BEYOND ATTENTION MASS: AUDITING THE OUTPUT SENSITIVITY OF SINK CONTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention sinks are token positions that repeatedly attract high attention, raising a basic interpretability question: does prominent attention imply a large effect on the model's output? We study this question through controlled deletions of first-token contributions in decoder-only language models. Matched contributions with the same natural write direction, text, and head produce different local deletion effects across receiving tokens at matched residual magnitude. Same-norm random-direction interventions provide direction-specific control evidence at Qwen3-8B layer 28 and Mistral-7B layer 27. We use readout-Fisher brightness, the variance of a contribution's direct logit perturbations under the unmodified model's predictive distribution, to audit this sensitivity. We then ask whether its predictive utility requires the correct correspondence between token probabilities and logit perturbations. Averaging brightness over random vocabulary permutations yields an analytic control without this correspondence. Adding the control to strong uncertainty and contribution baselines reduces held-out mean squared error for predicting the square root of deletion Kullback–Leibler (KL) divergence by 4.85% and 8.40%, respectively, using ridge regression; vocabulary-aligned Fisher yields a further 1.76% relative reduction in Qwen. A CNN/DailyMail replication corroborates matched deletion contrasts and aggregate ridge gains from the analytic control in both models. Complementary Llama-3.1-8B and Gemma2-9B experiments examine layer dependence and sensitivity under bounded nonlinear readout. Additional Fisher gains remain setting-dependent in predictor-family and eight-site audits. The results distinguish two judgments about sink importance: measuring a contribution's output effect and identifying why a readout score predicts it. The analytic control provides a concrete test of vocabulary-specific predictive utility.

1 INTRODUCTION

Attention sinks are token positions that persistently attract high attention weights across receiving tokens, often including the first token of a sequence (Xiao et al., 2024; Gu et al., 2025). Yet high attention need not imply a large effect on the next-token distribution: routing strength and output sensitivity answer different questions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019).

Prior and contemporaneous studies already describe multiple sink computations and context-dependent behavior (Guo et al., 2024b; Fesser et al., 2026; Choi et al., 2026). Our question is more controlled: does the deletion effect vary across receiving queries when the natural write direction is fixed and the scalar dose is matched? Here dose is the size of the subtracted residual contribution. A receiving query here means a token position whose state predicts the next token, not a user question.

First-token contributions at sink heads provide a controlled setting for this question. Within a fixed text, layer, and head, the projected first-token value is one fixed vector; different receiving queries obtain scalar multiples of it. Matching dose therefore isolates comparisons within a fixed head, source token, and write direction. Figure 1 illustrates this comparison with measured next-token probability changes from one matched pair.

We distinguish three quantities: raw residual dose, direct-readout sensitivity per unit squared norm, and the actual effect of deleting a contribution inside the network. Direct readout-Fisher brightness measures the variance of a contribution's logit perturbations through the unembedding, the linear

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.

map from hidden vectors to vocabulary logits. The variance is weighted by the clean next-token distribution, where clean denotes the model run without intervention. High and low scores per unit squared residual norm define readout-bright and readout-dark events, terminology related to low-output directions in Cancedda (2024). We test whether this direct score helps predict the effects of finite deletions propagated through the remaining network.

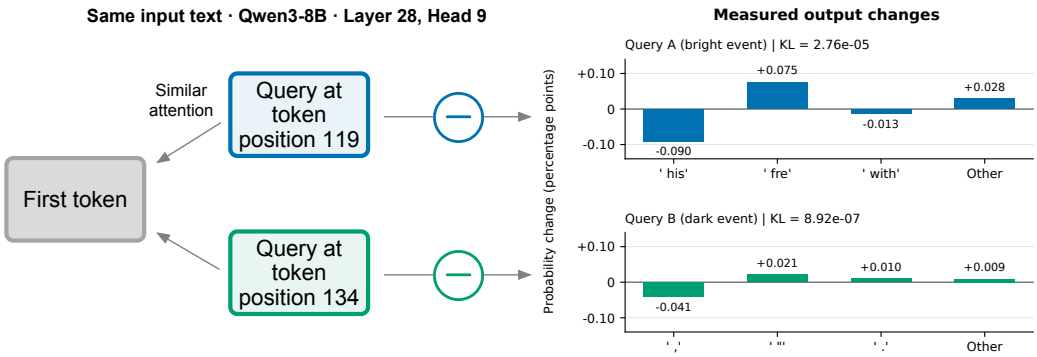

Figure 1: **Similar attention, different output effects.** Two receiving queries in the same text and head attend to the same first-token value direction, with closely matched attention mass and residual dose. Deleting their respective contributions produces different next-token probability changes (Qwen3-8B, layer 28, head 9). Bars show intervened minus clean probabilities in percentage points; Kullback–Leibler (KL) divergence is computed over the full vocabulary. Query positions are zero-indexed token positions in the input text. This illustrative pair was selected near the median deletion-effect gap among eligible dose-matched pairs; it is not an independent estimate of the population effect.

