, not specificity: the controls match neither residual displacement nor per-head output norm. Gemma's frozen-ρ loss is the smallest across models, yet larger than the loss from every matched random set. Next-token loss remains close to baseline across models, but is recorded rather than used as a validity gate (Appendix E). Contributions from heads outside the top-15 sets also covary with token index, so the selected sets are not exhaustive.

A post-hoc rerun selects heads for current-token surprisal. Their ablation causes larger frozen-ρ losses than matched random-head ablations in Pythia and Gemma. Llama shows separation only in the exploratory re-fit comparison; Qwen shows none under either diagnostic. This arm uses fewer random draws and includes no next-token-loss check (Appendix D).

7 USE: A SEPARATE BEHAVIORAL AUDIT

The preceding audits measure readout disruption and recoverability in base models. We now ask a separate question in instruct variants: does removing a fitted position basis change behavior beyond a matched control? The position constructs are fitted to these variants before intervention; post-edit recoverability is not measured in this behavioral battery. On Llama-Instruct multi-document QA, sequence-wide removal of the fixed position basis flattens the lost-in-the-middle dip beyond a matched control (Table 4).

The behavioral endpoint. The endpoint is the dip: the reduction in answer log-probability when the gold document is in the middle of the context rather than at either end. We use the lost-in-the-middle benchmark (Liu et al., 2024), with 20-document instances from Natural Questions (Kwiatkowski et al., 2019) and ten tested gold-document slots. Let $\ell_{q,p}$ be the mean teacher-forced log-probability of the first gold-answer alias. The dip D_q contrasts end slots with middle slots, and ΔD is its mean change under an intervention:

$$D_q = \tfrac{1}{2}(\ell_{q,0} + \ell_{q,19}) - \tfrac{1}{2}(\ell_{q,8} + \ell_{q,11}), \qquad \Delta D = \frac{1}{n_q}\sum_q \left(D_q^{\text{arm}} - D_q^{\text{identity}}\right). \tag{2}$$

A negative ΔD indicates a flatter profile, not necessarily better performance. A prompt-only greedy audit measures the dip in normalized exact match (EM) on the same questions.

The held-out comparison. On previously untouched questions, we compare full-U magnitude removal with a *decoder-null control*. Within U, the control edits the nullspace of a second, independently fitted scalar position decoder, preserving its output while matching treatment support

Arm / contrast	Log-probability	Exact match	Interpretation
Identity (level D)	$+0.1811$ $[+0.1057, +0.2609]$	0.0625 $[0.0125, 0.1188]$	baseline dip present
Full-U magnitude	-0.0407 $[-0.0657, -0.0187]$	-0.0500 $[-0.1000, -0.0063]$	minimum-effect rule missed
Decoder-null control	-0.0205 $[-0.0402, -0.0022]$	—	active; equivalence gate passed
Complement placebo	-0.0095 $[-0.0292, +0.0101]$	—	interval includes zero; containment passed
Magnitude − decoder-null	-0.0202 $[-0.0312, -0.0109]$	-0.0250 $[-0.0500, -0.0063]$	primary contrast; Holm pass

Table 4: Behavioral comparison on previously untouched Llama-3.1-8B-Instruct QA questions ($n_q = 80$), with sequence-wide edits at L^* over the fixed basis U. Estimates are dip changes ΔD with 95% paired question-bootstrap CIs; identity reports the baseline level D (Eq. 2). Negative changes mean a flatter profile. Exact match uses prompt-only greedy decoding on the same sample. The complement placebo matches rank and displacement outside $\mathrm{span}(U)$. Decoder-null equivalence uses a 90% TOST interval; complement containment uses a 95% interval, both within the committed ± 0.065 band. The primary log-probability contrast survives the committed Holm family ($m = 4$). Sample records and full statistics appear in Appendix F.2.

and realized displacement. The paired difference is negative under both endpoints (Table 4). The decoder-null control also flattens the dip: it passes its equivalence gate, but its 95% interval excludes zero. The paired contrast therefore measures the additional flattening relative to this active control.

Full-U magnitude removal does not meet our minimum-effect rule: its 95% interval is not entirely beyond the ± 0.065 band. We fixed the band in advance from an earlier ablation anchor on a block with a larger baseline dip. The direct paired treatment–control contrast was specified separately and does not use that band (Appendix F.2).

Support and dose. The QA transplant produces a nearly flat profile as donor position varies with the gold document fixed (Figure 2a). This span-restricted result remains descriptive because its *potency control*, which verifies that the endpoint responds to intervention, was sequence-wide. Within-span magnitude removal also has an interval containing zero (Appendix Table 21). On the marker-free ordinal tasks, transplants passing the content-leak check reach at most 0.9% of the same-span full-residual-swap reference (Appendix Table 20).

With sequence-wide support, scaling the basis produces a monotonic dose response even under displacement matching (Figure 2c; Appendix Table 21). Reversing the position code also flattens the dip. On a disjoint sample, the reversal effect retains its sign but is roughly half as large and does not meet the minimum-effect rule (Figure 2d; Appendix F).

Boundaries. Global displacement alone does not account for the response (Appendix F.4), and a sequence-wide shift of the coordinate does not translate the answer profile (Appendix F). Overall EM is similar across the fresh audit's arms (Appendix F.2); a flatter dip therefore does not establish an improvement in overall performance.

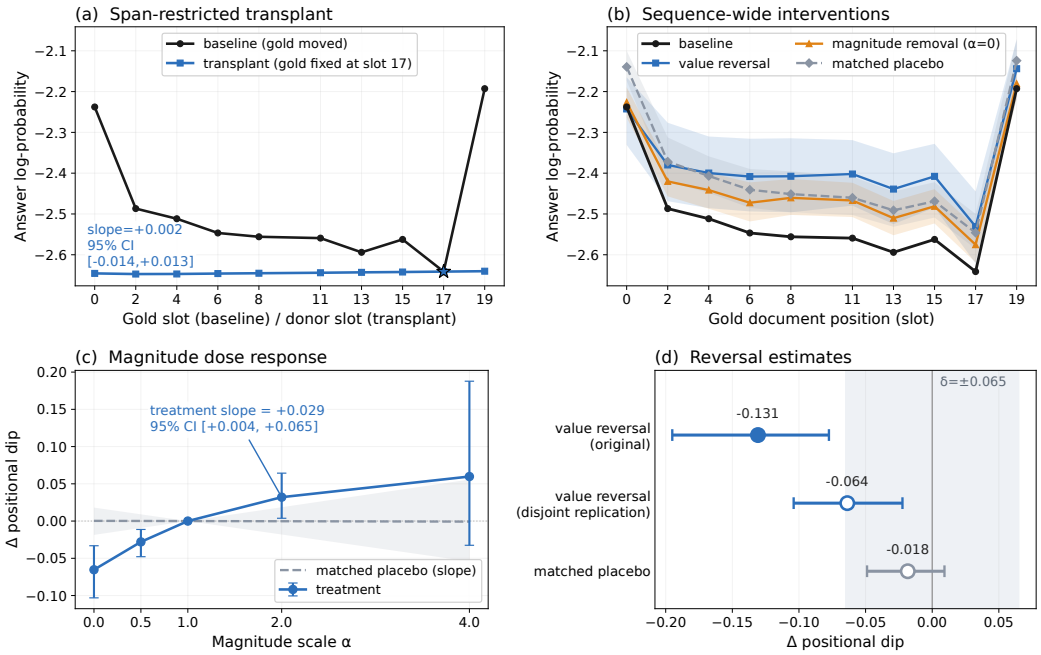


Figure 2: Support and dose in Llama's original behavioral batteries (records 0–79), plus the disjoint reversal replication. (a) Donor position varies with gold fixed at slot 17; the baseline instead moves the gold document. The transplant comparison remains descriptive because its potency control was sequence-wide. (b) Sequence-wide magnitude removal and value reversal alter the positional profile; the plotted complement placebo shifts the level without flattening the dip. (c) Dose response, showing the original treatment slope and the displacement-matched placebo slope; the matched treatment rerun is reported in Appendix Table 21. (d) Original and disjoint-sample reversal effects with 95% CIs and the minimum-effect band; the sign replicates, but the replication does not meet the rule.

8 Discussion and Limitations

The central result is that failure of a frozen readout need not imply loss of linearly recoverable position.

Unresolved calibration effects. For the Formation edits, we did not test intercept-only or affine recalibration of the frozen predictions. Re-fitting establishes retained linear recoverability, but does not show whether changing only the frozen readout's offset or scale would suffice.

No cell-level uncertainty for Formation. Each estimate uses eight evaluation documents. The per-document re-fit inputs were not retained, so uncertainty intervals cannot be reconstructed. We therefore limit comparisons to order-of-magnitude differences or the committed threshold. Threshold sensitivity (§4.4) does not replace those intervals.

Compound mask intervention. The visible-set clamp changes cardinality, recency, and visible content together, so it does not isolate counting. Gemma fails the content check; the mask-dependence claim rests on Pythia, Qwen, and Llama.

No optimal-erasure baseline. The Representation edit removes the *fitted* leading probe direction; we do not run least-squares-optimal linear erasure (Belrose et al., 2023). The rank needed to remove a fixed fraction of the position information decodable from the complement varies across refits (Appendix Table 26). Whether an optimal projection removes more information at equal rank remains open.

Writer diagnostics and controls. Qwen's exploratory disagreement involves different metrics and aggregation schemes, not re-fitting alone (§6). The controls leave output norm and residual displacement unmatched, so they assess relative contribution, not specificity.

Behavioral scope. The behavioral claim covers one Llama model, task, layer, basis, and decoder. No replicated treatment meets the committed minimum-effect rule. The threshold equals the earlier anchor's point estimate, from a block with a larger baseline dip (§7). The rule requires the entire interval below that value, not merely a matching point estimate. The Qwen audit stops because its complement-placebo interval extends beyond the committed band; downstream arms remain uninterpreted (Appendix F.6). Gemma's full-removal anchor interval includes zero, so its positive slopes remain descriptive (Appendix Table 21).

Rank-one behavioral arm inconclusive. Leading-direction ablation changes the dip less than full-basis removal does, but its paired treatment–control interval includes zero. Excluding a single question changes the conclusions of both that contrast and the rank comparison. These results do not establish a rank-one effect beyond the matched control (Appendix F.3).

What the decoder-null control preserves. The active control is confined to the nullspace of one scalar decoder inside U, and every column of U is fitted to decode position. The control may therefore perturb redundant position information rather than cause nonspecific residual disruption. A narrower nullspace control has a near-zero effect on another question block (Appendix F.2), but both nullspace width and sample differ between the audits.

8.1 A PRACTICAL PROTOCOL FOR EVALUATING EDITS

The combined findings suggest three checks, each tied to a different claim.

Compare readouts on equal terms. Fix the intervention site, target, probe family, and evaluation documents before scoring. Fit the post-edit probe on a separate partition shared across baseline, treatment, and controls. Score both probes on the same evaluation examples using the same metric and their respective unedited baselines. Retain per-document inputs for paired uncertainty estimates. This isolates re-fitting from the metric and aggregation changes present in the Qwen writer comparison (§6).

Distinguish recoverability from model use. A high re-fit score after the frozen score collapses shows that the information remains accessible to a new linear readout. It does not show that the model's unchanged downstream computation can use the altered representation as before: the probe has been retrained, but the model has not. Conversely, a low re-fit score describes the tested probe's performance on the available data, not information-theoretic erasure. Recoverability therefore does not establish behavioral harmlessness.

Give behavior its own controlled comparison. Match support and realized residual displacement, then estimate the paired treatment–control contrast. A control can preserve a selected decoder's output and still affect behavior, as the full-U audit demonstrates. Re-fit probes in both behavioral arms would distinguish preservation of that scalar readout from preservation of linearly decodable position; this remains unmeasured here. Finally, report the positional profile together with overall task performance. A smaller middle-position dip is not by itself a mitigation gain.

Retaining position information does not by itself mean that a mitigation has failed: the goal is better retrieval across document locations, not necessarily erasure. An edit could change how the model uses position while leaving it linearly recoverable, but task benefits still require a matched comparison. Conversely, a behavioral gain would not establish erasure. The success criterion must match the intervention's goal.

9 CONCLUSION

A position edit can disrupt a chosen coordinate while leaving most position information linearly recoverable. The Formation audit shows this directly through the contrast between frozen and re-fit position R^2 after sink-value blocking; visible-set edits produce larger re-fit losses. Leading-direction removal leaves position recoverable in every base model, while selected-head ablations cause larger losses than matched random-head ablations under both diagnostics in three models. Separately, Llama basis removal flattens the dip beyond an active decoder-null control, establishing a behavioral difference conditioned on the tested basis and decoder. The practical lesson is to test remaining recoverability after a frozen readout fails, and to audit any behavioral consequence with its own matched controls.

AI USE STATEMENT

We used generative AI tools for language editing, manuscript organization, consistency checks, figure-label revisions, and code debugging. We did not use generative AI tools to form research ideas or hypotheses, to design the experiments or the preregistered decision rules, or to run model experiments. AI-assisted revisions did not alter saved experimental outputs; corrections to derived summaries and labels were checked against those outputs.

ETHICS STATEMENT

This work analyzes publicly released language models on public text and involves no human-subject study or collection of private personal data. The interventions studied here could inform position-bias mitigation, but our results limit the conclusions that can be drawn from them. An edit can disrupt a frozen readout while leaving position information recoverable by a new probe. Using only that readout to assess information removal can therefore overstate what the edit removes. We make this limitation explicit to inform downstream use. On research integrity, Appendix B contains the preregistration lineage, the analysis ledger, the amendment log, and the recorded implementation deviations. They document which decision rules were fixed before their runs and which were amended, including amended criteria that our own results did not meet.

REPRODUCIBILITY STATEMENT

The experiments use public data: FineWeb-Edu and MDQA data derived from Natural Questions. For the Llama-Instruct multi-document QA experiments on records 0–499, Appendix Table 7 lists record blocks, sample roles, endpoints, gate order, and dependencies between analyses. Its caption identifies analyses outside the table's scope. Statistical conventions and the $\alpha = 0$ estimate lineage are reported in Appendix Table 8 and the experiment-specific result tables. The anonymous release will include code, frozen configurations, amendment logs, saved result summaries, seed metadata, and an environment-and-compute manifest. No private data are used.

REFERENCES

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 92408–92438, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e6d58fc68c0f3c36ae6e0e64478a69c0-Paper-Conference.pdf.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063. Curran Associates, Inc., 2023. doi: 10.52202/075280-2884. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d066d21c619d0a78c5b557fa3291a8f4-Paper-Conference.pdf.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024. doi: 10.18653/v1/2024.acl-long.263. URL https://aclanthology.org/2024.acl-long.263/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. URL https://doi.org/10.1162/tacl_a_00359.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f1b04face60081b689ba740d39ea8f37-Paper-Conference.pdf.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, 2022. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24892–24928. Curran Associates, Inc., 2023. doi: 10.52202/075280-1082. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4e85362c02172c0c6567ce593122d31c-Paper-Conference.pdf.

Junu Kim, Xiao Liu, Zhenghao Lin, Lei Ji, Yeyun Gong, and Edward Choi. Layernorm induces recency bias in transformer decoders. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28638–28652. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1430. URL https://aclanthology.org/2026.findings-acl.1430/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177. URL https://doi.org/10.1177/1948550617697177.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. Where does absolute position come from in decoder-only transformers? *arXiv preprint arXiv:2606.06160*, 2026. URL https://doi.org/10.48550/arXiv.2606.06160.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419. URL https://doi.org/10.1007/BF01068419.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 91212–91239, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e389b15166cf98966ba058965a8c17e3-Paper-Conference.pdf.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67756–67781. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25ad.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. Mitigate position bias in LLMs via scaling a single hidden states channel. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6092–6111, 2025. doi: 10.18653/v1/2025.findings-acl.316. URL https://aclanthology.org/2025.findings-acl.316/.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 60755–60775. Curran Associates, Inc., 2024. doi: 10.52202/079017-1943. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ffdbbe354893979367f93e2121e37dd-Paper-Conference.pdf.

Chunsheng Zuo, Pavel Guerzhoy, and Michael Guerzhoy. Position information emerges in causal transformers without positional encodings via similarity of nearby embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9418–9430. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.632/.

Prior line of work	Starting finding used here	Qualification in this paper	Additional contribution
Haviv et al. (2022); Kazemnejad et al. (2023)	absolute position can emerge without explicit position embeddings	our RoPE intervention changes cardinality and recency jointly, so it cannot isolate a pure token-count mechanism	cross-model frozen/re-fit tests of mask-associated recoverability
Zuo et al. (2025)	nearby-embedding similarity can yield position information in causal NoPE Transformers	the mechanism is not directly tested by our RoPE interventions or the SmolLM3 maintenance audit	an alternative formation mechanism that limits the conclusions we can draw from the hybrid model
Kim et al. (2026)	causal self-attention combined with LayerNorm can induce recency bias	recency bias is not identical to residual-stream absolute-position decodability	an architectural alternative motivating compound mask/normalization interpretations
Yu et al. (2025)	native hidden-state coordinates vary monotonically with absolute position, and selectively scaling one in last-token attention can alter position-biased retrieval	the coordinate is chosen by task calibration; this does not identify a learned residual subspace or a unique causal representation	frozen/re-fit probes and matched behavioral controls separate coordinate manipulability from representation and use
Ruscio et al. (2026)	attention-logit position leakage has causal-mask and residual-stream components; rank-one position-direction ablation reduces leakage by only 3%	the outcome is attention-logit leakage, not re-fit residual recoverability or downstream behavioral use	explicit frozen/re-fit separation plus fit-stability and matched writer and behavioral audits
concept-erasure work (Ravfogel et al., 2020; Belrose et al., 2023; Elazar et al., 2021)	removing one linear direction can leave a concept decodable. Erasure methods iterate nullspace projection until a fresh classifier cannot recover the attribute, or solve for erasure in closed form with minimal change to the representation. Elazar et al. (2021) pair erasure with a behavioral readout	we remove the *fitted* leading probe direction; we do not test an optimally erasing projection	the frozen/re-fit pair for upstream formation edits, plus a separately controlled behavioral audit of a decoded position coordinate
probing and causal intervention work (Hewitt & Liang, 2019; Belinkov, 2022; Geiger et al., 2021; Vig et al., 2020; Meng et al., 2022)	decodability alone does not establish causal relevance	interpreting residual edits requires potency, support, dose, and placebo checks	a multi-level audit separating Use, Representation, Writing, and Formation

Table 5: Claim-level relation to adjacent work. "Qualification" limits the conclusions drawn from a prior observation; it does not imply that the prior result is incorrect.