Our first finding is a controlled separation between write magnitude and deletion effect: the same natural sink direction produces different local output changes across receiving queries at matched residual magnitude. Actual random-direction placebos perturb the same layer and query with a random vector of equal norm. Natural-deletion contrasts exceed these controls at Qwen3-8B layer 28 (L28) and Mistral-7B layer 27 (L27; Section 5.1). At the same sites, a fixed-plan CNN/DailyMail follow-up replicates the positive matched natural-deletion contrasts in both models (Section 5.3.1).

We next test brightness on events sampled without consulting the score, rather than only on selected bright/dark pairs. Even if prediction improves, the gain could reflect logit scale and predictive uncertainty rather than the correspondence between vocabulary items and perturbations. We therefore ask a second question: *when Fisher brightness improves prediction, does that gain require matching each token's logit perturbation to its own output probability?*

Our second finding is constructive: an analytic control without vocabulary correspondence improves held-out ridge prediction. We obtain this analytic control by averaging brightness over random vocabulary permutations, retaining logit variance and a measure of how diffuse the predictive distribution is. Adding this representation reduces held-out prediction error by 4.85% in Qwen and 8.40% in Mistral using ridge regression, an L2-regularized linear predictor. Vocabulary-aligned brightness, hereafter real Fisher, yields a further 1.76% [0.64, 3.25] relative error reduction in Qwen (Section 5.2.2). Thus useful predictive features need not preserve vocabulary correspondence, although correctly aligned brightness can retain additional value.

The experiments separate the reach of these findings across corpora, predictors, and sites. Qwen and Mistral carry both central analyses; Llama tests layer dependence, and Gemma2 tests a bounded nonlinear readout. Additional real-Fisher gains remain setting-dependent; Section 5.3 reports the predictive, natural-key, and action-selection scope tests.

Contributions. We contribute (1) controlled evidence of query-dependent deletion effects at matched dose, using natural directions and actual-intervention placebos; (2) an exact analytic permutation control that separates feature utility from vocabulary correspondence; and (3) cross-corpus replication with complementary predictor, site, and model scope tests.

2 Related Work

Sink behavior and contribution measures. Sink studies examine rest states, residual compression, and information transport (Ran-Milo, 2026; Queipo-de Llano et al., 2025; Yoo et al., 2026). Fesser et al. (2026) distinguish adaptive no-op and broadcast computations, while Choi et al. (2026) examine global information in multimodal sink representations and task-dependent effects of key scaling. These analyses concern the computation performed or the information carried; our matched-event estimand is the local output effect of subtracting a specified natural contribution. Value-aware importance measures incorporate value magnitude (Guo et al., 2024c), whereas contribution weights additionally account for alignment with head or attention output (Cunningham & Cirone, 2026). This distinction motivates including their signed and absolute scores as competing predictors in Section 5.2.1.

Readout diagnostics and Fisher geometry. Logit-lens methods decode intermediate states, while direct logit attribution examines contributions to individual token logits (nostalgebraist, 2020; Elhage et al., 2021; Belrose et al., 2023). Unlike token-specific direct logit attribution, our score aggregates perturbations under the clean predictive distribution. Park et al. (2023) develop a non-Euclidean inner product for concept representations; our object is an individual residual contribution evaluated under a context-specific output distribution. Fisher-based steering and intervention-design work (Wang & Zhao, 2026; Zhou et al., 2026) uses geometry to choose interventions, whereas our protocol scores and deletes contributions already produced by the model.

Interpretation and action. Head-pruning and key/value (KV) cache-retention studies evaluate which components can be removed or retained under an action budget (Sok et al., 2026; Sandoval-Segura et al., 2025; Bui et al., 2025); this policy-level target differs from predicting single-event deletion KL. Work on subspace-intervention illusions and mechanism-recovery benchmarks motivates testing whether an intervention-based explanation identifies its proposed object (Makelov et al., 2023; Mueller et al., 2025). Our control tests the interpretation of predictive feature gains rather than recovering a causal circuit or optimizing an intervention policy. Appendix B provides the full discussion, including multimodal interventions, attention routing, and information-geometric connections.

3 Preliminaries: Direct Readout-Fisher Brightness

Let $h_{\ell,t} \in R^d$ be a residual stream state at layer ℓ and token position t, where d is the hidden dimension. Let $\mathcal{V}$ denote the vocabulary and $W_U \in R^{|\mathcal{V}|\times d}$ the unembedding. At event $e = (x, \ell, h, q)$, x is the input text, h the head index, and q the receiving query position. Let z_e be the clean model's actual final logits at q and $p_e = \mathrm{softmax}(z_e)$. We write p when the event is understood; it is not a distribution obtained by decoding the intermediate residual state. For a residual direction v, the direct logit perturbation is $W_U v$. The Fisher information of the categorical output in logit coordinates is $\Sigma_p = \mathrm{diag}(p) - pp^\top$. We define direct readout brightness as

$$B_p(v) = v^\top W_U^\top \left(\mathrm{diag}(p) - pp^\top\right) W_U v = \mathrm{Var}_{i\sim p}\left[(W_U v)_i\right]. \tag{1}$$

Equation 1 is the predictive variance of the induced logit perturbation, not a finite output KL. We call $B_p(v)$ total brightness and $B_p(v)/\|v\|_2^2$ unit brightness, using the latter to rank bright/dark event candidates. Unless a normalization-aware or soft-cap-aware variant is explicitly specified, these scores use W_U alone, without the final normalization or downstream Jacobian.

Direct-readout geometry. We use the standard categorical Fisher matrix (Park et al., 2026), rather than propose a new metric; it has a null direction for constant logit shifts. Pulling it back through W_U gives the quadratic term for a direct logit perturbation: its infinitesimal KL is $\alpha^2 B_p(v)/2 + o(\alpha^2)$. That relationship is definitional; whether brightness predicts a finite deletion propagated through the remaining network is an empirical question.

Sink event directions. For layer ℓ, head h, query position q, and first-token key position 0, we define a sink event contribution

$$v_{x,\ell,h,q} = a_{\ell,h}(q,0)\, W_O^{\ell,h} V_{\ell,h}(x_0), \tag{2}$$

where $a_{\ell,h}(q,0)$ is the attention weight from query q to the first token, $V_{\ell,h}(x_0)$ is the value vector at the first token, and $W_O^{\ell,h}$ denotes the head-specific slice of the output projection. Within a fixed

text, layer, and head, the write direction $W_O^{\ell,h} V_{\ell,h}(x_0)$ is fixed; queries vary in scalar sink mass and predictive distribution.

Dose, coordinate, and network effect. Write $u = W_O^{\ell,h} V_{\ell,h}(x_0)$, $v_e = a_e u$, $D_e = \|v_e\|_2$, and $b_e = B_{p_e}(u)/\|u\|_2^2$ for nonzero u. Then

$$B_{p_e}(v_e) = D_e^2 b_e, \qquad K_e(\alpha) = \mathrm{KL}\Big(p_e \,\|\, p_e^{(-\alpha v_e)}\Big). \tag{3}$$

The first equality is exact: it separates scalar raw dose from direct-readout sensitivity per unit squared norm. It explains why matching mass and norm controls one scalar factor within a fixed text/head, not two independent mechanisms. The second quantity is the Kullback–Leibler (KL) divergence from the clean next-token distribution to the distribution after subtracting the natural contribution inside the network; it is not defined by brightness. For a differentiable downstream computation with Jacobian J_e from the intervention site to the final logits,

$$K_e(\alpha) = \frac{\alpha^2}{2} v_e^\top J_e^\top \Sigma_{p_e} J_e v_e + o(\alpha^2). \tag{4}$$

This follows from the local categorical KL expansion. Replacing J_e by W_U recovers the direct-readout surrogate, not a theorem about the full network. Both p_e and J_e may change across queries; our matched comparisons do not isolate a causal effect of changing p_e alone.

Estimand and interpretation. The intervention estimand is the paired bright-minus-dark difference in $K_e(\alpha)$ after matching within text, layer, and head on scalar dose and measured query descriptors. It concerns the affected query, not a sequence-wide average. Deleting a specified contribution identifies its effect; comparing those effects across contexts does not identify a causal effect of brightness itself or eliminate every contextual confound. Direct and propagated KL are distinct counterfactual measurements, not additive causal components. Throughout, output functionality means distributional sensitivity, not beneficial task contribution: $B_p(v) = B_p(-v)$, and a large KL alone does not indicate whether prediction improves. The diagnostic therefore returns context-indexed scores and intervention contrasts, not fixed functional labels for heads.