Artifact	Current paper level	Frozen
Signature matrix + confirmation rules	Formation (§4)	before any intervention
Writer-sweep prespec	Writing (§6)	before the sweep
Behavioral-use battery, blind specification	Use (§7)	before preliminary validation runs
Behavioral-use battery, pre-run specification (5 amendments)	Use (§7)	before each affected run
Later amendments, itemized in the log below	Formation (§4) and Use (§7)	each before its own run

Table 6: Preregistration lineage. The level column uses the paper's current four-level terminology, while artifact names retain their archived wording. No separate Representation row appears because this table lists frozen preregistration artifacts rather than serving as a complete section inventory. The surprisal arm of §5 and §6 has no frozen artifact and is recorded in the amendment log alone. Audit amendments are commit-pinned in the amendment log and listed under *Amendments* below.

A Claim-Level Relation to Prior Work

Table 5 distinguishes the prior findings we build on from the additional tests and narrower conclusions of this paper.

B Preregistration: Lineage, Decision Rules, Amendments

Lineage. Every confirmatory analysis in this paper was scored against criteria frozen before the corresponding runs (Table 6). The code release will include the original, versioned preregistration documents and their history, linked to specific commits.

Combined confirmation rules (Formation). No single cell of Table 10 is sufficient for confirmation. As preregistered, G3 requires lowest-frequency-band collapse that survives the channel-specificity controls—high-frequency-band null, phase-shuffle null, non-position probes intact—

Analysis family	Records	n_q	Data role and dependence	Endpoint / gate family	Inferential status
Initial Llama MDQA batteries	0–79	80	Original behavioral block; separate transplant, magnitude, and reversal executions share questions	dip; experiment-specific potency/placebo rules	mixed lineage; each table states its frozen or amended rule
Layer audit, stage 1	0–79	80	Reuses the original block; reference rows are integrity checks	dip; Holm family over non-reference cells	committed first stage, not an independent replication
Disjoint reversal and layer audit, stage 2	80–159	80	Disjoint behavioral block; stage 2 is conditionally triggered	dip; disjoint replication and stage-specific Holm families	committed replication / terminal conditional stage
Qwen sequence-wide audit	160–279	120	Fresh Qwen behavioral block	sequential dip, potency, placebo, then downstream gates	halted at the placebo gate
Prefix-null, support, and refit audits	200–279	80	Shared Llama block across three audit families	dip; Holm or nominal families stated per audit	selectivity/robustness evidence, not independent replications
Instrument calibration	280–359	80	Forwarded for support/dose calibration; behavioral outcome not used for selection	intervention calibration and NLL checks	calibration-only
Full-U decoder-null audit	360–439	80	Previously untouched Llama behavioral block	dip; committed Holm $m = 4$	fresh decoder-conditioned contrast
Rank-one leading-direction ablation	360–439	80	Third analysis on this block; reuses the identity baseline and both full-U anchor arms from the decoder-null audit, and adds a new displacement-matched control	dip; archived analysis uses Holm $m = 2$, versus $m = 4$ for the full-U audit	rank contrast within one fixed basis; treatment-minus-control interval includes zero
Greedy-generation validation	360–439	80	Same questions as the full-U audit; new endpoint and prompt-only hook scope	normalized EM primary, token F1 secondary	endpoint transfer, not independent replication
Independent-decoder content audit	440–499	60	Disjoint prompts; no behavioral treatment comparison	frozen decoder under token/content controls	decoder-content validation only

Table 7: Analysis-level data and decision ledger for the Llama-Instruct multi-document QA experiments, records 0–499, with the fresh Qwen audit included for comparison. Other cross-model analyses, the ordinal-retrieval tasks, the translation family, and the FineWeb-Edu probe samples are specified in their own result tables and appendix protocols. The *Inferential status* column states which conclusions each analysis supports; analyses that share a record block test robustness or endpoint transfer. The anonymous release will provide the exact frozen artifacts, earlier amendments, timestamps, and checksums.

$\alpha = 0$ estimate	Records	Experiment family	Baseline / execution convention	Role
-0.052	0–79	transplant battery ablation control	paired baseline; predates the explicit identity arm	original full-sequence control
-0.065	0–79	magnitude-scaling battery	explicit $\alpha = 1$ identity within the scaling family	anchor that fixed the minimum-effect band
-0.082	0–79	initial reversal battery	separately executed paired ablation control	within-family comparison
-0.055	80–159	disjoint reversal replication	disjoint paired control	replication-family control
-0.041	360–439	full-U decoder-null audit	explicit identity on a fresh block	headline selectivity treatment

Table 8: Lineage of the printed Llama $\alpha = 0$ estimates. They use the same operator class but come from different experimental families, executions, record blocks, or baseline conventions, not repeated runs of an identical evaluation. Only the -0.065 scaling-family anchor defines the committed band. Gemma's two printed $\alpha = 0$ estimates are related in the caption of Table 19.

together with apparent-position tracking. G1 requires sink-value collapse that also degrades the re-fit probe, a G3 null under derotation, and true-index tracking. G2 requires scale-invariant probe survival under sink-value blocking together with count tracking under the sink-mass clamp. The placebo-adjusted frozen-readout decomposition is retained only as the ordinal summary in Appendix Figure 4. The controls are a mass-matched placebo, sham subtraction, count-matched scramble, high-frequency-band derotation, phase shuffle, and frozen/re-fit probe pairing. Realized sink mass and attention entropy are logged for each cell. The rules printed here are the operational versions of the frozen predictions. The original curve-shape requirements—a linear ramp for G3, a concave-saturating curve for G1, and a convex $1/t$ shape for G2—proved less decisive. The re-rotation (true-index, I4) requirement for G1 and G3 was not evaluated at scale because perplexity diverged. The reported verdicts therefore use the remaining predeclared requirements. The preregistered secondary range readout, the max–min extent of $C_\beta(t)$ across positions, is archived with the results. The interventions displace this curve rather than shrink it, so frozen-coordinate displacement is the primary readout.

Amendments. We distinguish amendments frozen before the affected runs from three post-hoc additions: the position writer audit's re-fit companion, the band-sensitivity check, and the surprisal arm. The behavioral-use battery went through five logged amendments before its full runs: hi-

Claim family	Status	Evidential boundary
Frozen/re-fit readout pair	committed 0.5 drop threshold; direction of disagreement depends on the threshold	23 of 28 cells agree. All 5 disagreements are frozen-only at the committed threshold; at 0.45, three are frozen-only and two are re-fit-only. Three disagreeing cells keep perplexity within 15% of baseline. All 28 cells use eight evaluation documents per model from a 16-document fit/evaluation grid, with no cell-level uncertainty. The writer comparison below uses different metrics and aggregation schemes.
Representation redundancy and concentration	descriptive; four base models, one removal operator	A probe re-fit after removal of the leading fitted probe direction retains 74–88% of baseline position R^2 in all four models. Concentration holds in Gemma alone, where one dimension carries 54% [52, 55] of the frozen decoder's R^2 against 5–11% elsewhere. The removed direction is the fitted one rather than a least-squares-optimal erasure. The behavioral arms cover ranks 6–11 and the full 32-direction basis; the rank-one behavioral arm is reported separately and is inconclusive (Appendix F.3).
Lowest-frequency RoPE carrier	preregistered band-specific test	Negligible effect in four models; all-plane potency does not establish the realized phase span of the tested band, and Llama's tested band failed its manipulation check. Other low-frequency carriers remain untested.
Mask and cardinality dependence	preregistered compound intervention; amended re-fit readout	Re-fit loss is 0.45–0.62 across all four models, exceeding the committed 0.5 threshold in three. Three pass the content check; Gemma is content-confounded. The original rule applies the 0.5 threshold to the frozen-probe drop, which measures readout displacement. Qwen's clamp and position-0 cells fall below it, at 0.358 and 0.381, while their re-fit drops are 0.453 and 0.476. Thus, the four-model result describes a re-fit magnitude range, not four threshold passes. The clamp changes cardinality, recency, and visible content jointly.
Writer-set comparison (position)	preregistered + held-out audit	Ablating the preselected sets causes larger losses than ablating sets matched on layer and head count in three models; output norm and displacement are unmatched. Qwen does not reject the frozen-readout null.
Qwen position re-fit writer effect	exploratory reanalysis	Separate Holm-corrected family on archived arrays; re-fit-readout separation under the matched null, not a retroactive confirmatory verdict. Frozen-ρ and re-fit-R^2 differ in metric and aggregation, so their disagreement does not isolate re-fitting.
Surprisal-channel control	post-hoc; two of four levels	A re-fit probe retains 82–95% of baseline surprisal R^2 after removal. Re-ranked head sets reject the frozen-ρ exclusion null in Pythia and Gemma at Holm-$p = 0.013$; the corresponding values are 0.286 for Llama and 0.336 for Qwen. Llama's re-fit effect differs from matched-set effects at Holm-$p = 0.020$ in the exploratory analysis. This arm was specified after the position results were known and does not test Formation or Use. It uses 300-draw nulls, limiting p-value resolution, and has no next-token-loss check. Nulls match head count, not output dose. The probe peaks 10–15 layers deeper, confounding target with depth (Appendix D).

Table 9: Full claim status and evidential boundaries (first part). The continuation covers the remaining audits, including behavioral use. Historical criterion labels in Table 14 are retained verbatim but do not imply behavioral use.

erarchical uncertainty for mixture shares; a donor-selectivity readout added after a validation run exposed a placebo-confounded slope; outcome controls and a same-span full-residual-swap denominator; validity-checked pair selection; and scope matching after the span-scope prediction failure. Further amendments added the symmetric content check of §4 and the application of the preregistered 0.5 drop threshold to $\mathrm{drop}_{\mathrm{refit}}$ as well as $\mathrm{drop}_{\mathrm{frozen}}$; that threshold amendment was frozen before the Formation re-scoring and motivated by Qwen's clamp and position-0 cells (§4.4). The amendment list also covers the full-sequence value reversal of §7, a displacement-matched placebo re-run, operator-potency and content-baseline diagnostics, a common-context mixture re-run with

Claim family	Status	Evidential boundary
Span-restricted use	mixed	The initial QA transplant has a scope-mismatched potency control; ordinal full-residual-swap comparisons and span-scope scaling provide the strongest evidence.
Position-subspace stability	refit-stability audit; non-unique and fit-sensitive	The fixed Llama U decodes position out of sample ($R^2 = 0.86$–0.88), but independently refit subspaces have 0.17–0.31 affinity, minimum principal angles of 29°–42°, and regularization-sensitive ranks.
Support configuration	committed global-dose-matched audit	Six contiguous-minus-dispersed contrast intervals exclude zero nominally. Total global displacement does not explain the difference, but calibrated α and maximum per-token displacement differ, and placebo equivalence is not established. A geometric mechanism remains untested.
Sequence-wide magnitude intervention	decoder-conditioned differential; minimum-effect rule missed	On fresh records, magnitude removal changes dip by -0.041 $[-0.066, -0.019]$, versus -0.021 $[-0.040, -0.002]$ for the decoder-null perturbation; the paired difference is -0.020 $[-0.031, -0.011]$ (Holm $m = 4$). The control interval lies inside the committed band, but the control still changes behavior. Across five coverage-selected prefixes, effects are negative in $5/5$, complement-control contrast intervals exclude zero in $4/5$, and no 95% treatment interval lies entirely below -0.065. The decoder-null control lies inside U, every column of which is fitted to decode position. We therefore cannot rule out perturbation of redundant position information by the control (§8).
Greedy-generation accuracy validation	secondary; prompt-only, same sample	On the same Llama block, a normalized-EM dip is present; magnitude removal changes it by -0.050 $[-0.100, -0.006]$ and exceeds the decoder-null arm by -0.025 $[-0.050, -0.006]$. The paired token-F1 contrast has the same direction, but its interval includes zero.
Instruct decoder content robustness	disjoint frozen-decoder audit	On held-out Llama-Instruct MDQA prompts, random-token and order-shuffle controls retain the readout, while token-only R^2 is near zero. This supports the tested decoder's absolute-position interpretation under these controls; other position information remains untested.
Sequence-wide value reversal	directional, dependent on sample and U	Prefix-null and complement controls pass their null rules and checks, but the disjoint 95% treatment interval does not lie entirely below -0.065. The same is true for all five corrected prefix replays.
Qwen sequence-wide behavioral audit	sequential audit; halted	A fresh $n_q = 120$ sample passes dip and $\alpha = 0$ potency gates, but the norm-matched complement placebo fails the committed 95%-CI containment gate. Downstream arms are not interpreted.
Readout maintenance through NoPE layers	descriptive; one model, one layer profile	In SmolLM3, position remains decodable through most interleaved NoPE layers, with no systematic drop at the boundaries (Appendix Figure 5). Eight of the nine boundaries change R^2 by at most 0.031; the final NoPE layer instead has $R^2 = -0.163$. Where decodability first arises is untested.
Residual translation	committed amendment	The decoded-coordinate manipulation check passes, but the answer profile does not translate under fixed physical indices.

Table 9: Full claim status and evidential boundaries (continued). The k_{85}-prefix and fresh 32-coordinate decoder-null audits are in Appendix F.2; support–dose and fit-stability analyses appear in Appendices F.4 and F.5. Historical criterion labels in Table 14 are retained verbatim but do not imply behavioral use.

independent document samples, and a paired value–magnitude comparison. Later additions covered per-arm displacement accounting, a sample-size audit of the Qwen dip, base–instruct channel alignment (§3), the disjoint-sample reversal replication, and the translation family. The replication effect has the same sign and half the magnitude; we give both results equal prominence (§7). A pilot at n_q=30 was inconclusive. The final revision also includes five later selectivity and robustness analyses: a k_{85}-prefix decoder-null control, its fresh extension over all 32 U coordinates, a global-dose-matched support grid, a refit/coverage stability family, and a sequential Qwen use au-

Intervention (archived frozen-readout prediction)	G1: sink value	G2: mask/count	G3: low-frequency phase
I1 sink-value block	collapse	amplitude only	∼invariant
I2a visible-window clamp	∼invariant	tracks visible-set size	∼invariant
I2b position-0-key removal	collapse	partial	∼invariant
I3 lowest-frequency-band derotation	∼invariant	∼invariant	collapse
I4 query-only re-rotation	true index	true index	apparent position

Table 10: Original preregistered signature matrix. These are the archived predictions, not the final causal labels. In particular, the visible-window operator also changes recency topology, and the re-rotation results were not interpretable at scale because perplexity diverged. The main text therefore uses the narrower mask and cardinality interpretation and the operational decision rules below.

dit. Their samples, grids, and decision gates were fixed before the corresponding behavioral runs. Two further validations test greedy-generation accuracy on the full-U audit block and the frozen independent decoder under content controls on a disjoint MDQA block. The layer-contingency and position writer audits were likewise frozen before their runs with outcome labels fixed in advance. The layer-contingency design was fixed before calibration; numeric thresholds were then fixed after calibration but before behavioral data were collected. The position writer audit re-evaluates the preselected top-fifteen head identities on held-out documents with thousand-draw nulls and a re-fit readout (§6, Appendix E). A post-hoc companion to the position writer audit (offline, no new runs) applied the same re-fit-damage test to the already logged arrays for all four models as a separate exploratory family with its own Holm correction; it is labeled exploratory wherever it appears and changes no preregistered verdict (§6). The surprisal arm of §5 and §6 is also post-hoc, but it differs from that companion analysis. The companion re-scores arrays already logged under a frozen protocol; the surprisal arm fits a new channel and runs fresh forward passes for probe fitting, removal, and the writer audit. It appears in no frozen preregistration artifact and in no row of Table 6; §5, §6, Table 9, and Appendix D carry the post-hoc label.

Layer-audit staging. Stage 1 triggered the preregistered second stage on the two reversal cells: the anchor signature persisted at every adjacent layer and operator, ablation was supported at both adjacent layers, and the matched control also changed behavior in the reversal arms (Table 29). The frozen design left one convention open, and we fixed it between the two stages: a support verdict requires unanimity across a cell's scored persistence readouts, the escalation trigger accepts any single persisting readout, and mixed cells are reported with their full readout vector. Fixing it could not change stage 1, where every readout was uniformly persistent. Stage 2 used the disjoint sample, with interpretation branches fixed before the run. The control effect reproduced at one adjacent layer but not the other, and no non-reference layer showed a value effect distinguishable from the placebo. The ablation effect below the reference layer replicated; the effect above it retained its direction but did not meet the support rule. The design ends after two stages, so cells that would otherwise trigger another stage are reported as informational (Table 31). In Stage 2, these are the $l = 3$ cells. The reported 95% intervals are two-sided, whereas the Holm-adjusted tests use the preregistered one-sided alternative; the two summaries need not agree. Pooled values across the two stages are descriptive only under the frozen sequential rule and are archived with the results.

Implementation deviations. Two implementation errors affected runs, but neither affected a reported label. The first stability replay mistakenly passed the full 32-column basis to every nominal rank cell. Before further forward passes, an implementation audit exposed the scope of the error, and the final verdict uses only the corrected rank-truncated replay in Appendix F.5. The full-32 output remains archived as an implementation-deviation record and is not used in any manuscript claim. The first aggregation implementation passed a summary status to the support gate and the escalation trigger, which both require the cell-level instrument status. This prevented both decisions from being triggered. We caught the error before reporting any labels, corrected the implementation, and verified the fix against the frozen rules with regression tests.

Misses. One preregistered prediction failed, and we revised two earlier interpretations. The failed prediction is the span-scope dose–response, which the main text reports (§7). The earlier manuscript-level conclusion that the span-restricted value edit had no behavioral effect—not a frozen prediction—was revised after a test frozen before execution produced a directional effect at matched full-sequence scope (§7). We had earlier read Yu et al. (2025)'s published scaling grid as amplifying the position coordinate. It does not amplify: their searched factors are $\{0.5, 0, -0.5, -1\}$,

which attenuate, zero, and reverse the channel, and their per-model selections never exceed 0.2. Their operator scales the raw channel, whereas ours scales the centered U coordinate, and our grid $\alpha \in \{0, 0.5, 1, 2, 4\}$ meets theirs only at 0 and 0.5 while extending to amplification rather than sign reversal. The two dose responses are therefore not directly comparable, and we make no claim that ours reproduces theirs.

C Operator Definitions from the Implementation

Statistical conventions. Intervals use percentile bootstraps at the sampling unit specified for each analysis. Confirmatory Use contrasts draw $20{,}000$ question-level resamples, paired by question. Monte Carlo tail probabilities use the add-one correction $p_{\mathrm{MC}} = (k+1)/(B+1)$ for k null-tail draws in B resamples; a zero count is reported at resolution $1/(B+1)$, never as $p = 0$. The initial behavioral batteries used $2{,}000$ resamples. Each affected table states which intervals we recomputed at $20{,}000$ and which retain the legacy value. Equivalence gates use two one-sided tests (Schuirmann, 1987; Lakens, 2017); multiplicity corrections follow the analysis-specific Holm families (Holm, 1979).

Every intervention uses a forward hook in the same TransformerLens implementation, so the probe and the edit share a residual basis. Each definition below identifies the hook point, allowing the equation to be checked against the implementation. We use the following notation throughout.

- $h_l(t)$, a d-vector: the post-block residual stream (`hook_resid_post`) at layer l, position t; d is the model width. L^* is the run-specific frozen intervention layer; (β, μ) the frozen coordinate's direction and mean, with $C_\beta(t)$ as in Eq. 1.
- For attention at layer l, head i: $S^{l,i}_{q,k}$ the pre-softmax score (`hook_attn_scores`), $A^{l,i}_{q,k}$ the post-softmax pattern (`hook_pattern`), $v^{l,i}_k$ the value, and the head output $z^{l,i}_q = \sum_k A^{l,i}_{q,k}\, v^{l,i}_k$ (`hook_z`).
- Post-rotary query/key vectors (`hook_rot_q`, `hook_rot_k`) carry $n = \texttt{rotary_dim}/2$ rotation planes; plane j occupies the split-half column pair $(a_j, b_j) = (j,\ j+n)$; columns $\geq$ `rotary_dim` are pass-through and never touched. θ_j is the model's own per-plane angular frequency, read from its rotary buffer (so any NTK-by-parts rescale is inherited, not recomputed). At position t, the model has already applied $R(t\theta_j)$.
- *Writer region*: unless stated, every edit is applied at all layers $l \in \{0, \ldots, L^*\}$, all query positions, and all heads.

Formation evaluation protocol. The 16-document grid uses the first eight FineWeb-Edu windows for fitting and the last eight for evaluation, at $2{,}048$ tokens per window (§4). The frozen readout retains its original coefficients and intercept. Each re-fit probe uses that condition's edited fitting activations with ridge penalty $\lambda_{\mathrm{r}} = 10^{-2}$; its baseline is fitted on unedited activations from the same fitting partition. Both probes' R^2 values are pooled over the same held-out document–token pairs, and perplexity is pooled over those evaluation documents. The archived grid does not record the exact document identities or stream offset, and the retained summaries do not support cell-level intervals.

A rotation of a plane pair (a, b) by angle ϕ is

$$R(\phi): \quad x_a \mapsto x_a \cos\phi - x_b \sin\phi, \qquad x_b \mapsto x_b \cos\phi + x_a \sin\phi. \tag{3}$$

I3—Lowest/high-frequency-band derotation

Edit. On the post-rotary query *and* key (`hook_rot_q`, `hook_rot_k`), for every plane j in a band $\mathcal{B}$, apply the exact inverse rotation $R(-t\theta_j)$ of the position-t rotation, using $c_j = \cos(t\theta_j)$, $s_j = \sin(t\theta_j)$:

$$x'_{a_j} = x_{a_j} c_j + x_{b_j} s_j, \qquad x'_{b_j} = x_{b_j} c_j - x_{a_j} s_j, \qquad j \in \mathcal{B}. \tag{4}$$

This returns the plane to its pre-rotary value, undoing the band's explicit RoPE rotation. $\mathcal{B} = \mathcal{B}_{\text{low}}$ is the lowest-frequency band; the negative control is $\mathcal{B}_{\text{high}}$, the $|\mathcal{B}_{\text{low}}|$ highest-frequency planes (plane-count matched).

Held fixed / free. Planes $j \notin \mathcal{B}$ keep their RoPE phase, and other vector components are unchanged at the hook. The softmax, grouped-query-attention (GQA) key repeat, and downstream computations are not directly edited. Eq. 4 undoes only the explicit rotation in band $\mathcal{B}$; position information already present in the pre-rotary vectors can remain.

Manipulation check. The round trip (derotate then re-rotate = identity) is unit-tested independently of the model against a NumPy reference. The band comparison includes two controls: high-frequency-band derotation and an energy-matched phase-shuffle content control. The latter uses the same plane pair, but replaces $-t\theta_j$ with the extra angle $\theta_j\,(\pi(t) - t)$ for a fixed position permutation π, disrupting position-monotone phase while preserving band energy. The all-plane derotation control verifies that broad derotation can affect the readout; it does not establish the potency of the tested lowest-frequency band (§4). The tested band rotates by only 0.0006 rad in Llama, failing its manipulation check and leaving that null uninterpretable. Its realized rotation spans 0.002–0.257 rad across the other three models, so their nulls apply to this band rather than to low-frequency carriers generally.

I2—VISIBLE-WINDOW CLAMP (AVERAGE SINK MASS FIXED)

This operator combines two hooks applied in sequence at layers $0..L^*$.

Edit 1 (window mask, pre-softmax). On $S^{l,i}$, block every causal key outside the trailing window of width w, but keep position 0:

$$S'_{q,k} = \begin{cases} -10^9 & k \le q,\ \ \neg(q - w + 1 \le k \le q),\ \ k \neq 0 \\ S_{q,k} & \text{otherwise,} \end{cases} \tag{5}$$

after which the model's own softmax renormalizes over $\{0\} \cup [q - w + 1, q]$.

Edit 2 (P0 clamp, post-softmax). Narrowing w increases sink mass through renormalization. To counter this, a second hook on $A^{l,i}$ fixes the sink weight at a constant τ and rescales the remaining weights:

$$A'_{q,0} = \tau, \qquad A'_{q,k} = A_{q,k}\,\frac{1 - \tau}{\sum_{k' \ge 1} A_{q,k'}}\ \ (k \ge 1). \tag{6}$$

Here τ is a *single scalar*: the baseline full-context realized P0 weight, averaged over all heads, query positions, and layers $\le L^*$.

Held fixed / free. The sink's post-softmax mass is held at τ regardless of w, position, or head; the competing keys are those in the trailing window. We sweep w over $\{256, 512, 1024\}$. Every cell in Tables 2 and 14 uses $w = 512$, fixed before the sweep rather than selected from its results. The preregistered count-matched scramble uses the same width (Appendix E). The number of readout agreements in §4.4 depends on w. At the committed threshold, $w = 256$ gives 25 agreements and three disagreements; $w = 1024$ gives 23 and five, with disagreement in Gemma's clamp cell instead of Pythia's position-0 cell. All disagreements are frozen-only at all three widths. Because the set is a trailing window, competitor identity and recency topology vary with w. The operator therefore reduces sink-mass variation while jointly changing visible-set cardinality and recency topology; it does not isolate token count. Because τ is a global mean, the clamp imposes a stronger constraint than holding each configuration at its own full-context value. There is also a boundary case: at $q{=}0$, the sink is the only causal key, so Eq. 6 leaves the row summing to $\tau < 1$. Every row with $q \ge 1$ retains its own key and sums to 1. The realized P0 weight and attention entropy are logged per cell.

I1—SINK-VALUE BLOCKING (HEAD-SELECTIVE)

Edit. On the post-softmax pattern $A^{l,i}$ for a head set H (all heads, or a selected subset), zero the position-0 column *without renormalizing*:

$$A'^{\,l,i}_{q,0} = 0 \quad (i \in H), \qquad z^{l,i}_q = \sum_{k \ge 1} A^{l,i}_{q,k}\, v^{l,i}_k. \tag{7}$$

Only the $v^{l,i}_0$ contribution is removed from the head output; every remaining weight is unchanged, so the row sums to $1 - A_{q,0}$. A robustness variant renormalizes the remaining weights to sum 1. It leaves Pythia and Qwen near null on both readouts and reproduces Llama's dissociation. Pythia's

frozen and re-fit drops are 0.010 and -0.005, respectively. Qwen's corresponding drops are -0.014 and -0.007; Llama's are 0.276 and 0.043. In these three models, the absolute relative change in perplexity is at most 0.044. Gemma is content-confounded: its relative perplexity change is $+0.876$, above the 0.15 limit. Its frozen drop is 224.770, whereas its re-fit drop is 0.204 (27% of its 0.763 re-fit baseline). We therefore exclude this variant's Gemma cell from formation claims. The headline sink-value cells use the non-renormalizing edit of Eq. 7, and the 28-cell count excludes this arm.

Held fixed / free. All non-sink attention weights, the values, RoPE, and positions are unchanged; only the value copied from the sink is removed from the residual write.

Manipulation / specificity check. The mass-matched placebo zeros, per query and head, the *non*-sink key whose weight is closest to the sink's, removing approximately the same mass. Both frozen and re-fit scores are recorded to distinguish readout disruption from loss of linear recoverability.

EXCL-P0—POSITION-0-KEY REMOVAL

Edit. The window mask of Eq. 5 with $k{=}0$ *not* exempted: every causal key outside the trailing window is blocked, *including position 0*, and the softmax renormalizes over the window alone:

$$S'_{q,k} = -10^9 \;\text{ for }\; k \le q,\; \neg(q - w + 1 \le k \le q). \tag{8}$$

This removes the sink from the attention set once position 0 falls outside the trailing window ($q \ge w$); earlier queries still retain it. Unlike I1, the edit changes the available keys rather than only deleting the sink's value contribution post hoc (Eq. 7).

Held fixed / free. The trailing-window competitors are fixed; position 0 has no exemption from the window mask. The archived count-matched control draws a random causal subset of size w per query and then retains P0. It matches nominal window width, not exact realized cardinality: after the initial prefix, it retains w keys if P0 was already drawn and $w + 1$ otherwise. This control was intended to separate the number of visible tokens from recency and degradation. It exceeds the content-damage threshold in three models and is not used to support a causal claim (Appendix E).

I4—RE-ROTATION FAMILY T_δ (QUERY-ONLY)

Edit. On the post-rotary query *only* (`hook_rot_q`), for a token subset $\mathcal{T}$, a head set H, and planes $\mathcal{B}$ (each all-inclusive if unset), apply an extra forward rotation $R(\delta\theta_j)$ of Eq. 3:

$$\begin{aligned} x'_{a_j} &= x_{a_j}\cos(\delta\theta_j) - x_{b_j}\sin(\delta\theta_j), \\ x'_{b_j} &= x_{b_j}\cos(\delta\theta_j) + x_{a_j}\sin(\delta\theta_j), \qquad t \in \mathcal{T},\; i \in H,\; j \in \mathcal{B}. \end{aligned} \tag{9}$$

Non-selected heads are restored to their unedited value.

Why query-only. Rotating query and key symmetrically on the same subset preserves the intra-subset relative phase, $(m{+}\delta) - (n{+}\delta) = m - n$, and produces no shift, which could falsely suggest a G3 null. Keeping the key at its true phase leaves P0 as an unchanged reference. $\delta = \pm 128$ is the graded in-range arm; $\delta = \pm 512$ is an out-of-range robustness arm.

Held fixed / free. True token index, order, causal mask, and the KV cache are untouched; keys stay at true phase. Only the query's *apparent* phase moves, allowing us to test whether the channel follows apparent position ($\approx \delta$ shift, G3) or the true index (no shift, G1/G2).

Manipulation check. The realized channel-mean shift is compared to the shift predicted from the baseline coordinate's rate of change per position (follow-ratio ≈ 1 implies apparent-position; ≈ 0 implies true-index); the symmetric query-and-key arm is the artifact control (must show no shift). **Caveat:** this family was not interpretable at scale because perplexity diverged; its signature was not used in any confirmed verdict (Appendix B).

POSITION-SUBSPACE VALUE TRANSPLANT (SPAN-SCOPE)

Multi-document QA slots. Each 20-document instance places the gold document at one of ten slots, $p \in \{0, 2, 4, 6, 8, 11, 13, 15, 17, 19\}$. The end slots and middle pair define the dip in Eq. 2; all ten slots enter the profile and transplant-slope readouts.

Subspace. $U = [u_1, \ldots, u_K]$, a $d \times K$ matrix, is a greedy orthonormal position subspace fit at L^*: each u_m is the best ridge position-decoder in the complement of $\{u_{1..m-1}\}$. Rank k_{85} is the smallest k for which projecting out $U_{:k}$ removes $\geq 85\%$ of the position R^2 recoverable from the complement.

Edit. On `hook_resid_post` at L^*, over the gold span $[g_0, g_1)$, add a fixed displacement projected into the subspace:

$$h'(t) = h(t) + \delta, \qquad \delta = U_k U_k^\top (r_D - r_R), \quad k = k_{85},\ t \in [g_0, g_1), \tag{10}$$

where r_R is the span-mean residual of the gold at its true slot and r_D is the span-mean residual at the donor slot, both measured in the unedited forward pass. This shifts the span-mean k-dimensional position coordinate to the donor's value while preserving within-span differences.

Slope readout. The gold stays physically at a fixed anchor slot while the donor slot sweeps all ten positions. Write $\bar{\ell}(s) = \frac{1}{n_q} \sum_q \ell_{q,s}$ for the mean answer log-probability with the gold physically at slot s, and $\bar{\ell}_{\mathrm{tp}}(s)$ for the same quantity with the gold left at the anchor and its position coordinate transplanted to slot s. The *transplant slope* is the ordinary-least-squares slope of $\bar{\ell}_{\mathrm{tp}}$ on $\bar{\ell}$ across the ten slots, bootstrapped over questions. A slope of 1 is consistent with tracking the baseline profile, but does not establish that the profiles match. A slope of 0 indicates no linear tracking, not necessarily unchanged behavior. An interval contained in ± 0.2 scores the cell null. Two anchor slots were run per model, and Table 19 reports the deeper one.

Held fixed / free. $\delta \in \mathrm{span}(U_k)$, which leaves the $d - k$ orthogonal complement and all positions outside $[g_0, g_1)$ untouched. The operator is **not norm-preserving**: $\|h'(t)\| \neq \|h(t)\|$ in general (additive projected displacement).

Controls. *cross*: donor drawn from a different question's gold (position, not content). *placebo*: a same-norm displacement in the complement of U (norm-matched to $\|\delta\|$; must not move behavior). *coverage*: repeat at k_{90}, k_{95}. *ablate* (positive control, Eq. 13) must flatten the dip.

Manipulation check. An independent decoder $\hat{q}(r) = \mathrm{ic} + w^\top (U_{:k}^\top r)$, fit on the *test* split, must match the donor value within a fixed tolerance: $|\hat{q}(r_{\mathrm{post}}) - \hat{q}_{\mathrm{target}}| < \max(0.5,\ 0.1\,|\Delta\hat{q}|)$. The code enforces this as a hard assertion.

Marker-free ordinal retrieval and full-residual-swap reference

Task construction and readout. Each model's tokenizer filters a shared word pool to items that are exactly one token when preceded by a space. Each instance contains one gold item, one reference competitor, and randomized fillers. The reference item and its physical slot remain fixed while the other items are randomly assigned to slots, separating item identity from position. Items are separated only by newlines, without numbered markers. The word-list query is "Word number q is:"; the sentence-preamble query is "Sentence number q is:", including the same template for the first slot. The word-list task uses zero-based slot 8 (the ninth item) as its fixed anchor, and the sentence-preamble task uses slot 0. For gold item g and fixed reference r, the teacher-forced readout is

$$M_q = \log P(g \mid q) - \log P(r \mid q). \tag{11}$$

Sweeping the gold across physical slot s and querying ordinal $s + 1$ gives the baseline position profile. Because the reference text and slot do not move, variation in this profile comes from gold-item retrieval rather than a moving competitor.

Full-residual-swap reference. For a preselected slot pair (p, q), the reference intervention exchanges the two items' complete d-dimensional residual vectors at L^* in both directions. The input tokens, causal mask, and downstream RoPE phases remain unchanged. This reference measures the effect reachable by an unrestricted residual edit at the same one-token spans. Let $\Delta\ell_r$ be the intervention-induced change in the teacher-forced log-probability of item r under the target query. Donor selectivity is

$$S_{p \to q} = \Delta\ell_p - \frac{1}{N - 2} \sum_{r \notin \{p, q\}} \Delta\ell_r, \tag{12}$$

excluding the target-slot item q from the distractor average. The reported fraction is the position-subspace transplant's mean S divided by the full-residual-swap mean S, with numerator and denominator evaluated on the same bootstrap resamples.