4 Experimental Setup

Complementary comparisons. We call a layer–head pair a site; each primary setting comprises the audited heads at Qwen L28 or Mistral L27. Matched deletions compare a fixed source direction across receiving queries; score-blind prediction samples events without consulting brightness and evaluates on held-out texts. The stronger-baseline and analytic-control reanalyses reuse the original prediction data, whereas the outcome-blind site audit fixes new sites before measuring outcomes (Section 5.3.2).

Models and data. We study Qwen3-8B, Mistral-7B-v0.3, Llama-3.1-8B, and Gemma2-9B on WikiText-2 text segments of length 512. The fixed-plan cross-corpus follow-up additionally evaluates Qwen and Mistral on CNN/DailyMail news articles at the same maximum length (Section 5.3.1). Qwen and Mistral carry both primary designs. Llama tests localization sensitivity; Gemma2 tests a nonlinear, bounded readout with tied embeddings and final-logit soft-capping (Team et al., 2024). Tied embeddings share input-embedding and output-readout weights; soft-capping smoothly bounds the final logits with a scaled hyperbolic tangent. Unless stated otherwise, all interventions use teacher-forced forward passes: the model receives the recorded text prefix rather than its own sampled continuation. Precision-controlled runs use 32-bit floating-point (fp32) weights; whether TensorFloat-32 (TF32) matrix multiplication is enabled is a separate execution setting. For the main within-head comparisons, we report 95% confidence intervals (CIs) from a text-cluster bootstrap: complete texts are resampled while their intervention pairs remain grouped. Auxiliary readout-control audits report matched-pair gaps, with text-cluster intervals when multiple seeds are aggregated.

Metrics. The primary intervention outcome is local baseline-to-intervened KL at the affected query. Sequence-averaged KL dilutes this effect over other positions; negative log-likelihood (NLL) of the observed next token is secondary because a deletion can help or harm prediction. Supplementary geometry analyses measure cosine over-mixing, cone angle, and spectral rank (Appendix A.12).

Layer localization. The original localization used brightness audits and head-level first-token suppression with same-count and mass-matched controls, focusing on Qwen L26 and Mistral L27.

The primary Qwen event evidence instead uses the subsequent seed-split-selected L28 confirmation; these selected sites are not a representative sample of layers. Llama's pre-specified causal rule and the complete localization details are in Appendix A.4. Head-level suppression redistributes attention through softmax renormalization, unlike the value-only event subtraction below; the two interventions identify different actions.

Within-head event intervention. Within each text, layer, and head, we rank events from Equation 2 by unit brightness. We pair readout-bright and readout-dark candidates by nearest-neighbor matching on log attention mass, log raw residual norm, clean predictive entropy, and normalized query position. Unmatched candidates are discarded. Mass and norm control one collinear dose factor because the natural direction is fixed. The original matched panels have aggregate attention-mass and raw-norm log-ratios within ± 0.1; this aggregate balance does not guarantee a per-pair caliper, meaning a maximum allowed matching difference. For the Qwen layer-28 seed-split confirmation, we also report a strict re-slice of the already-run intervention pairs. Every retained pair has an absolute bright/dark log-ratio of at most 0.1 for both attention mass and residual norm. We subtract the single-head projected contribution at the attention module's forward output, before residual addition and the same-layer multilayer perceptron (MLP). For grouped-query attention (GQA), multiple query heads share a key/value (KV) head; subtraction follows the query-head-specific output projection and does not alter shared cache entries. We then apply

$$h'_{\ell,q} = h_{\ell,q} - \alpha v_{x,\ell,h,q}, \qquad \alpha \in \{0.25, 0.5, 1.0\}, \tag{5}$$

and measure local and global output changes. Matching controls residual norm rather than brightness itself, avoiding the tautology of giving dark events a larger raw perturbation to equalize the direct Fisher score.

5 Results

The results answer two questions in sequence: can matched natural writes have different deletion effects, and does predicting those effects identify the role of vocabulary correspondence? We then test scope along separate axes: a new corpus at the original sites, outcome-blind sites and held-out layers, and additional model and action settings.