Validity and verdict rules. The baseline location effect must exceed a position-label permutation null ($p < 0.05$ and position variance at least 0.02). Full-U ablation must reduce profile spread, whereas a same-rank complement ablation must not. Physical input swap and full-residual swap must each produce a positive selectivity interval, and the complement placebo must remain within 0.2 times the full-residual-swap effect. A separate different-word donor control rejects a condition when the nominal matched transplant exceeds that cross-content arm, indicating content leakage. Conditional on these gates, the verdict is positive when the transplant/full-residual ratio interval lies above 0.2 and both the predicted donor-item rise and target-item drop are present. An interval with a positive lower bound below 0.2 is labeled weak/partial positive. An interval contained in $[-0.2, +0.2]$ with neither item-level signature is equivalent to null, and other cases are inconclusive.

POSITIVE CONTROL / α=0 ANCHOR—U-ABLATION

Edit. On `hook_resid_post` at L^*, over all positions, set the U-coordinate to its per-sequence mean:

$$h'(t) = h(t) - \big(z_t - \bar{z}\big)U^\top, \quad z_t = (h(t) - \mu)\,U, \quad \bar{z} = \tfrac{1}{S}\textstyle\sum_t z_t. \tag{13}$$

This removes variation across positions in $\mathrm{span}(U)$ while preserving the $d - K$ complement and the per-sequence mean. It is the behavioral potency control: if it does not flatten the dip, this intervention site lacks a demonstrated effect on the endpoint and the associated null comparison is not interpreted.

α-SCALING OF POSITION-SUBSPACE MAGNITUDE (FULL-SEQUENCE)

Edit. On `hook_resid_post` at L^*, over *all* positions, scale the U-coordinate by α:

$$h'(t) = h(t) + (\alpha - 1)\,\big(z_t - \bar{z}\big)U^\top, \quad z_t = (h(t) - \mu)\,U, \;\; \bar{z} = \tfrac{1}{S}\textstyle\sum_t z_t. \tag{14}$$

$\alpha = 1$ is the identity; $\alpha = 0$ reduces *exactly* to the U-ablation of Eq. 13 (the potency anchor). The original placebo repeats Eq. 14 with U_c, a random orthonormal basis in the complement of U of the same rank.

Held fixed / free. Rank and α are matched between treatment and placebo. In the original run, the displacement magnitude was **not** held fixed, because the residual's coordinate variance differs between U and U_c; that placebo is matched on rank and α only. The audit re-run (§7) addresses this mismatch by rescaling the complement displacement at each position to the treatment's realized $\|\Delta h\|$ (measured ratio 1.00).

Slope readout. The dose-response slope is the ordinary-least-squares slope of ΔD (Eq. 2) on α over the retained grid points, bootstrapped over questions. A positive slope means a larger position-coordinate magnitude deepens the dip, and shrinking the coordinate flattens it.

Manipulation and anchor checks. An answer negative-log-likelihood (NLL) ratio above 1.5 flags an α as out of distribution and excludes it from the slope. The α=0 potency anchor passes only if its 95% interval is negative and its mean lies within ± 0.065 of the archived anchor value -0.065; this reproduces the anchor and is distinct from the minimum-effect rule.

GLOBAL-DOSE-MATCHED SUPPORT CONFIGURATION

Edit and calibration. The masked operator applies Eq. 14 only at a fixed set of positions. Support fractions 0.25 and 0.5 are instantiated as either one contiguous block or a dispersed mask with the same cardinality; the gold span and full support are additional references. For each cell, α is calibrated and then frozen to realize one of three common full-residual Δh RMS targets, $\{0.5, 1.0, 1.5\}$.

Held fixed / free. The calibration matches global residual RMS across masks, not the local per-edited-token displacement distribution. The realized maximum per-token $\|\Delta h\|$ values range from 1.11 to 42.51, and the frozen α values range from -0.44 to -6.19 across the analysis grid. α is the direct multiplier in Eq. 14: these negative values extrapolate beyond the $\alpha = 0$ ablation and reverse the centered U coordinate within the edited mask. Thus, the analysis matches global intervention strength but allows local dose to vary.

Controls. The complement-subspace placebo uses the same positional mask and matched global residual RMS. All 18 treatment cells satisfy the 1.5 answer-NLL-ratio threshold. Paired question-level contrasts use 20,000 bootstrap resamples on the shared records 200–279; the six geometry intervals are nominal, without familywise correction.

NORM-PRESERVING VALUE REVERSAL (FULL-SEQUENCE)

Edit. On `hook_resid_post` at L^*, over all positions $t \in \{0, \ldots, T-1\}$, reverse the U-coordinate profile along the context while preserving each position's magnitude:

$$z'(t) = z(T-1-t) \cdot \frac{\|z(t)\|}{\|z(T-1-t)\|}, \qquad z(t) = (h(t) - \mu)\, U, \tag{15}$$

with the complement of U untouched.

Held fixed / free. The per-position magnitude profile $\|z(t)\|$ is held exactly fixed; only the value (direction and ordering) of the position code moves. Within this operator, a behavioral change cannot be explained by a change in the per-position U-magnitude profile. Attribution to the reordered U coordinate additionally requires the matched controls described below.

Controls. The α=0 ablation (Eq. 13) is the scope-matched positive control; the placebo applies the same reversal construction in a complement subspace rescaled to the treatment's realized per-position $\|\Delta h\|$. The operator and decision rule were frozen before the run; verdicts appear in §7.

SHIFT FAMILY—NORM-PRESERVING RESIDUAL-COORDINATE TRANSLATION (FULL-SEQUENCE)

Edit. At L^*, let $z(t) = (h(t) - \mu)U$. For $\Delta \in \{\pm T/8, \pm T/4\}$ and positions whose shifted index lies in range, set

$$z'(t) = z(t + \Delta)\frac{\|z(t)\|}{\|z(t + \Delta)\|}, \tag{16}$$

with boundary values clamped and behavioral readout restricted to the valid window. The complement of U is unchanged.

Interpretation and controls. The edit translates the decoded residual coordinate while preserving each position's U-magnitude profile. It does not alter physical token indices, the causal mask, or downstream RoPE phases. An independent decoder verifies the realized shift; a complement-subspace placebo is matched to the treatment's realized per-position $\|\Delta h\|$, and next-token loss must remain within its committed threshold. A behavioral null limits the sufficiency claim: translating this coordinate alone does not shift behavior when physical positions remain fixed.

LAYER-MATCHED RANDOM-HEAD CONTROL

Writer ranking. Each head's channel contribution is $c^{l,i}(t) = \beta^\top\big(z_t^{l,i} W_O^{l,i}\big)$ (for sandwich-norm models the post-attention RMS scale and gain are folded in exactly). Heads are ranked by $|\mathrm{cov}_t(c^{l,i}, t)|$ on the fit documents; a per-layer self-check confirms $\sum_i c^{l,i}$ + bias reconstructs the attention contribution to $\approx 10^{-6}$.

Ablation operator. On `hook_z`, zero the writer heads: $z_q'^{l,i} = 0$ for $i \in H_l$.

Matched control. Draw $N_{\text{rep}} = 20$ random head sets with the *same per-layer counts* as the top-k set, sampling heads without replacement within each set from the pool

$$\mathcal{P} = \{(l, i) : l \le L^*\} \setminus \{\text{tested top-}k\}, \tag{17}$$

and ablate each identically. At the single preregistered k=15, writing is *localized* if and only if three conditions hold: the top-k set's surviving monotonicity ρ is below that of *every* valid random draw, all 20 draws are feasible, and the empirical $p = (\#\{\rho_{\text{rand}} \le \rho_{\text{top}}\} + 1)/21$ is at most 0.05. The pool excludes only the tested top-k. This null therefore tests whether the selected set has a larger effect than its non-top alternatives—the comparison reported in the main text. The held-out audit below adds an inclusion null, which asks how the selected set compares with a typical eligible set.

HELD-OUT POSITION WRITER-HEAD AUDIT—PROTOCOL

Frozen inputs. The audit evaluates the preselected top-15 heads without re-ranking. Each model's (l, i) list is loaded unchanged from the archived sweep artifact. The frozen probe parameters (β, μ, L^*, intercept) are also loaded unchanged; no new construct is fit. Heads were ranked on the sweep's fit documents and are evaluated on a disjoint sample, so no head is ranked and scored on the same data.

Evaluation data. The evaluation data are FineWeb-Edu (train split, streamed in dataset order): 48 single-document windows of 2,048 tokens each, taken at stream offset 500 (documents 500–547). That sample is disjoint from the construct fit at offset 0, from the sweep ranking at offset 200, and from the layer-contingency audit's offline partitions at documents 0–175. The fit-stability splits of Appendix F.5 use stream offsets 200, 400, and 600, but they use Llama-3.1-8B-Instruct at 4,096 tokens. Their document indices and this audit's refer to differently filtered streams, so the indices do not establish text disjointness. The readout context is 2,048 tokens in all four models, matching the archived sweep: the audit varies the document sample alone. For Gemma, the construct-fit context (4,096) differs from the readout context; the other three were fit at 2,048. The archived sweep already operated under that disclosed mismatch.

Nulls. Both nulls draw 1,000 random head sets with the same per-layer counts as the frozen top set from one deterministic stream (seed 0). Heads are sampled without replacement within each set. They match layer composition and head count, not per-head output norm, aggregate output dose, or realized residual displacement. The exclusion null (primary) draws from the pool of Eq. 17, eligible heads at layers $\leq L^*$ minus exactly the fifteen tested heads, and instantiates the hypothesis that the top set is indistinguishable from its non-top alternatives. The inclusion null (robustness) draws from the full eligible pool. Because draws are layer-count-matched, its expected overlap with the top set is $\sum_l \text{count}_l^2 / n_{\text{heads}}$, or $\approx$ 3.5, 2.8, 2.5, and 2.6 heads for Llama, Pythia, Qwen, and Gemma. This overlap makes the inclusion null less distinct from the top set by construction; it compares the top set with a typical eligible set, rather than exclusively with non-top alternatives. Realized mean overlaps are 3.60, 2.80, 2.57, and 2.60, reconstructed by replaying the seeded sampler (per-draw overlaps were not stored) and validated against the stored feasibility and distinctness diagnostics. A draw is infeasible if any layer has fewer eligible heads than its required count. Verdicts require all 1,000 draws to be feasible, and we also check that the draws are distinct. The achieved count sets the p-value resolution; every model had 1,000 feasible, distinct draws. The surprisal writer audit of §6 draws 300 sets per null under this same layer-matched sampler, giving it a resolution of $1/301$ and a smallest attainable Holm-adjusted p of 0.013 at $m = 4$ (Appendix D).

Readouts. Each condition (the top set and every draw) is ablated with the sweep's operator above, and the residual is read at L^*. The runner computes three readouts, retaining per-draw null arrays for two. The primary axis readout is the frozen-ρ monotonicity of the frozen-β decode, computed by correlating the document-averaged curve over all 48 audit documents with token index. The secondary readout is the frozen decoder's R^2, retained for the baseline and top set but not as a per-draw null. The representation readout is a re-fit R^2: a ridge probe fit per condition on that condition's ablated activations, and on un-ablated activations for the baseline, with the shared ridge estimator used by every probe in this paper ($\lambda_r = 10^{-2}$). The re-fit uses a document-level $70/30$ split (34 fit, 14 evaluation, seed 0) held identical across the baseline, the top set, and every draw; all re-fit differences share one split, and no probe is fit and scored on the same documents. Thus the tested frozen and re-fit diagnostics change the metric, aggregation, and evaluation sample as well as the probe. Their disagreement does not isolate re-fitting; a comparison with a common metric and evaluation split is not recoverable from the retained null arrays. The next-token LM-loss ratio on the audit documents is logged as a content-damage check, not a gate.

Statistics and verdicts. Baselines (ρ and re-fit R^2) are measured on the audit sample itself; archived values are carried as reference only. The decision statistic is the damage $E_{\text{top}} = \rho_{\text{base}} - \rho_{\text{top}}$ against the null damages E_r, with one-sided empirical $p = (1 + \#\{E_r \geq E_{\text{top}}\})/(N + 1)$ at $N = 1{,}000$ feasible draws (resolution $1/1{,}001$). Holm correction ($m = 4$) covers this arm's four frozen-ρ exclusion tests only; the inclusion null is reported separately as robustness, and the surprisal writer audit corrects within families of its own (Appendix D). The axis verdict requires E_{top} to exceed the exclusion null's 99th percentile with Holm-adjusted $p < 0.05$. The representation verdict is conditional on an axis pass: re-fit damage must then exceed the 99th percentile of re-fit damage over the same exclusion draws, with no separate Holm correction. The symmetric re-scoring of the

position re-fit family in §6 (its own Holm, $m = 4$) is an exploratory post-hoc reanalysis of these same logged arrays; the surprisal arm of §6 corrects two further $m = 4$ families of its own, computed from separate runs (Appendix D).

D SURPRISAL-CHANNEL CONTROL: PROTOCOL AND DIFFERENCES

The surprisal control applies the Representation and Writing audits to a different ridge-decoder target. Including this target change, three protocol differences separate the representation arm from the position removal run, and fourteen separate the writing arm from the position writer audit. The position construct is a ridge decoder from the post-block residual $h_l(t)$ at layer l to token index t. The surprisal construct is a ridge decoder from the same $h_l(t)$ to the current token's surprisal $y(t) = -\log p(x_t \mid x_{<t})$, taken from a clean forward pass of the unedited model. The fit otherwise follows the position pipeline: FineWeb-Edu single-document windows, the shared ridge estimator ($\lambda_{\mathrm{r}} = 10^{-2}$), and a document-level $70/30$ split at seed 0 (34 fit, 14 evaluation of 48 documents). We sweep all layers and fix L^* at the one with the highest held-out R^2. Surprisal is undefined at the first token. Every surprisal ridge fit and ablation readout below therefore uses positions $1, \ldots, S-1$, whereas the position arm uses all S. The readout-versus-position diagnostic in the next paragraph is the sole exception: the routine shared by both arms computes a document-averaged readout at all S positions. We write (β_s, μ_s, L_s^*) for the frozen surprisal construct and $C_s(t) = (h_{L_s^*}(t) - \mu_s)^\top \beta_s$ for its readout, which is Eq. 1 with the surprisal-fitted construct substituted. The selected layers are $L_s^* = 14$, 12, 27, and 17 for Llama, Pythia, Gemma, and Qwen, respectively. These models have 32, 32, 42, and 28 layers; the selected probes achieve held-out $R^2 = 0.722$, 0.670, 0.762, and 0.705.

The surprisal readout is not a relabeled position readout. The document-averaged C_s trajectories have Spearman correlations of -0.130, -0.098, -0.045, and -0.070 with absolute position, compared with at least $+0.993$ for the position channel in every model. L_s^* is also deeper than the position L^* in all four models (14 versus 1, 12 versus 2, 27 versus 12, 17 versus 3), so target and depth are confounded in every comparison between the two arms. Pythia alone has the same highest-correlating residual dimension, 1180, under both targets. The selected dimension changes from 79 to 1132 in Gemma, 1288 to 1790 in Qwen, and 3405 to 3585 in Llama.

Representation arm—removal operator. Both arms remove the fitted leading direction using $P = I - \hat{\beta}\hat{\beta}^\top$ at L^*, with $\hat{\beta} = \beta/\|\beta\|$. The surprisal implementation constructs P exactly as the position implementation does and imports its pooled-moment and projection helpers unchanged. Since $P\beta = 0$, the edit removes the frozen probe's direction, making its prediction constant across positions. No post-edit frozen R^2 is reported in either arm: the position survival plotted in Figure 1b and the surprisal survival discussed in §5 are both re-fit ratios.

Representation arm—sample, split, and denominator. The surprisal removal run uses 48 single-document windows at stream offset 0 (documents 0–47), matching the position removal run on documents, window length, and split. Windows contain 4,096 tokens for Gemma and 2,048 for the other models. The $34/14$ document split at seed 0 is the same one used to fit the channel. The re-fit probe is trained in the β-orthogonal complement on the 34 fit documents and scored on the 14 evaluation documents. The survival fraction divides this re-fit R^2 by the frozen decoder's R^2 on the same 14 documents. That denominator differs from the pooled pre-removal R^2 (0.71–0.78 on surprisal, 0.85–0.98 on position); the artifacts store the ratio and the pooled value, not the split denominator. Table 11 reports survival and single-dimension concentration side by side for both targets. The discrepancy between the pooled and moment-based calculations is $\leq 1.4 \times 10^{-15}$ in all four models.

Representation arm—differences from the position removal run. The runs differ in three respects: the regression target, the layer at which the construct is fit and removed (L_s^* versus the position L^*), and the exclusion of position 0, one of 2,048 or 4,096 positions. The two arms share documents, window lengths, document split, ridge penalty, removal operator, and bootstrap draws.