5.1 Controlled natural deletions reveal context-dependent effects

At fixed text/head, variation in b_e reflects the receiving query's predictive distribution, not a changing first-token vector (Equation 3). To measure its prevalence before bright/dark pairing, we compute the ratio of the 90th to the 10th percentile of unit brightness within each text/head group. The full candidate pools at Qwen L28 and Mistral L27 have median ratios across these groups of $1215\times$ and $131\times$, respectively. Random pairs within the same text/head exceed a $10\times$ brightness ratio with probabilities 42.2% and 33.5%. Table 16 reports the sampling frame and the other audited sites. These statistics describe score variation, not the prevalence of causal-effect differences.

Matched natural deletions. A head-level contrast could result from selecting different heads; the event experiment instead fixes text, layer, head, and first-token write direction. Bright and dark events are matched on raw dose and measured query descriptors, then deleted one at a time through the actual model. Qwen L28 supplies the strict pair-caliper fp32 confirmation; Mistral L27 supplies a complementary matched comparison with a separate fp32 repeat (Table 1). We distinguish the original aggregate balance from a per-pair caliper, rather than treating all historical panels as identically matched.

Natural direction versus generic perturbability. Different contexts may be sensitive to many directions, not specifically the natural sink write. The actual placebo therefore applies a same-norm random residual perturbation at each selected event's layer and query, and compares the bright–dark contrast with that of natural deletion. At Qwen L28, the random-direction contrast is 1.53×10^{-4} $[0.97, 2.23]\times 10^{-4}$, and the actual-minus-random contrast is 4.08×10^{-4} $[2.85, 5.49]\times 10^{-4}$. The Mistral L27 placebo audit gives an actual-minus-random contrast of 1.88×10^{-4} $[1.36, 2.44]\times 10^{-4}$ in its original matched panel (Table 11); it is not a placebo comparison on the separate fp32 repeat. The natural directions therefore contribute a larger bright–dark effect contrast than the tested same-norm random controls at these selected sites.

Setting	Measurement panel	Local KL gap [95% CI]
Qwen L28	fp32, strict dose caliper	5.60 [4.30, 7.11]
Mistral L27	Original matched panel	3.19 [2.70, 3.75]
Mistral L27	Separate fp32 repeat	2.98 [2.57, 3.45]

Table 1: Bright-minus-dark local deletion KL at $\alpha = 1$, in units of 10^{-4}. Intervals resample text clusters. These rows retain their respective matching and precision settings; they are not pooled into a common effect estimate. Table 8 reports the original-panel and Qwen strict-audit counts; the Mistral fp32 repeat is a separate precision check, not the panel used for the original placebo contrast.

What the interventions establish. Controlled-dose deletion reveals output-effect heterogeneity within a fixed natural direction. Its possible sources include both the predictive distribution and the downstream Jacobian in Equation 4; the experiment measures their combined effect rather than isolating brightness as a cause. Local KL is the appropriate measurement for this single-query intervention; a smaller sequence-averaged KL does not negate that local contrast, but neither metric establishes useful task information. Historical cross-model plots and dose-balance summaries are retained in Appendix A.18; the Qwen L26 point in those plots is not the current L28 confirmation.

5.2 What do predictive gains identify?

The deletion experiment establishes effect heterogeneity, but brightness selected the pairs. We therefore first test its predictive utility on score-blind events, then ask whether that utility requires matching each vocabulary item's logit perturbation to its own probability.

5.2.1 Score-blind prediction beyond a strong baseline

To separate incremental prediction from brightness-based pair selection, we fix Qwen3-8B L28 and Mistral-7B L27 and sample first-token events without consulting scores or intervention outcomes. For each model, 64 WikiText-2 training-split articles calibrate the diagnostic regressions and 64 disjoint articles evaluate them, with normalized-content and token-prefix deduplication. Each article contributes eight sampled queries per audited head, yielding 2,048 calibration and 2,048 evaluation events per model and dose. The sampling frame includes all eligible first-token events at these heads, not only events above a high-attention threshold. These prediction experiments use fp32 weights with TF32 matrix multiplication disabled. Each event is independently deleted at $\alpha = 1$ (primary) and 0.25 (secondary).

The target is $\sqrt{K_{\text{actual}}}$, where $K_{\text{actual}} = K_e(\alpha)$ is the measured deletion KL. Prediction risk is text-equal mean squared error (MSE): squared errors are averaged within each evaluation text, then equally across texts. Relative MSE reduction is $100(R_{\text{baseline}} - R_{\text{extended}})/R_{\text{baseline}}$; positive values mean lower prediction error, not improved language-model accuracy. Pooled reduction computes this ratio after aggregating prediction risks across heads or sites; head-equal (macro) reduction instead averages their separately computed relative reductions with equal weights. We organize prediction around three comparisons: (1) strong baseline versus baseline+Fisher; (2) strong baseline versus baseline+analytic permutation expectation S; and (3) baseline+S versus baseline+S+Fisher. They test incremental utility, utility without vocabulary correspondence, and residual utility conditional on that control, respectively; none is a causal decomposition.

Comparison 1: Fisher beyond a strong baseline. The strong baseline covers four groups: dose, clean-state uncertainty, signed and absolute contribution weights (Cunningham & Cirone, 2026), and simple readout descriptors. These include predictive concentration and uniform-vocabulary logit variance, so the comparison tests brightness beyond elementary uncertainty and logit-scale features; the complete feature list is in Appendix A.1.1. The extended predictor adds direct-readout total and unit brightness. Ridge uses standardized transformed features, individual squared terms, head indicators, and calibration-text cross-validation; evaluation outcomes are not used in fitting. We compare ridge and histogram gradient boosting (HGB), with article-grouped calibration cross-validation and calibration/evaluation bootstrap refitting. Calibration refers to fitting the diagnostic predictor, not updating the language model; each joint bootstrap refit repeats predictor fitting and hyperparameter selection on resampled calibration texts. Table 2 reports the joint-refit analysis: the

Predictor	Qwen L28	Mistral L27
Ridge	6.21 [4.26, 9.38]	9.82 [6.91, 12.63]
HGB	4.15 [−5.56, 10.52]	−3.07 [−9.54, 5.25]

Table 2: Primary-site relative MSE reduction (%) from adding Fisher total/unit to the strengthened baseline, for $\sqrt{K_{\text{actual}}}$ at $\alpha = 1$. Intervals use 200 joint calibration/evaluation article-bootstrap refits, including calibration model selection, rather than the fixed-fit intervals in Table 5.

Setting	Predictor	Add analytic S	Add B given S
Qwen L28	Ridge	4.85 [3.25, 7.85]	1.76 [0.64, 3.25]
Mistral L27	Ridge	8.40 [5.73, 10.82]	1.83 [−1.32, 3.09]
Qwen L28	HGB	0.44 [−8.56, 5.29]	3.32 [−1.91, 9.74]
Mistral L27	HGB	0.58 [−4.52, 6.97]	−0.49 [−8.21, 6.28]

Table 3: Analytic permutation-expectation audit in the primary settings. Entries are relative MSE reductions (%) for $\sqrt{K_{\text{actual}}}$ at $\alpha = 1$, with 200 calibration/evaluation article-bootstrap refits. The first contrast uses baseline risk as denominator; the second uses baseline+S risk. They are not additive components of a causal explanation.

ridge increment remains positive, whereas the HGB intervals cross zero. These reanalyses reuse the original evaluation articles and are not independent prospective confirmations.

Shuffled Fisher features also improve ridge prediction, motivating the next comparison: a score may be useful even after its vocabulary correspondence is broken. Brightness combines logit spread, predictive concentration, and the correspondence between vocabulary items and perturbations. A direction can have large overall logit variation but little variation among likely tokens. Because spread and concentration can themselves predict deletion sensitivity, a Fisher gain does not isolate correspondence. We therefore hold events, outcomes, probabilities, and perturbation values fixed, but randomly reassign perturbations across vocabulary items. Their permutation expectation supplies the control below. Appendix A.1 reports the unit-only and finite-shuffle sensitivity results and the earlier smaller-baseline analysis.

5.2.2 An analytic control for vocabulary correspondence

What does a vocabulary shuffle preserve? To determine what survives the shuffle, we compute brightness's expectation over vocabulary permutations. For uniform random vocabulary permutation π, writing $s = W_U v$ and V for vocabulary size gives

$$E_\pi Var_{i\sim p}(s_{\pi(i)}) = \frac{V}{V-1} Var_{\text{uniform}}(s) \left(1 - \sum_i p_i^2\right). \tag{6}$$

Here $1 - \sum_i p_i^2$ is predictive dispersion, the complement of probability collision: it is small for a concentrated distribution and larger for a diffuse one. Thus shuffled scores preserve a combination of directional logit scale and predictive dispersion; they are not information-free noise. To obtain Equation 6, center s by its uniform mean and let its uniform variance be σ^2. Under permutation, each coordinate has second moment σ^2 and each distinct pair has cross-moment $-\sigma^2/(V-1)$; expanding the weighted variance yields the expression. Our finite average of 16 permutations approximates, but is not equal to, this expectation.