Writing arm—head selection. The surprisal writer audit ranks heads within the run rather than loading archived identities. It uses the same per-head decomposition as the position ranking, $c^{l,i}(t) = \beta_s^\top (z_t^{l,i} W_O^{l,i})$, with the post-attention RMS scale and gain folded in exactly for sandwich-norm models. Heads at layers $l \leq L_s^*$ are then ranked by $|\mathrm{cov}_t(c^{l,i}, y)|$ against surprisal rather than against token index. The per-layer reconstruction error is $\leq 3.8 \times 10^{-5}$ of the attention con-

	R^2 retained after removal (%)		Single-dimension fraction (%)	
Model	Position	Surprisal	Position	Surprisal [95% CI]
Llama	73.6	89.0	5.3	14.6 [13.4, 15.9]
Pythia	87.8	81.6	10.5	8.7 [7.7, 9.7]
Gemma	86.1	89.8	53.6	14.4 [12.6, 15.9]
Qwen	81.0	94.9	8.6	29.8 [27.4, 31.9]

Table 11: Representation audit for position and surprisal at each target's selected layer. Retained R^2 is the post-removal re-fit score divided by the frozen probe's score on the same evaluation documents. The single-dimension fraction instead divides the best dimension's squared pooled correlation with the target by the frozen decoder's pooled R^2. Surprisal intervals use 2,000 document-bootstrap resamples at seed 0. The comparisons confound target with depth.

tribution's standard deviation. The run loads 48 single-document windows at stream offset 200 and divides them in stream order: documents 200–223 are used to rank heads, and documents 224–247 to score them. Ranking and scoring are therefore disjoint by document inside this arm. Both halves are disjoint by index from the surprisal channel fit at offset 0. In Llama, Pythia, and Qwen, both are also disjoint from the position audit's evaluation window at offset 500. In those same three models, one other offset-200 sample overlaps them: the archived position sweep read 16 documents there (documents 200–215), all of which fall inside this arm's ranking half. The fit-stability splits of Appendix F.5 also use offset 200, but only for Llama-3.1-8B-Instruct at 4,096 tokens, a differently filtered stream. Gemma's 4,096-token samples likewise use a different stream from the position sweep's 2,048-token windows. Eligible documents are filtered by each model's tokenizer and the window length, so "documents 200–223" refers to different texts in each model. For the same reason, Gemma's offset-200 ranking and scoring samples cannot be shown to be disjoint from the offset-500 position audit by index alone. The top-15 sets this ranking selects share one head in total with the archived position top-15 sets: 0 of 15 in Llama, Pythia, and Qwen and 1 of 15 in Gemma. The expected overlap under layer-matched draws is $\sum_l k_l^{\mathrm{s}} k_l^{\mathrm{p}} / n_{\text{heads}}$ heads per model: 0.00 in Llama, 0.06 in Pythia, 0.38 in Gemma, and 0.04 in Qwen, totaling 0.47. Here k_l^{s} and k_l^{p} are the two arms' per-layer counts and n_{heads} is the model's heads per layer.

Writing arm—ablation and readouts. Each condition zeroes the selected heads on `hook_z` with the operator of the layer-matched random-head control above, and the residual is read at L_s^*. Three readouts are recorded per condition. The primary axis readout is ρ, the Spearman correlation between C_s and the clean-model surprisal label over all pooled (document, position) pairs of the 24 scoring documents. The representation readout is a re-fit R^2: a ridge probe fit per condition on that condition's ablated activations under $\lambda_{\mathrm{r}} = 10^{-2}$, with a document-level split of the scoring half (17 fit, 7 evaluation, seed 0) held identical across the baseline, the top set, and every draw. The frozen decoder's R^2 is recorded as a diagnostic without a per-draw null. It falls from 0.731 to 0.675 in Llama, 0.680 to 0.578 in Pythia, 0.763 to 0.655 in Gemma, and 0.724 to 0.177 in Qwen. Qwen's 0.55 drop is the largest, yet neither tested family rejects its null (Table 12). The two tested statistics are the frozen-ρ damage $E_{\text{top}} = \rho_{\text{base}} - \rho_{\text{top}}$ and the re-fit damage $E_{\text{top}}^{\text{refit}}$, with baselines measured on the scoring half itself.

Writing arm—nulls. Both nulls draw 300 random head sets with the same per-layer counts as the selected top set from one deterministic stream (seed 0), the exclusion null first. Heads are sampled without replacement within each set. The exclusion null (primary) draws from the pool of Eq. 17, and the inclusion null (robustness) from the full eligible pool. The full pools contain 480, 416, 448, and 504 heads for Llama, Pythia, Gemma, and Qwen, compared with 64, 96, 208, and 112 in the position arm. The deeper L_s^* enlarges every pool. Expected inclusion overlap with the top set is $\sum_l \mathrm{count}_l^2 / n_{\text{heads}}$, or 1.03, 1.16, 2.44, and 1.25 heads in the Llama, Pythia, Gemma, Qwen order used throughout this section; the position protocol above lists its own overlaps in a different order. Replaying the seeded sampler gives realized means of 1.03, 1.23, 2.46, and 1.20. All 300 draws are feasible in both nulls in all four models, and the replay confirms all 300 are distinct in every case. As in the position audit, the draws match layer composition and head count, not per-head output norm, aggregate output dose, or realized residual displacement.

Writing arm—statistics and the two Holm families. With $N = 300$ feasible draws, the one-sided empirical $p = (1 + \#\{E_r \geq E_{\text{top}}\})/(N + 1)$ has a resolution of $1/301 = 0.003$. We report two

(a) Frozen-ρ readout

Model	L^*	$\rho_{\text{base}} \to \rho_{\text{top}}$	Damage	Excl. 99th	Incl. 99th	Holm-p
Llama	14	$0.837 \to 0.811$	0.027	0.054	0.056	0.286
Pythia	12	$0.819 \to 0.749$	0.070	0.041	0.038	0.013
Gemma	27	$0.831 \to 0.745$	0.086	0.029	0.071	0.013
Qwen	17	$0.831 \to 0.575$	0.256	1.022	1.022	0.336

(b) Re-fit R^2 readout and stored verdict

Model	Drop	Excl. 99th	Incl. 99th	Holm-p	Stored verdict
Llama	0.036	0.033	0.032	0.020	not-localized†
Pythia	0.028	0.025	0.024	0.020	**representation-localized**
Gemma	0.046	0.015	0.035	0.013	**representation-localized**
Qwen	0.119	0.397	0.402	0.336	not-localized

(c) Raw empirical p values

	Frozen-ρ readout		Re-fit R^2 readout	
Model	Exclusion	Inclusion	Exclusion	Inclusion
Llama	0.143	0.156	0.007	0.007
Pythia	0.003	0.003	0.010	0.010
Gemma	0.003	0.003	0.003	0.003
Qwen	0.336	0.359	0.336	0.359

Table 12: Post-hoc surprisal writer-head audit: 24 documents rank the heads and 24 disjoint documents score them. Panel (a) reports pooled readout–surprisal rank correlation and its loss; panel (b) reports loss of re-fit R^2. Null columns give the 99th percentiles of 300 layer-count-matched draws, excluding or including the selected heads in the pool. Each Holm column corrects a separate four-model exclusion family; panel (c) gives the raw exclusion and inclusion p values. Stored verdicts require a frozen-axis pass before the re-fit gate; raw and adjusted p give the same labels. †Llama's re-fit-only separation is exploratory. The preceding protocol records all sample and depth differences from the position audit; the following paragraphs explain the stored labels and remaining protocol differences. No next-token-loss check was run, and the nulls do not match output dose.

Holm families of $m = 4$, each covering one statistic across the four models' exclusion tests. The frozen-ρ family rejects the exclusion null for Pythia and Gemma; the separate re-fit family also rejects it for Llama (Table 12). At 300 draws, the smallest attainable adjusted p at $m = 4$ is 0.013, and the position audit's 0.004 is out of reach in this arm at any effect size. Panel (c) lists the raw empirical p values; inclusion-null results are reported without correction as robustness checks. Re-fit damage exceeds both null 99th percentiles in Llama, Pythia, and Gemma, but neither in Qwen. For Qwen, 100 of 300 exclusion draws cause at least as much frozen-ρ damage as the top set, and 100 cause at least as much re-fit damage.

Writing arm—the stored verdict field. The runner stores one verdict per model from the locked claim taxonomy, whose specification requires a Holm-adjusted exclusion p. It passes the raw p instead, and the stored field was therefore not Holm-gated. The correction does not change any verdict: 0.003 and its adjusted 0.013 both meet $\alpha = 0.05$, whereas 0.143 with its adjusted 0.286 and 0.336 with its adjusted 0.336 do not. The stored verdicts are representation-localized for Pythia and Gemma and not-localized for Llama and Qwen. Llama's re-fit damage of 0.036 exceeds both null 99th percentiles, but its stored label remains not-localized because the re-fit gate is conditional on an axis pass. §6 therefore reports this result under the exploratory re-fit family.

Writing arm—differences from the position writer audit. Fourteen protocol differences separate this arm from the held-out position writer audit above.

1. The audit uses a deeper layer. L_s^* is 14, 12, 27, and 17, versus the position probe's 1, 2, 12, and 3. This enlarges the eligible pools from 64, 96, 208, and 112 heads to 480, 416, 448, and 504. Target and depth are confounded in every comparison between the two arms.

2. Head sets are re-ranked inside the run, whereas the position audit loads preselected identities verbatim from an archived sweep artifact.
3. Heads are ranked by covariance with surprisal, while the position ranking uses covariance with token index.
4. Documents come from stream offset 200, split 24 for ranking and 24 for scoring; the position audit scores 48 documents at offset 500 against a set ranked in a separate run.
5. Gemma's windows are $4{,}096$ tokens, against $2{,}048$ in all four models of the position audit; the other three models here read $2{,}048$ in both arms.
6. ρ is a rank correlation over pooled (document, position) pairs against surprisal, whereas the position audit rank-correlates a document-averaged readout curve of length S against token index. The two arms' ρ values are not the same statistic and are not comparable in magnitude.
7. Position 0 is dropped from every fit and every ablation readout, because surprisal is undefined at the first token.
8. The re-fit split is 17 fit and 7 evaluation documents, against 34 and 14.
9. Each null draws 300 sets, against $1{,}000$, which sets the empirical-p floor at $1/301 = 0.003$ and the $m = 4$ Holm floor at 0.013.
10. No next-token-loss content check is computed. The position writer-audit artifacts record LM-loss ratios of 1.005–1.033; the surprisal implementation neither computes nor records this quantity. The nulls match layer composition and head count, not per-head output norm, aggregate output dose, or realized residual displacement. General damage from ablation is therefore unmeasured in this arm.
11. The stored verdict field was gated on the raw exclusion p rather than the Holm-adjusted one, as described above.
12. The artifacts have no configuration block. Position artifacts record $\{n_{\text{eval}}, \text{offset}, \text{seq_len}, \text{test_frac}, \alpha, n_{\text{draws}}, \text{seed}\}$. Surprisal artifacts record the model, regression target, layer, heads per layer, draw count, seed, top-k, and selected head list. They omit offset, window length, re-fit split fraction, α, and document count.
13. Draw distinctness is printed at run time and not stored. Position artifacts record a distinct-draw count; we recovered the surprisal counts by replaying the seeded sampler.
14. The arm is post-hoc. It appears in no preregistration artifact, and Table 6 names five frozen artifacts, none of them a surprisal specification.

The document count rests on the run directory, not the artifacts. We report $n_{\text{docs}} = 48$ for the writer arm on two pieces of evidence: the results directory is named for that count, and the release script's surprisal writer-audit step takes the count as a parameter, with a default of 24 that this run overrode. An independent 24-document run of the same step is archived beside it, with different head orders and a different ρ baseline. This confirms that the directories contain different runs, but does not independently establish the larger run's document count. No stored field records the document count, and the artifact supplies no third line of evidence.

Status. This arm is a post-hoc control and carries no preregistered verdict. §6 labels Qwen's position re-fit comparison exploratory and reports Llama's surprisal re-fit separation on that same footing. The two differ in provenance: the position companion re-scored arrays a preregistered run had already logged, whereas this arm fits a new channel and runs three fresh stages—probe fit, removal, and writer audit. This control tests whether concentration, redundancy, and writer localization are specific to the position channel. Its results describe the surprisal channel; they do not provide additional evidence about position.

E Criterion Coverage, Probe Validity, and Base–Instruct Transfer

Table 14 applies the committed content check to every Formation cell, identifying which cells can support the mask and cardinality claim (§4).

Model	Intervention	Frozen drop	Re-fit drop	Δppl	Content check
Pythia	exclude position 0	0.741	0.473	+0.002	passes
Llama	exclude position 0	1.403	0.463	+0.065	passes
Gemma	sink-value block	17.404	0.059	+0.031	passes
Qwen	high-frequency control	1.996	−0.057	+0.259	fails
Gemma	exclude position 0	119.628	−0.081	+0.285	fails

Table 13: Formation disagreements: exactly one readout's loss in position R^2 crosses the preregistered 0.5 threshold, extended to the re-fit readout by a logged amendment. Δppl is the perplexity ratio minus one; the content check passes below 0.15. A failed check makes the cell content-confounded. Both scores use eight evaluation documents per model from the 16-document grid, without cell-level intervals.

Threshold sensitivity. At the committed 0.5 drop threshold, the two readouts agree in 23 of 28 cells: twenty fall below the threshold on both, and the visible-set clamp exceeds it on both in Pythia, Gemma, and Llama. The five disagreements are all frozen-only. At a lower threshold of 0.45, three disagreements are frozen-only and two are re-fit-only. Pythia's and Llama's position-0 cells now agree, while Qwen's position-0 and clamp cells now disagree. Qwen's re-fit drops are 0.476 and 0.453, respectively, against frozen drops of 0.381 and 0.358. All four re-fit drops that change agreement status lie within 0.003–0.026 of 0.45, too close to assess without uncertainty estimates. For comparison, Gemma's mass-matched placebo alone changes a re-fit drop by 0.023. Of the five disagreements at the committed threshold, three pass the content check. Their re-fit drops are 8% (Gemma's sink cell), 49% (Pythia's position-0 cell), and 61% (Llama's position-0 cell) of the corresponding re-fit baseline. Thus a below-threshold re-fit drop can still represent substantial loss of recoverable position.

Content and probe controls. Two further controls increase perplexity as much as or more than their models' position-specific cells, yet remain below the 0.15 content-damage threshold. Pythia's high-frequency control has a perplexity change of $+0.120$, versus $+0.002$ and $+0.006$ in those cells. Llama's placebo has a change of $+0.083$, equal to its clamp cell and above its position-0 cell at $+0.065$. Tables 15 and 18 establish that the base-model probe is not a residual-norm readout and that decodability precedes L^* (§3). Archived base-model FineWeb controls preserve the position readout under token randomization and content-order shuffling, while a token-embedding-only diagnostic carries negligible position signal. A disjoint frozen-decoder audit reported with the full-U selectivity results (Appendix F.2) now supplies the corresponding content controls for the Llama-Instruct MDQA decoder.

Rank k_{85} is the smallest prefix of U whose removal reduces position R^2 recoverable from the complement by 85% (Appendix C).

Table 17 re-evaluates the frozen position writer-head sets on held-out documents with thousand-draw nulls and a re-fit readout. Figure 6 plots the per-draw dissociation underlying Qwen's post-hoc label (§6); the protocol is defined in Appendix C. Per-model effect sizes are in Appendix G. The inclusion null is concordant with the exclusion null in every model. In Qwen, the exclusion 99th percentile (1.153) exceeds the top-set effect (0.095): random sets damage the frozen rank readout more than the top set does, while the re-fit column shows the reverse. The post-hoc reanalysis applies a separate Holm correction across the four re-fit drops, giving $p = 0.004$ in every model under both nulls. Qwen's re-fit R^2 falls from 0.803 to 0.354. This separation depends on the diagnostic: the frozen-ρ and re-fit-R^2 tests differ in metric and evaluation aggregation. Next-token LM-loss ratios on the audit documents are 1.005–1.033 throughout, a logged content-damage check rather than a gate. Thus, the measured probe effects are accompanied by only small changes in next-token loss.

Count-matched scramble control. The preregistered scramble randomized which causal predecessors remained available at the same nominal window width, $w = 512$. Its post-sampling retention of P0 makes realized cardinality differ by at most one key from the trailing-window control (Appendix C). Its pooled frozen-coordinate drops were 0.005 (Pythia), 0.069 (Qwen), 1.179 (Gemma), and 0.092 (Llama); the corresponding perplexity ratios were 1.30, 10.49, 1.28, and 1.13. Three of four ratios exceed the committed 1.15 perplexity-ratio threshold, and only pooled statistics were retained. We report this control without an uncertainty interval and do not use it to separate cardinality from recency or content disruption.