Comparison 2: utility without vocabulary correspondence. We define S_e as the right-hand side of Equation 6 and use total and unit versions, S_e and $S_e/\|v_e\|^2$. We test whether this representation improves prediction and whether real Fisher B adds value beyond it. We compare baseline, baseline+S, and baseline+S+B, alongside dispersion-only and finite-shuffle controls, using the same events and article splits. This analysis was designed after inspecting the preceding results; it is explanatory and post-hoc, not independent confirmation.

Comparison 3: real Fisher conditional on the analytic control. Table 3 establishes ridge utility of the analytic representation in both primary settings. After adding the analytic control, real Fisher further improves Qwen ridge prediction. The other model–predictor conditional intervals cross zero,

placing the resolved additional gain in the Qwen ridge setting. In particular, a positive interval in Qwen and a zero-crossing interval in Mistral do not themselves establish a statistically resolved difference between the models. Adding finite-shuffle features conditional on S gives ridge gains of -0.17% [-1.87, 1.03] in Qwen and -0.04% [-0.83, 0.38] in Mistral, neither establishing an advantage nor proving equivalence.

Because S is determined by existing descriptors, its gain demonstrates utility of an explicit feature representation for the fitted predictor, not additional information beyond those descriptors. Feature transformations and regularization can change that utility; the comparison does not isolate a physical interaction mechanism or an "explained percentage" of deletion effects. Appendix A.15 reports the transformation details, dispersion-only comparisons, and refitting sensitivity.

5.3 Scope across corpora, sites, and actions

We test the reach of the WikiText-2 findings along two distinct axes: replication on another corpus at the original sites, and prediction at outcome-blind sites and held-out layers.

5.3.1 New corpus, original sites

Do the two findings recur beyond WikiText-2? We repeat the score-blind audit on CNN/DailyMail articles at the same Qwen L28 and Mistral L27 sites, fixing the sampling and analysis plan before collecting the new outcomes. Each model uses 64 calibration and 64 evaluation articles; predictors are refitted on the new calibration split, so this tests cross-corpus replication rather than frozen-predictor transfer. Secondary strict-dose pairs yield positive bright-minus-dark local-KL gaps in both models: $1.73\,[0.86, 2.86] \times 10^{-4}$ in Qwen and $1.04\,[0.42, 1.89] \times 10^{-4}$ in Mistral at $\alpha = 1$. These contrasts cover 17.77% and 27.54% of evaluation events and do not repeat the random-direction placebo.

The predictive distinction also persists: adding analytic S reduces pooled ridge MSE by 3.42% [2.60, 7.19] and 4.09% [1.98, 9.24], respectively. Real Fisher added after S gives 2.13% [-0.73, 5.18] in Qwen and 2.18% [0.23, 5.39] in Mistral. The latter interval is positive, but both head-equal conditional intervals and all corresponding pooled HGB intervals cross zero. Thus the new corpus corroborates natural-deletion heterogeneity and pooled ridge representation utility, not a uniform or predictor-independent Fisher advantage (Appendix A.14).

5.3.2 Outcome-blind sites and held-out layers

The preceding prediction tests retain sites selected during earlier intervention studies. We now ask whether their predictive increments transfer to outcome-blind sites and unseen layers; this is a separate generalization question from the matched deletion contrast. To test dependence on the originally selected settings, a prospective protocol fixes depths $\lfloor r(L-1) \rfloor$, $r \in \{0.25, 0.5, 0.75, 0.9\}$, where L is the number of layers, and deterministically samples two heads per depth using a fixed hash before observing scores or outcomes. For Qwen the layers are 8, 17, 26, and 31; for Mistral they are 7, 15, 23, and 27. We sample 64 calibration and 64 evaluation articles per model, excluding 125 articles represented in the supplied prior manifests, with two score-blind queries per site/article and $\alpha = 1$ in fp32. This yields 2,048 deletions per model; all sampled sites are retained, without requiring that every head have strong sink attention. The protocol is prospective for this follow-up, not a claim that its design predates the preceding results.

The new-text test trains on calibration articles from all sampled sites. The held-layer test trains only on the other three layers' calibration articles and evaluates the omitted layer on evaluation articles. All feature sets receive relative depth but no head-identity indicators; the stronger baseline and shuffle controls are retained. Our primary summary gives equal weight to each site's relative MSE reduction, rather than letting a high-risk site dominate pooled risk. Bootstrap draws resample whole articles jointly across sites and refit calibration predictors, including hyperparameter selection.