Model	Intervention	Frozen-probe drop	Re-fit-probe drop	Δ perplexity	Committed criterion label
Pythia	placebo	0.007	0.000	+0.001	no effect
Pythia	phase shuffle	0.000	−0.000	+0.000	no effect
Pythia	high-frequency control	0.040	−0.001	+0.120	no effect
Pythia	lowest-frequency derotation (G3)	0.000	0.001	+0.000	no effect
Pythia	sink-value block (G1)	0.034	0.004	+0.004	no effect
Pythia	exclude position 0 (G1-attention)	0.741	0.473	+0.002	position-specific
Pythia	visible-set clamp (G2)	0.788	0.524	+0.006	position-specific
Qwen	placebo	0.010	0.009	+0.000	no effect
Qwen	phase shuffle	0.001	0.001	+0.000	no effect
Qwen	high-frequency control	1.996	−0.057	+0.259	general damage
Qwen	lowest-frequency derotation (G3)	−0.000	−0.001	+0.000	no effect
Qwen	sink-value block (G1)	−0.005	−0.046	+0.019	no effect
Qwen	exclude position 0 (G1-attention)	0.381	0.476	+0.000	no effect†
Qwen	visible-set clamp (G2)	0.358	0.453	+0.015	no effect†
Gemma	placebo	−0.025	−0.023	+0.008	no effect
Gemma	phase shuffle	−0.003	−0.006	+0.000	no effect
Gemma	high-frequency control	0.020	−0.006	+0.022	no effect
Gemma	lowest-frequency derotation (G3)	0.003	−0.007	+0.000	no effect
Gemma	sink-value block (G1)	17.404	0.059	+0.031	position-specific
Gemma	exclude position 0 (G1-attention)	119.628	−0.081	+0.285	general damage
Gemma	visible-set clamp (G2)	9.745	0.618	+0.202	general damage
Llama	placebo	0.006	−0.001	+0.083	no effect
Llama	phase shuffle	0.000	−0.000	+0.000	no effect
Llama	high-frequency control	0.221	0.000	+0.250	no effect‡
Llama	lowest-frequency derotation (G3)	−0.000	−0.000	−0.000	no effect
Llama	sink-value block (G1)	0.409	0.070	+0.021	no effect
Llama	exclude position 0 (G1-attention)	1.403	0.463	+0.065	position-specific
Llama	visible-set clamp (G2)	0.847	0.557	+0.083	position-specific

Table 14: The symmetric content check evaluated on all 28 Formation cells (preregistered frozen-probe drop threshold 0.5; amended content threshold = perplexity ratio $-1 = 0.15$). The 16-document grid reserves eight documents for fitting and eight for evaluation; exact document identities are not recorded in the archived results. Baseline position R^2 differs by model: frozen 0.979, 0.794, 0.672, and 0.858, and re-fit 0.964, 0.644, 0.763, and 0.762 for Pythia, Qwen, Gemma, and Llama. The absolute 0.5 threshold therefore corresponds to 52% of Pythia's re-fit baseline and 78% of Qwen's. A negative re-fit drop means the intervention's re-fit R^2 exceeds the shared baseline. Per-document re-fit inputs were not retained, so no entry has cell-level uncertainty; negative values are not interpreted as improvements. The threshold-sensitivity analysis in this appendix gives the split counts and losses relative to each model's re-fit baseline. The committed labeling rule applies the thresholds sequentially. A frozen-probe drop below 0.5 is labeled *no effect*; above that threshold, a cell is *position-specific* if Δ perplexity stays below 0.15 and *general damage* otherwise. "Position-specific" is the verbatim historical diagnostic label and does not denote behavioral decoder-null selectivity. †Both Qwen cells stay below the 0.5 cut on both readouts, although their re-fit drops (0.476 and 0.453) exceed their frozen drops (0.381 and 0.358). These point estimates motivate reporting both scores; neither cell crosses the committed threshold. ‡Two control cells exceed the content threshold, but their labels flag only one. Qwen's high-frequency control ($+0.259$) also exceeds the frozen-drop threshold and is labeled *general damage*. Llama's ($+0.250$) exceeds neither readout's drop threshold and is labeled *no effect*; this means that the position-drop threshold was not met, not that content damage was absent.

F Behavioral Intervention Verdicts and Summary

Tables 19, 20, 21, and 22 report per-condition verdicts for the behavioral interventions of §7. The two-stage intervention-layer audit appears in Tables 29 and 31, with instrument checks in Table 30. Instrument checks pass in all twelve cells: every persistence readout passes, manipulation and matching checks are exact, and next-token-loss ratios are 0.95–1.03. In Stage 1, the reversal rerun at the reference layer reproduces the archived treatment effect (-0.131 $[-0.195, -0.078]$, Table 22) to the printed precision. The reference-row placebos are new runs of the complement sham. The reversal placebo is therefore -0.008 in Table 29, versus the archived -0.018 in Table 22; its reference contrast uses the new value. Only the treatment arms were required to reproduce the archived effects. Stage-1 manipulation and $\Delta\|h\|$-matching checks also pass (median ratio 1.000). Table 32 marks

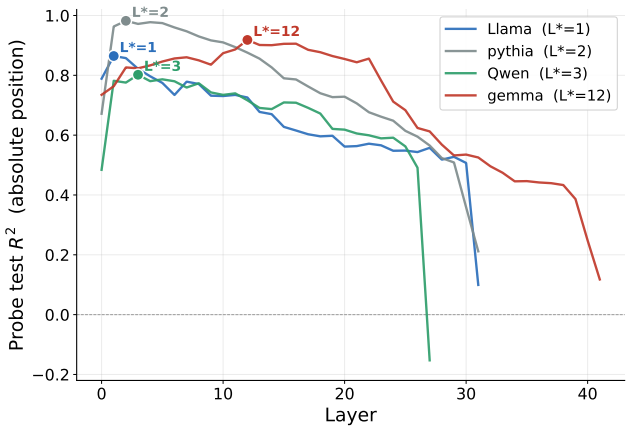


Figure 3: Layer profiles of absolute-position decodability in the four base models. The peak layer L^* is the argmax of the plotted test R^2; it defines the fixed intervention coordinate and is not reported as an unbiased estimate of the maximum.

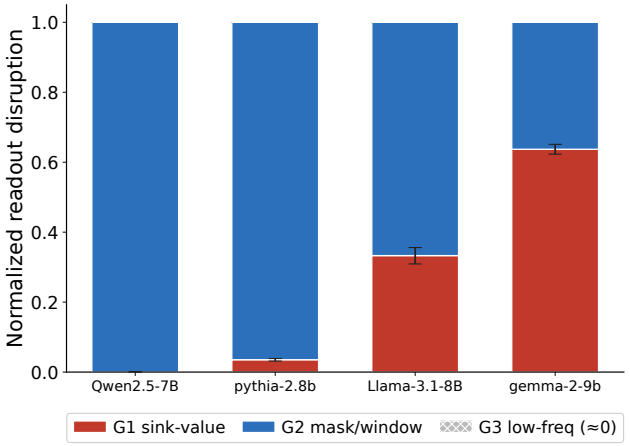


Figure 4: Frozen-readout disruption shares from a separate 48-document Formation re-run, with 24 evaluation documents per model and 95% document-bootstrap intervals on the sink share. For each operator, we subtract the mass-matched placebo's nonnegative frozen-R^2 loss and clip the remainder at zero; G2 averages the three window widths before the three components are normalized to sum to one. These are readout-dependent ordinal summaries: sink-value interventions are displacement-dominated, and the sink share collapses under a re-fit-based decomposition. The bars are not mechanistic variance fractions, zero-valued components are boundary-censored, and the cross-model ordering by sink share is descriptive at $n = 4$.

which cells of the §7 design were evaluated and which were not. The span-restricted QA transplant remains descriptive because its potency control was sequence-wide; the ordinal-task comparisons follow their own validity gates. Sequence-wide magnitude and value interventions move the Llama dip for the fixed U. In the corrected five-prefix replay, raw magnitude effects are negative in $5/5$, paired complement-contrast intervals exclude zero in $4/5$, and no 95% treatment interval lies entirely below -0.065. The full-sequence value effect has the same sign across two samples, and its corrected five-prefix replay differs from the complement control in $5/5$ variants. Its magnitude still depends on the sample and U, and no variant's 95% treatment interval lies entirely below -0.065. A residual edit at L^* produces about half the effect of physical relocation: the full-residual swap recovers 46–52% of the physical-swap effect in all four ordinal conditions, whereas the span-restricted subspace transplant is near zero (Figure 7). The pathways responsible for the remaining half are unknown.

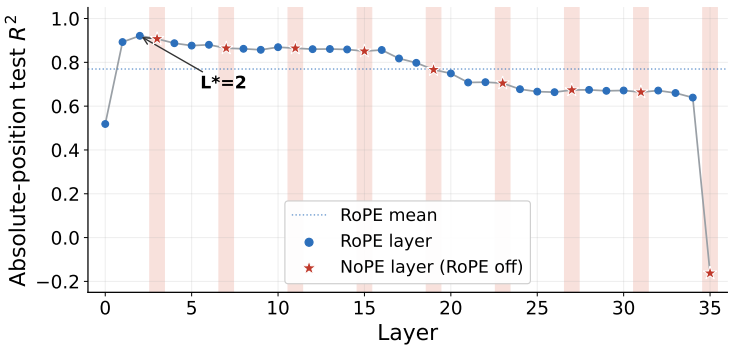


Figure 5: Hybrid maintenance test. Each point is a ridge probe fit independently at that layer; stars and shaded bands mark the nine NoPE layers, the dotted line the RoPE-layer mean $R^2 = 0.77$, and L^*=2 the peak-R^2 layer. In SmolLM3, position remains decodable through most interleaved NoPE layers, with no systematic drop at RoPE-to-NoPE boundaries; where decodability first arises remains unknown. Eight of the nine NoPE boundaries change R^2 by at most 0.031; at the final boundary, R^2 falls from $+0.639$ to -0.163.

Model	L^*	$\cos(\beta, \bar{h})$	$\cos(\beta, \dot{h})$	$\mathrm{corr}(C_\beta, \|h\|)$	pos-from-$\|h\|$ R^2	partial$(C_\beta, t \mid \|h\|)$	norm-std R^2/ρ
Pythia	2	$+0.018$	$+0.013$	-0.050	0.002	$+0.990$	0.948/0.980
Qwen	3	-0.043	$+0.000$	-0.030	0.002	$+0.883$	0.794/0.908
Gemma	12	-0.017	$+0.006$	-0.153	0.003	$+0.922$	0.880/0.951
Llama	1	$+0.048$	$+0.000$	-0.042	0.001	$+0.927$	0.874/0.948

Table 15: Scale diagnostics at L^* (§3). β aligns with neither the mean-residual direction $\bar{h}$ nor the mean trajectory derivative $\dot{h}$. The frozen coordinate is weakly correlated with residual norm, and norm alone does not predict position. The partial correlation with position remains high after controlling for norm, and the probe remains effective after position-wise norm-standardization.

F.1 Displacement accounting and the translation family

The raw reversal effect is accompanied by a larger realized displacement than the ablation anchor. For question q and arm a, let $\mathcal{S}_q$ denote the intervened token positions; we define cumulative sequence-wide U-space displacement and its normalized behavioral effect as

$$D_{a,q} = \sum_{t \in \mathcal{S}_q} \|\Delta z_{a,q}(t)\|_2, \qquad E_{a,q} = \frac{\Delta \,\mathrm{dip}_{a,q}}{D_{a,q}}. \tag{18}$$

Reconstructing the unedited forward pass at L^* on the same queries gives pooled per-position median $\|\Delta z\|_2$ values of 0.260 for reversal and 0.187 for the ablation anchor. The median paired cumulative-dose ratio is 1.425 $[1.420, 1.428]$. Each question contributes its own ratio of summed displacements, and 1.425 is the median of those 80 values. Per 1,000 cumulative-displacement units, E is -0.156 $[-0.234, -0.092]$ for reversal and -0.139 $[-0.205, -0.082]$ for ablation. The paired difference is -0.018 $[-0.051, +0.011]$. These intervals use the same paired question-level bootstrap ($n_q = 80$, $B = 20{,}000$); token positions are not resampled. We therefore do not detect a difference in this displacement-normalized diagnostic. Because the normalization does not control intervention geometry, it does not establish displacement as the causal explanation for the raw effect difference.

The translation family of §7 shifts the decoded coordinate without a corresponding translation of the answer profile. The realized decoded shift is 0.79–0.98 of its target Δ across $\{\pm T/8, \pm T/4\}$, whereas placebo shifts are approximately zero and perplexity ratios are 0.99–1.01. The behavioral profile nevertheless does not translate at any Δ. No cell is excluded, and at least two shift magnitudes remain valid in each direction.

Pair	$\cos\beta$	proj^2	cross-R^2 (b→i / i→b)	within-fit (b / i)	full-32 affinity	min. angle
Llama	$+0.767$	0.61	0.60 / 0.69	0.86 / 0.85	0.060	39°
Gemma	$+0.909$	0.84	0.84 / 0.62	0.84 / 0.82	0.062	30°
Qwen	$+0.764$	0.66	0.76 / 0.72	0.77 / 0.75	0.091	29°

Table 16: Base–instruct bridge on a matched corpus. Direction alignment is $\cos\beta$; proj^2 is the squared Euclidean projection of the unit base-model decoder direction into the instruct k_{85} prefix, not a covariance-weighted variance fraction. Cross-decoding refits intercept and scale on a fitting partition and scores against within-fit references on the held-out partition. Affinity averages $\cos^2$ over all 32 saved principal angles between the full bases; the final column gives their minimum angle. These are full-basis measures, not prefix-to-prefix geometry, which cannot be recovered by truncating the saved angle spectra. Aligned leading directions coexist with low full-basis affinity; independent-fit variability limits attribution of the subspace difference to instruction tuning alone.

Model	$\rho_{\text{base}} \rightarrow \rho_{\text{top}}$	E_{top}	Holm-p	excl. 99th	re-fit drop / 99th	incl. p	Separation
Llama	$0.998 \rightarrow 0.033$	0.965	0.004	0.004	0.538 / 0.039	0.002	**both readouts**
Pythia	$1.000 \rightarrow 0.807$	0.193	0.004	0.018	0.478 / 0.008	0.004	**both readouts**
Gemma	$0.998 \rightarrow 0.976$	0.022	0.004	0.002	0.139 / 0.013	0.001	**both readouts**
Qwen	$0.997 \rightarrow 0.902$	0.095	0.315	1.153	0.449 / 0.241	0.274	re-fit only^{P}

Table 17: Held-out position writer-head audit. The preselected top-15 head identities are kept fixed and evaluated on disjoint documents against 1,000 layer-matched random sets per null. Every model uses 2,048-token windows, matching the original sweep's readout context. The exclusion null (top heads removed from the pool) is primary and Holm-corrected over the four models; the inclusion null (full eligible pool, overlap-aware) gives incl. p. "Both readouts" additionally requires that a probe re-fitted on the ablated activations lose decodability beyond the null's 99th percentile (a conditional second gate, no separate Holm). The readouts differ in metric, aggregation, and evaluation subset; their disagreement does not isolate re-fitting, and the nulls do not match output dose. ^{P}Post-hoc label from an exploratory symmetric reanalysis of the logged re-fit readout. The protocol, data partitions, null definitions, and re-fit split appear in Appendix C. Here ρ is the rank correlation between the document-averaged readout curve and token index.

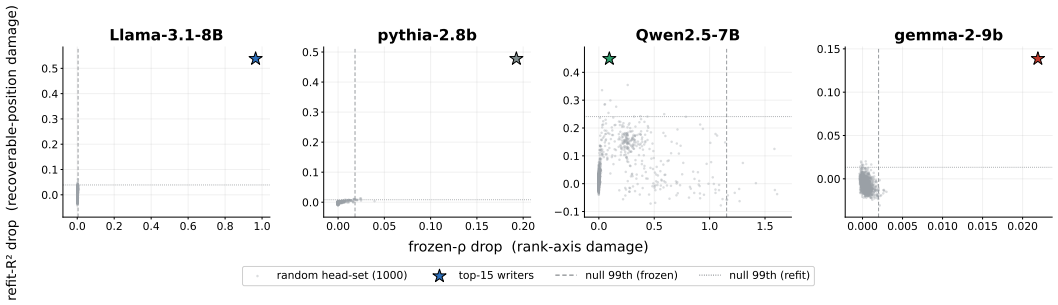


Figure 6: Per-draw dissociation in the held-out position writer audit. The vertical axis shows re-fit-R^2 drop; the horizontal axis shows frozen-ρ drop. Dots represent 1,000 layer-matched random head sets from the exclusion null, and the star marks the preselected top-15 set. Dashed and dotted lines mark the frozen and re-fit null 99th percentiles. In Llama, Pythia, and Gemma, the top set exceeds both thresholds. In Qwen, it misses the frozen-ρ threshold yet clears the re-fit threshold: 31% of random sets damage the frozen axis at least as much as the top set, while none reaches its re-fit damage (exploratory reanalysis of logged arrays; §6). The axes differ in metric and evaluation aggregation as well as in the probe.

Model	layer-0 R^2 [CI]	layer-0 ρ	refit L^*	R^2 at refit L^*
Pythia	$0.918\ [0.914, 0.922]$	0.990	$4^{\ddagger}$	0.981
Qwen	$0.416\ [0.378, 0.453]$	0.934	3	0.840
Gemma	$0.735\ [0.721, 0.747]$	0.986	12	0.918
Llama	$0.756\ [0.744, 0.766]$	0.983	1	0.864

Table 18: Early decodability: position probes at the first block's output (layer 0), after block-0 attention. This diagnostic refits probes independently of the frozen coordinate on its own document sample, so values differ slightly from the original layer-profile analysis. ‡The independent refit selects layer 4 for Pythia, whereas the frozen coordinate of §3 is read at layer 2; the layer-0 readout does not depend on this choice. These values measure decodability at the earliest post-block site. The archived embedding-only baseline carries no position signal (test R^2 between -0.070 and -0.040 in all four models).

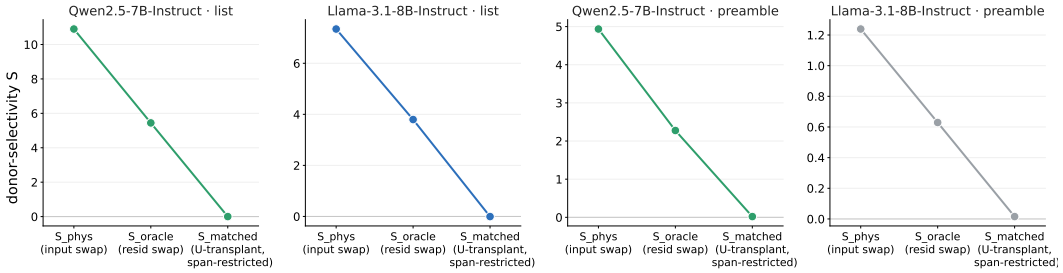


Figure 7: Span-restricted intervention ladder: physical input swap, full-residual swap at L^*, and position-subspace value transplant. Each panel is one model · ordinal task, and the y-axis is donor selectivity S (Eq. 12). Panel scales are independent: S_{phys} spans 1.2 to 10.9 across the four arms. The physical–residual gap reflects pathways changed by physical relocation but not by the L^* residual swap; it cannot be attributed uniquely to RoPE phase. Left to right, the residual swap reaches 50, 52, 46, and 51% of the physical swap. The residual–subspace gap compares the tested U-value intervention with a same-span edit that exchanges the complete residual vectors. Gray marks the Llama sentence-preamble arm, which the content-leak control excludes (Table 20); the remaining arms take their model's color.

Multi-document QA	dip	ablation ΔD	transplant slope	outcome
Llama-It	$0.34\ [0.19, 0.52]$	$-0.052\ [-0.08, -0.03]$	$0.002\ [-0.014, 0.013]$	descriptive (scope mismatch)
Qwen-It	$0.16\ [0.01, 0.30]$	$-0.000\ [-0.04, 0.04]$	$-0.056\ [-0.109, 0.005]$	no conclusion (positive control null)
Gemma-IT	$0.37\ [0.21, 0.57]$	$-0.040\ [-0.13, 0.06]$	$0.040\ [-0.065, 0.158]$	inconclusive (positive trend)

Table 19: Value transplant on multi-document QA. Complement placebos are null in all conditions; manipulation checks pass. Ablation is full-sequence; transplant and placebo are span-restricted (mismatch disclosed in §7). The small Llama slope is descriptive; it does not establish a null span effect under scope-matched controls. Qwen's row is the earlier transplant-battery estimate at n_q=120 (the others use n_q=80); at n_q=80, its baseline is $0.093\ [-0.088, 0.273]$. It is distinct from the later sequential audit on fresh records 160–279, whose full-U ablation is -0.055 and whose complement-control gate fails (Table 28). Gemma's ablation column is the transplant battery's execution of the full-sequence U-ablation of Eq. 13; the magnitude battery's execution of the same operator gives $-0.019\ [-0.13, +0.11]$ (Table 21). Both intervals contain zero: the potency control fails under either execution. The Llama dip interval was recomputed from its retained question-level baseline with 20,000 resamples. The Qwen and Gemma dip intervals, and the three ablation-gate intervals, retain the initial battery's 2,000-resample bootstrap because the required question-level arrays were not all retained; transplant-slope intervals use 20,000 resamples. The transplant slope is defined in Appendix C.

Ordinal task · model	transplant (% of swap reference)	ratio CI	controls	outcome
word list · Qwen	0.0%	$[-0.0, +0.0] \ni 0$	gates passed	**ratio-equivalent to null**
word list · Llama	-0.1%	$[-0.002, -0.001]$	gates passed	**ratio-equivalent to null**
sentence preamble · Qwen	$+0.9\%$	$[+0.006, +0.012]$	no target-item drop	weak positive
sentence preamble · Llama	$(+2.5\%)$	—	matched > cross	**excluded** (content leakage)

Table 20: Donor-selectivity outcomes on the marker-free ordinal tasks, relative to the same-span full-residual-swap reference defined in Appendix C. All selected pairs pass the phenomenon, positive-control, and swap-reference gates; the Llama sentence-preamble condition is subsequently excluded by its content-leak control. Both word-list conditions are ratio-equivalent to null because the rule requires containment in the $[-0.2, +0.2]$ equivalence band, not inclusion of zero. Llama's interval excludes zero at -0.1% of the swap reference, but this magnitude falls within the equivalence band.

Magnitude (full sequence)	α=0 control	slope	placebo	outcome
Llama-It	-0.065 $[-0.10, -0.03]$	$+0.029$ $[+0.004, +0.065]$	$-0.025 \ni 0$	**confirmed**
re-run, matched $\|\Delta h\|$	—	$+0.035$ $[+0.008, +0.069]$	-0.000 $[-0.018, +0.019]$	**dose-controlled**
Gemma-IT	-0.019 $[-0.13, +0.11]$	$+0.089$ $[+0.037, +0.145]$	$+0.010 \ni 0$	positive slope[‡]

Table 21: Dose–response of the dip to global U-magnitude scaling, $\alpha \in \{0, 0.5, 1, 2, 4\}$. Span-scope scaling does not establish a dose response (-0.006, CI $\ni 0$): a preregistered prediction failure that motivated scope matching. The original complement placebo matched rank and α but not injected displacement; the audit re-run rescales it to the treatment's realized per-position $\|\Delta h\|$ (measured ratio 1.00). α=0 intervals from the initial magnitude battery use its legacy 2,000-resample flattening bootstrap; slope intervals and the displacement-matched re-run use 20,000 question-level resamples. [‡]Potency-inconclusive: Gemma's α=0 control does not establish dip flattening, so the slope remains descriptive. The dose-response slope is defined in Appendix C.

Value reversal (full seq.)	dip	ΔD	α=0 control	placebo	outcome
Llama-IT	0.342	-0.131 $[-0.195, -0.078]$	-0.082 $[-0.121, -0.048]$	-0.018 $[-0.049, +0.009]$	**effect-size rule passed**
replication (80–159)	0.250	-0.064 $[-0.104, -0.022]$	-0.055 $[-0.085, -0.026]$	-0.003 $\ni 0$	directional only[a]
pooled (n_q=160)	—	-0.097 $[-0.136, -0.063]$	—	—	directional[a]
Gemma-IT	—	-0.156 $[-0.331, +0.003]$	$[-0.130, +0.078] \ni 0$	—	no conclusion[b]
Qwen-IT	0.093 $[-0.088, +0.273]$	$+0.025$ $[-0.101, +0.156]$	—	—	no conclusion[c]

Table 22: Norm-preserving value reversal at full-sequence scope (n_q=80). The operator reverses the U-value profile along the context while holding each position's magnitude fixed; the placebo is $\|\Delta h\|$-matched. Brackets give 95% intervals, and $\ni 0$ marks one containing zero. Treatment, positive-control, and placebo intervals use 20,000 paired question-level resamples. The Qwen baseline-dip interval is the initial 2,000-resample phenomenon gate; its question-level baseline was not retained. The replication row is the preregistered disjoint-sample re-run, frozen before execution. Outcome legend: [a]the minimum-effect rule is not met, i.e. the 95% interval does not lie entirely below -0.065; [b]potency of the α=0 control is not established; [c]the baseline dip is not established at n_q=80, with a fresh n_q=120 audit in Appendix F.6.

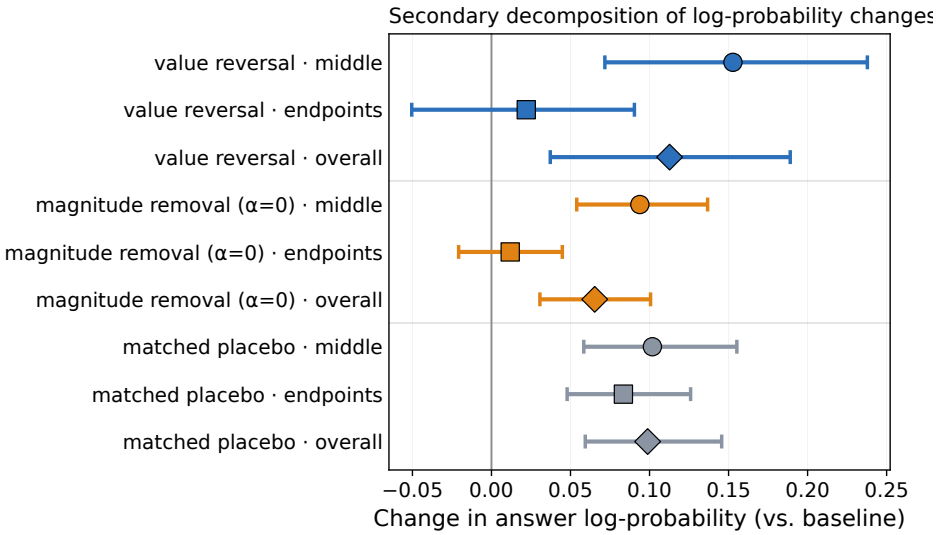


Figure 8: Secondary decomposition of the original Llama full-sequence audit (records 0–79; $n_q = 80$). Markers show mean changes relative to baseline and bars show paired question-level 95% bootstrap intervals. Endpoints are slots 0 and 19 and middle is slots 8 and 11 (Eq. 2). Overall averages all ten slots. The dip reduction under value reversal and magnitude removal is driven primarily by middle-position recovery rather than endpoint degradation. Overall log-probability also shifts under the matched placebo, so that shift alone does not establish a treatment-specific benefit. Dip change remains the primary behavioral endpoint; this decomposition is outside the preregistered confirmatory analysis.

F.2 DECODER-NULL SELECTIVITY AUDITS

k_{85}-prefix nullspace audit. A disjoint selectivity audit on records 200–279 ($n_q = 80$) tests whether the full-U magnitude effect survives a decoder-null control in the k_{85} prefix subspace. For the fixed Llama basis used in this audit, $k_{85} = 7$. The independent scalar position decoder has rank one on this prefix, leaving a tested nullspace of rank $7 - 1 = 6$. Magnitude removal changes the dip by -0.055 $[-0.102, -0.020]$. In contrast, a placebo constructed in the independent decoder's nullspace within $U_{k_{85}} \subset U$, matched on support and norm, changes the dip by only -0.004 $[-0.035, +0.025]$. Its 90% interval satisfies the preregistered TOST equivalence test. The placebo preserves the independent decoded-position readout to 7×10^{-8}, compared with a full-U treatment change of 1.51. The realized displacement ratio is 1.000, content cosine is 0.9999, and the negative-log-likelihood (NLL) ratio remains 0.95–0.98. All three members of the audit family satisfy their Holm-corrected decision rules ($m = 3$). Under these controls, the full-U magnitude effect differs from the decoder-null perturbation matched on support and norm in the k_{85} prefix subspace, as well as from the complement-subspace placebo. The nullspace check preserves this scalar decoder's readout, not necessarily all linearly recoverable position information.

Nullspace audit over all 32 U coordinates. The extension uses a frozen, previously untouched behavioral block (records 360–439; $n_q = 80$) and fits an independent scalar position decoder over all $K = 32$ coordinates of the fixed basis U. The fixed $[-0.065, +0.065]$ band was specified before both selectivity audits, using the absolute dip change of the earlier full-sequence $\alpha = 0$ potency anchor. It defines the comparison scale for the claim; it was neither derived from the fresh audit nor intended as a universal practical-null boundary. That anchor used records 0–79, whose baseline dip was 0.296, compared with 0.181 on the fresh audit block. The baseline dip is $+0.1811$ $[+0.1057, +0.2609]$. Full-U magnitude removal changes the dip by -0.0407 $[-0.0657, -0.0187]$, whereas a perturbation in the decoder's rank-31 nullspace, matched on support and on per-position norm, changes it by -0.0205 $[-0.0402, -0.0022]$. Its 90% interval is contained in the committed TOST band, but its 95% interval excludes zero, making it an active control rather than a behaviorally inert placebo. The primary evidence is therefore the committed direct paired contrast, -0.0202 $[-0.0312, -0.0109]$. No bootstrap draw crossed zero; the add-one correction gives $p_{\mathrm{MC}} = 1/(B+1) = 4.99975 \times 10^{-5}$ at $B = 20{,}000$, the Monte Carlo resolution floor. The complement placebo changes the dip by -0.0095 $[-0.0292, +0.0101]$. The Holm family ($m = 4$) comprises two directional effect decisions and two interval-containment placebo decisions. The directional pair is full-U magnitude removal against the identity arm and the paired treatment-minus-decoder-null contrast. The placebo pair is the decoder-null perturbation and the complement placebo, each against the committed band. All four pass after correction.

The full-U nullspace perturbation matches realized displacement at 1.0000, has off-U leakage 7.5×10^{-9}, and changes the tested decoded-position readout by 3.8×10^{-8}, compared with 0.558 under treatment. It places 0.887 of its energy in columns 8–32 and uses the same fixed 32-column basis as the headline full-sequence operator; no fallback samples occur. The independent decoder has out-of-sample $R^2 = 0.873$ (train $R^2 = 0.910$). These checks support a treatment–control difference on records 360–439, conditional on this scalar decoder and the fixed 32-column basis U. The separate behavioral contribution of columns 8–32 has not been tested.

Content robustness of the independent decoder. The fixed full-U decoder is evaluated without refitting on disjoint Llama-Instruct MDQA records 440–499 ($n_q = 60$). Prompt-construction checks preserve length and beginning-of-sequence structure and enforce the relevant token-multiset constraints. Its natural-prompt R^2 is 0.689 $[0.666, 0.712]$. Defining survival as control R^2 divided by natural-prompt R^2, iid random-token replacement retains 1.106 $[1.069, 1.147]$ and token-order shuffling retains 1.057 $[1.028, 1.088]$. A token-only diagnostic has $R^2 = -0.044$ $[-0.052, -0.037]$. All four frozen gates pass. Both survival ratios exceed 1, with intervals excluding it. One possible explanation is that content variation in natural prompts partly masks the position readout. Random-token replacement and order shuffling alter this variation while leaving token index intact, which could explain the frozen decoder's slightly higher scores on the controls. The frozen gates place one-sided lower bounds on survival, so ratios above 1 pass. We interpret these ratios as showing no loss of the tested readout, not as added position information. These results support a content-robust absolute-position readout under the tested MDQA controls, which cover random-token replacement, token-order shuffling, and a token-only diagnostic on records 440–499.

Metric	Identity dip	Magnitude change	Magnitude − decoder-null
Normalized EM	0.0625 $[0.0125, 0.1188]$	-0.0500 $[-0.1000, -0.0063]$	-0.0250 $[-0.0500, -0.0063]$
Token F1	0.0298 $[0.0083, 0.0551]$	-0.0153 $[-0.0319, -0.0010]$	-0.0093 $[-0.0208, +0.0001]$

Table 23: Prompt-only greedy-generation validation on the full-U Llama audit block ($n_q = 80$). The magnitude-minus-decoder-null interval excludes zero for normalized EM; the token-F1 contrast has the same direction, but its interval includes zero. Per-arm EM dips are 0.0625, 0.0125, 0.0375, and 0.0500 for identity, magnitude, decoder-null, and complement, respectively. Overall EM ranges from 0.522 to 0.530 across the four arms, reported descriptively. Because the sample is reused and the hook scope differs from the teacher-forced analysis, this is endpoint validation rather than an independent replication.

Band half-width η	full-U decoder-null	complement	Both placebos contained?
0.020	no	no	no
0.030	no	yes	no
0.040	yes	yes	yes
0.050	yes	yes	yes
0.065 (committed)	yes	yes	yes

Table 24: Post-hoc equivalence-band sensitivity for the audit over all 32 U coordinates using archived 90% paired-bootstrap intervals: $[-0.0367, -0.0052]$ for the decoder-null control and $[-0.0260, +0.0069]$ for the complement placebo. "Yes" denotes the nominal 5% TOST criterion: interval containment in $[-\eta, +\eta]$. The minimum containing half-widths are 0.0367 and 0.0260. Only $\eta = 0.065$ and its Holm $m = 4$ family were committed; other rows are descriptive. Both columns use 90% intervals for comparability, matching the decoder-null control's TOST gate. The complement placebo's actual gate instead requires 95% containment. Its 95% interval, $[-0.0292, +0.0101]$, raises that column's minimum containing half-width to 0.0292 without changing any row's verdict. The paired treatment–decoder-null contrast, -0.0202 $[-0.0312, -0.0109]$, does not depend on η.

Greedy-generation accuracy validation. The accuracy audit uses the same records 360–439 and the same identity, full-U magnitude, decoder-null, and complement arms as the full-U selectivity audit. All hooks act on the prompt prefill only; generated tokens are not directly edited. Greedy outputs are scored by normalized exact match (primary) and token F1 (secondary), and uncertainty uses 20,000 paired question-level bootstrap draws. Table 23 reports the endpoint transfer.

Reversal boundary. The prefix-nullspace audit does not extend the decoder-conditioned verdict to reversal under the fixed effect-size rule; the 32-coordinate audit evaluates magnitude, not reversal. Its within-U and complement placebos satisfy their null criteria and its manipulation checks pass, but the disjoint-sample treatment interval does not lie entirely below -0.065. We retain the reversal claim as directional and sample-dependent.

F.3 Rank-one leading-direction ablation

The preceding behavioral battery did not compare intervention ranks within a fixed basis. We therefore tested rank-one ablation directly. The operator ablates the leading direction of the fixed basis, $U_{:1}$, sequence-wide at L^* on Llama-3.1-8B-Instruct, using records 360–439 ($n_q = 80$). This is the third analysis on that block. The identity baseline and both full-U anchor arms are reused unchanged from the decoder-null audit, without re-running them. The rank comparison therefore reuses the comparators' sampling error; it does not provide an independent replication of those arms.

Arms. Rank-one ablation changes the dip by -0.0176 $[-0.0266, -0.0093]$. A displacement-matched control placed in the same decoder's nullspace changes it by -0.0055 $[-0.0144, +0.0017]$, an interval containing zero. No effect is detected for that control, unlike the full-U decoder-null control; this does not establish exact behavioral invariance. All manipulation checks pass. The realized displacement ratio is 1.000, and the decoded readout changes by 8.2×10^{-10} under the control versus 0.569 under treatment. Off-axis leakage is 2.7×10^{-14}, content cosine is 0.9996–1.0000, and next-token loss ratios are 0.988–0.996. These diagnostics use four documents and have no intervals.

Inconclusive treatment–control contrast. The paired treatment-minus-control contrast is -0.0122 $[-0.0249, +0.0022]$. Because the interval includes zero, this result does not establish a rank-one treatment–control difference under the paper's stated rule. The treatment also misses the minimum-

Support	Dose	$\Delta D_{\mathrm{contig}} - \Delta D_{\mathrm{disp}}$ [nominal 95% CI]
0.25	0.5	$+0.065$ $[+0.009, +0.135]$
0.25	1.0	$+0.213$ $[+0.062, +0.415]$
0.25	1.5	$+0.386$ $[+0.162, +0.649]$
0.50	0.5	$+0.083$ $[+0.020, +0.158]$
0.50	1.0	$+0.235$ $[+0.096, +0.410]$
0.50	1.5	$+0.456$ $[+0.223, +0.731]$

Table 25: Paired contiguous-minus-dispersed differences in ΔD on Llama-3.1-8B-Instruct multi-document QA records 200–279 (n_q=80). ΔD is the paired change from identity in the middle-position drop of answer log-probability (Eq. 2). Support is the fraction of context positions edited, and Dose is the calibrated full-residual Δh RMS target. Each row compares one contiguous block against a dispersed mask of the same cardinality at a matched global RMS; the calibrated α and the maximum per-token $\|\Delta h\|$ still differ between the two masks. All six intervals exclude zero at the nominal 95% level, without a familywise correction over the six.

effect rule: its magnitude is roughly a quarter of the ± 0.065 band half-width. For the paired contrast, the exploratory one-sided test gives an add-one-corrected p_{MC} of $915/20001 = 0.04575$; this test was not preregistered for this arm. The archived analysis uses a family of $m = 2$, smaller than the full-U audit's $m = 4$; adjusted p becomes 0.09 at $m = 3$ and 0.14 at $m = 4$. Rank-based tests distinguish the arms (Wilcoxon $p = 0.0002$; sign test $52/80$), but the committed estimator uses a percentile bootstrap on means. We report this disagreement and retain the committed estimator for the conclusion.

Rank ordering, and its single-question dependence. The rank-one minus rank-32 contrast within this fixed basis is $+0.0231$ $[+0.0009, +0.0504]$. The rank-one arm produces 0.43 $[0.19, 0.96]$ of the full-U effect, but this fraction is imprecisely estimated. The rank contrast and the treatment–control contrast are both sensitive to individual questions. Dropping one particular question changes the treatment–control contrast to -0.0171 $[-0.0276, -0.0077]$, whose interval excludes zero. The same omission changes the rank contrast to $+0.0138$ $[-0.0032, +0.0303]$, whose interval no longer excludes zero. Across leave-one-question-out analyses, the zero-exclusion verdict changes for the rank contrast in 11 of 80 cases, and for the treatment–control contrast in one. The headline full-U contrast in Table 4 changes verdict in none of the 80 cases, so this fragility is specific to the rank-one comparisons. The two contrasts also differ in precision. The rank-one treatment and control correlate at -0.31, so pairing increases the standard error by 14%. The full-U pair correlates at $+0.90$, and pairing reduces its standard error by 66%. The precision gain from pairing in the full-U audit therefore does not carry over to this arm.

What the arm does show. Rank-one removal moves the decoded position readout by 0.569, within 2% of the 0.558 the full-U arm produces, while changing behavior by less than half the full-U amount. Thus, similar readout changes accompany different behavioral point estimates within a single basis. The readout values come from separate four-document diagnostics in different runs, so this comparison is descriptive. The intermediate ranks 6–11 come from independently refit subspaces on records 200–279 (Table 27) and cannot be included in this fixed-basis ordering. The within-basis rank comparison therefore covers only two points.

F.4 Global-dose-matched support configuration

The support-configuration analysis uses MDQA records 200–279 ($n_q = 80$), the same behavioral records as the prefix-nullspace audit, and a grid frozen after a separate calibration stage. The full-residual Δh RMS is matched across contiguous and dispersed masks at support fractions 0.25 and 0.5. The paired contiguous-minus-dispersed contrasts are given in Table 25:

Thus total global residual displacement alone is insufficient to explain the contiguous–dispersed difference. Control validity varies across cells: the gold and contiguous complement controls pass the committed 95%-CI containment gate, whereas the dispersed and full-support control intervals are too wide. Paired treatment–control intervals exclude zero in 11 of 18 cells, including every cell whose marginal treatment interval excludes zero, but this does not validate the entire grid. Moreover, global-RMS matching leaves two quantities geometry-dependent: the calibrated α and the maximum per-token $\|\Delta h\|$. The results show dependence on support configuration under matched global dose, but do not isolate a geometric mechanism.

Diagnostic	Result	Interpretation
Fixed-U held-out R^2	0.86–0.88	The fixed subspace continues to decode position on every independent split.
Refit-U held-out R^2	0.53–0.62	Every independently refit subspace decodes position, but less accurately than the fixed subspace.
Fixed-direction ρ	0.94	The fixed direction stays monotone out of sample.
Affinity with fixed U	0.17–0.31	Independently refit subspaces are weakly aligned with the fixed subspace.
Minimum principal angle	29°–42°	No refit subspace comes within 29° of the fixed U.
Selected k_{85}	2–13	The selected rank is sensitive to the ridge penalty.

Table 26: Representation-level stability audit for the learned Llama position subspace (Llama-3.1-8B-Instruct, 4,096-token windows). Stable out-of-sample decoding coexists with low subspace affinity and rank sensitivity, indicating that the particular learned U is not uniquely identified.

Corrected U	k	Magnitude ΔD [95% CI]	Reversal ΔD [95% CI]	Complement placebo [95% CI]
fixed fit @ k_{90}	8	$-0.021\,[-0.040, -0.003]$	$-0.064\,[-0.108, -0.025]$	$+0.000\,[-0.017, +0.017]$
fixed fit @ k_{95}	11	$-0.026\,[-0.046, -0.008]$	$-0.071\,[-0.119, -0.029]$	$+0.002\,[-0.014, +0.019]$
split 0 @ k_{85}	6	$-0.031\,[-0.051, -0.013]$	$-0.037\,[-0.060, -0.014]$	$+0.016\,[+0.002, +0.033]$
split 1 @ k_{85}	7	$-0.038\,[-0.062, -0.016]$	$-0.057\,[-0.090, -0.028]$	$-0.014\,[-0.032, +0.003]$
split 2 @ k_{85}	9	$-0.029\,[-0.045, -0.015]$	$-0.051\,[-0.085, -0.022]$	$-0.003\,[-0.020, +0.013]$

Table 27: Corrected behavioral replay through five coverage-selected prefix subspaces. The raw 95% treatment intervals exclude zero in the negative direction for both arms in all five variants. Every complement placebo satisfies the committed 95%-CI containment gate, including the split-0 placebo, whose interval is nominally positive at $+0.016$ $[+0.002, +0.033]$.

F.5 Position-subspace fit stability

The fixed subspace U is not the only one that decodes position. Independently refit subspaces reach $R^2 = 0.53$–0.62 out of sample, yet have only 0.17–0.31 affinity with U. We fit those alternatives on three independent FineWeb-Edu splits (stream offsets 200, 400, and 600), across ridge penalties $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ and the k_{85}, k_{90}, and k_{95} coverage rules. All nine representation fits are valid (Table 26).

For the corrected behavioral replay, five subspaces were fixed before evaluation: three split-specific k_{85} refits at $\lambda = 10^{-2}$, together with fixed-fit k_{90} and k_{95} variants. Each operator uses its coverage-selected prefix rather than the full 32-column basis. All five bases have orthonormal columns and define distinct projectors that differ from the full-U projector. The replay uses MDQA records 200–279 ($n_q = 80$), the same records as the prefix-nullspace audit: it tests refit robustness rather than providing another disjoint behavioral replication.

After paired placebo subtraction, nominal reversal intervals exclude zero in all five variants, while magnitude intervals do so in four; the fixed-fit k_{90} variant is inconclusive at -0.021 $[-0.044, +0.002]$. These paired intervals describe separation from controls without a separate familywise correction. In no variant does a treatment's 95% interval lie entirely below the committed -0.065 threshold. The earlier full-32 replay, in which the nominal rank did not enter the operator, is retained as an implementation-deviation record and excluded from this verdict.

Together with the two decoder-null audits, these results support a behavioral difference conditional on basis and decoder, without establishing a unique mechanism or one that is stable across refits. For fixed full U, magnitude removal differs from a decoder-null perturbation matched on support and norm on the fresh sample; the prefix audit additionally supplies a near-zero k_{85}-nullspace control. Across five coverage-selected prefixes, the negative direction is consistent, but no treatment interval meets the minimum-effect rule and fixed-fit k_{90} magnitude is not separated from its complement control. Low affinity among valid position decoders is consistent with redundant representation; this audit examines readout subspaces, not the heads that contribute to them.

F.6 Qwen cross-model use audit

A fresh disjoint Qwen2.5-7B-Instruct run used records 160–279 ($n_q = 120$), with the sample size and sequential decision gates fixed before execution. The baseline dip gate passes. The earlier

Sequential gate	Status	Result
Baseline dip present	PASS	CI excludes zero
$\alpha = 0$ ablation potent	PASS	-0.055 $[-0.105, -0.006]$
Complement placebo null	FAIL	$+0.030$ $[-0.016, +0.073]$
NLL and downstream arms	not reached	not interpreted

Table 28: A priori sequential-gate audit in Qwen2.5-7B-Instruct on a fresh disjoint sample (records 160–279; $n_q = 120$). Execution halted at the first failed gate.

Layer	Operator	Instr.	Treat ΔD	Placebo	Contrast	Holm-p	Behavioral
1	ablate	persisted	-0.110 $[-0.167, -0.062]$	-0.011	-0.099 $[-0.160, -0.047]$	0.0004	**supported**
1	reversal	persisted	-0.164 $[-0.259, -0.086]$	-0.044	-0.120 $[-0.212, -0.044]$	0.0010	control effect[c]
2 (ref.)	ablate	persisted	-0.082 $[-0.121, -0.048]$	-0.010	-0.072 $[-0.113, -0.038]$	—	directional[r]
2 (ref.)	reversal	persisted	-0.131 $[-0.195, -0.078]$	-0.008	-0.123 $[-0.186, -0.069]$	—	directional[r]
3	ablate	persisted	-0.073 $[-0.117, -0.035]$	-0.015	-0.059 $[-0.106, -0.018]$	0.0034	**supported**
3	reversal	persisted	-0.076 $[-0.137, -0.022]$	-0.029	-0.047 $[-0.105, +0.007]$	0.0452	control effect[c]

Table 29: Stage 1 of the intervention-layer audit (Llama-3.1-8B-Instruct, records 0–79, n_q=80; design and thresholds frozen in two stages; Holm m=4 over the $l \neq L^*$ cells). Treat ΔD and contrast carry 95% intervals; "persisted" means every persistence readout passes at that cell. [r]Reference-layer rows rerun the archived arms as a pipeline-integrity check. They are excluded from the Holm family by design and are therefore ineligible for the "supported" label. [c]Matched-control effect detected: the sham independently changes behavior in the treatment direction. Under the frozen rule, this prevents value attribution even where the contrast interval excludes zero.

$n_q = 80$ estimate therefore leaves the dip unestablished rather than showing it absent, and because that comparison changes both sample size and record block, it does not isolate power as the cause. The archived Qwen $n_q = 80$ row of Table 22 remains unscored because its baseline dip was not established on that sample. The $\alpha = 0$ ablation is also potent, changing the dip by -0.055 $[-0.105, -0.006]$. The norm-matched complement placebo changes the dip by $+0.030$ $[-0.016, +0.073]$; its 95% interval is not contained in the committed band $[-0.065, +0.065]$.

The audit halts at the complement-placebo gate. The NLL gate is not reached, and downstream value, reversal, and translation arms are not interpreted as either positive or negative evidence. The run establishes a Qwen baseline dip and a potent full-sequence ablation, but not placebo-separated attribution to the fitted position subspace or an independent cross-model behavioral verdict.

The gate also fails if the band is rescaled to Qwen's own $\alpha = 0$ anchor of -0.055: the complement placebo's upper endpoint, $+0.073$, lies outside both bands. Every complement placebo in this paper is gated on 95% containment, including Llama's five coverage-selected prefix complements and the support-audit controls (Appendices F.5 and F.4); decoder-null controls are gated on TOST over the 90% interval.

F.7 Intervention-layer audit and experiment summary

Tables 29 and 30 report the first-stage outcomes and instrument checks; Table 31 gives the disjoint second stage. Table 32 collects the behavioral verdicts across tasks and edit types.

Layer	Operator	k_{85}	Anchor persistence	One-block persistence	NLL ratio
1	ablate	10	$0.951\ (\geq 0.75)$	$0.914\ (\geq 0.652)$	1.024
1	reversal	10	$0.131\ (\leq 0.25)$	$0.145\ (\leq 0.341)$	1.025
2 (ref.)	ablate	7	—	$0.916\ (\geq 0.652)$	0.974
2 (ref.)	reversal	7	—	$0.135\ (\leq 0.341)$	0.955
3	ablate	7	—	$0.850\ (\geq 0.652)$	0.962
3	reversal	7	—	$0.171\ (\leq 0.341)$	0.962

Table 30: Instrument detail behind Table 29. k_{85} is the rank selected for each independently fitted layer subspace under the shared coverage rule; the reference layer's archived and rebuilt values coincide. Rank ratios across layers remain within the preregistered twofold band, so no fixed-rank reanalysis is triggered. Persistence metrics are compared with thresholds fixed before behavioral evaluation. Ablation passes when the realized fraction of the intended signature is high; reversal passes when normalized error relative to the target is low. The anchor column applies only below the reference layer. Displacement matching is exact (median ratio 1.000, 95th-percentile deviation 0.000), and next-token-loss ratios are well below the 1.5 threshold.

Layer/operator	Stage 1	St. 2 placebo	St. 2 contrast	Holm-p	Resolution
1 ablate	supported	$[-0.020, +0.008]$	$-0.083\ [-0.115, -0.052]$	0.0002	**replicated in Stage 2**
1 reversal	control effect	$[-0.027, +0.016]$	$-0.082\ [-0.120, -0.044]$	0.0002	sign-consistent[B]
2 ablate (ref.)	directional	$[-0.002, +0.025]$	$-0.066\ [-0.100, -0.033]$	—	reference
2 reversal (ref.)	directional	$[-0.013, +0.028]$	$-0.071\ [-0.115, -0.028]$	—	reference
3 ablate	supported	$[-0.023, +0.001]$	$-0.019\ [-0.052, +0.014]$	0.259	not replicated (directional)
3 reversal	control effect	$[-0.042, -0.010]^{\dagger}$	$-0.017\ [-0.046, +0.014]$	0.259	control replicated[A]

Table 31: Stage 2 of the intervention-layer audit: the full battery on a disjoint sample (records 80–159, n_q=80; baseline dip 0.250 versus Stage 1's 0.342). Interpretation branches were fixed before the run; Holm correction uses m=4 over the $l \neq L^*$ cells, as in Stage 1. Both Stage 2 columns carry 95% intervals, and reference rows are the archived arms on the disjoint records. †The only placebo interval excluding zero: the sham again changes behavior in the treatment direction. [A]Prespecified branch: matched complement displacement is behaviorally potent at a non-reference layer; placebo-separated value attribution is limited to L^*. [B]Prespecified branch: the treatment retains its sign, but the control result is unstable across samples. No placebo-separated result is claimed at this non-reference layer; the Stage-1 control effect still limits the interpretation.

Behavior · model	value (span)	value (full sequence)	magnitude (span)	magnitude (full sequence)
QA · Llama	descriptive; scope mismatch	directional (2 samples); dependent on sample and U	CI includes zero	**decoder-conditioned behavioral differential**; no re-fit prefix meets the minimum-effect rule
QA · Qwen	consistent with null (no conclusion)	halted at placebo gate	not run	potent; attribution unresolved after halt
QA · Gemma	inconclusive (positive trend)	no conclusion (control)	not run	trend, no conclusion
word list · Llama	ratio-equivalent to null	not run	not run	not run
word list · Qwen	ratio-equivalent to null	not run	not run	not run
sentence preamble · Llama	excluded (content leakage)	not run	not run	not run
sentence preamble · Qwen	weak positive (0.9%)	not run	not run	not run

Table 32: Coverage map of the behavioral intervention design of §7: rows are behavior·model pairs, columns cross value and magnitude edits to the fitted position subspace U at span-restricted and full-sequence support. Cells give committed verdicts; "not run" marks an unexecuted cell. Per-cell results and remaining verdict rules appear in Tables 19, 20, 21, and 22. The minimum-effect rule is not met when the 95% treatment interval does not lie entirely below -0.065. "Halted at placebo gate" means that the complement placebo failed its 95% containment gate, leaving downstream arms uninterpreted. "Decoder-conditioned behavioral differential" denotes a treatment effect larger than that of a matched independent-decoder-null control that also has a detectable effect. The supported quantity is therefore the paired treatment–control difference; the entire treatment effect cannot be attributed to position specifically. Residual-coordinate translation under fixed physical indices is a fifth edit type, not shown here: it shifts decoded position without a corresponding translation of the answer profile (§7).

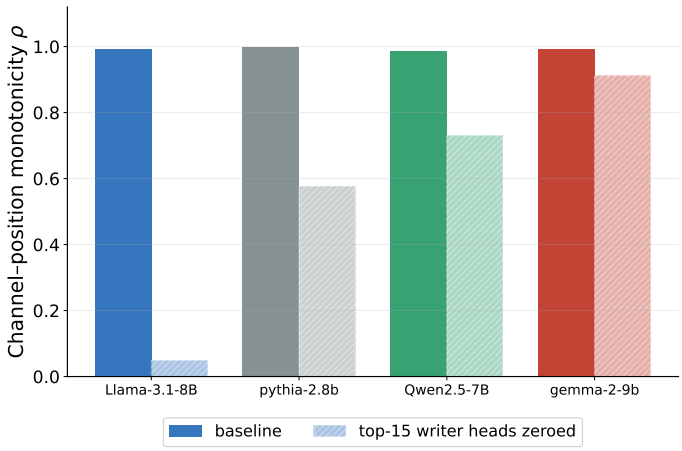


Figure 9: Effect size of top-15 writer-head ablation in the four base models. Each pair compares baseline channel–position monotonicity ρ with the value after the top-15 writer heads are zeroed. The bars come from the archived selection sweep; Table 3 scores ρ damage on the held-out position audit's disjoint documents, and the magnitudes differ between the two runs. The figure shows effect size only; separation from alternative head sets is evaluated on held-out documents using controls matched on layer and head count (Table 3 and Figure 6).

G Writer-Head Effect Sizes

The raw effect size of top-15 writer-head ablation (Figure 9) complements the matched-null comparison in §6: a large raw effect need not imply separation from alternatives matched on layer and head count.