Table 4 shows limited new-text evidence but no stable held-layer advantage. New-text ridge point estimates are positive at seven of eight sites in each model, but only Mistral's equal-site aggregate interval excludes zero; this is not seven significant sites. Qwen's secondary pooled new-text ridge reduction is 2.50% [0.76, 9.55], which does not replace its unresolved primary equal-site result. For Mistral, shuffle alone improves new-text ridge prediction by 5.77% [1.09, 14.18]; adding real Fisher conditional on shuffle yields 3.72% [-0.86, 7.08]. All eight model/predictor/regime intervals for the

Model	Predictor	New text	New text + held layer
Qwen	Ridge	3.04 [−2.53, 11.73]	−1.48 [−17.16, 7.66]
Qwen	HGB	5.34 [−17.62, 11.27]	−6.15 [−26.50, 13.92]
Mistral	Ridge	6.77 [0.45, 13.55]	−6.99 [−29.01, 16.38]
Mistral	HGB	5.30 [−6.22, 12.43]	−1.40 [−17.14, 11.90]

Table 4: Outcome-blind site audit: site-equal relative MSE reduction (%) from adding Fisher total/unit to the strong baseline. Target: $\sqrt{K_{\text{actual}}}$. The 95% intervals use 200 calibration/evaluation article-bootstrap refits and are conditional on the eight sampled sites, not confidence intervals over the population of layers or models. Negative point estimates with zero-crossing intervals do not establish harm.

increment conditional on shuffle cross zero. The resolved gains are confined to particular new-text comparisons; the held-layer and conditional intervals remain inconclusive, rather than establishing equivalence or absence of information. The analytic-control reanalysis reaches the same bounded conclusion: all eight model/predictor/regime intervals for real Fisher conditional on S cross zero (Appendix A.15).

5.3.3 Additional models and action boundaries

Beyond the complete Qwen/Mistral audits, Llama's weak pre-specified L20 and stronger exploratory L31 characterize layer dependence, while Gemma2's soft-capped-readout contrast tests sensitivity under bounded nonlinear readout (Appendix A.3). Normalization and small-perturbation checks validate the local readout calculation, not equality with the full downstream Jacobian (Appendices A.9 and A.10). Natural-key comparisons do not establish a first-token-specific predictive advantage, and selection/eviction probes do not establish a policy advantage over the tested baselines; full results, coverage, and numerical limitations remain in Appendices A.17 and A.16, and Tables 18 and 19.

6 Limitations

(1) **Causal scope.** Deleting a contribution identifies its output effect, not a causal effect of brightness independently of context. Predictive distributions and downstream Jacobians both vary across queries; normalization and finite-difference checks do not identify a propagation mechanism. (2) **Sampling and generality.** Selected primary sites and score-based pairs do not establish population prevalence. The eight-site audit covers a fixed depth grid, Llama L31 is exploratory, and natural-key controls have low coverage. Neither sink-exclusive sensitivity nor broad cross-layer predictive superiority is established; all experiments are teacher-forced. (3) **Conditional predictive evidence.** Gains depend on the baseline and predictor. The original analytic-control analysis is post hoc; the fixed-plan news replication refits predictors at the same sites and is not proof that vocabulary alignment is irrelevant. Calibration-refit intervals retain Monte Carlo uncertainty, and some controls use observed next-token information. Historical precision settings also affect effect magnitudes, motivating the separate primary Qwen L28 audit. (4) **No established policy advantage.** Output KL measures distributional change, not task benefit. Selection and eviction probes do not establish an advantage over the tested policies, so the diagnostic does not certify pruning, compression, or steering utility. Appendix C retains the detailed coverage, numerical, and statistical qualifications.

7 Conclusion

This study establishes two distinctions in sink analysis. Matched natural writes produce different deletion effects across receiving queries, and a representation without vocabulary correspondence improves pooled ridge prediction of those effects. The natural-deletion contrasts and analytic-control gains recur on news articles; actual random-direction controls strengthen the original-site intervention evidence. Additional real-Fisher gains remain corpus-, model-, and predictor-dependent. These results support a concrete audit protocol: control dose when measuring output effects, then compare baseline, baseline+S, and baseline+S+Fisher when attributing predictive utility to vocabulary correspondence.

AI USE STATEMENT

In this work, we used generative AI tools for grammar correction and rewriting. We have not used generative AI tools for generating synthetic numerical results to substitute for measured experimental outcomes, and AI-assisted recruitment or evaluation of human participants are not applicable to this work. Additionally, we used generative AI tools for literature discovery, manuscript organization and language editing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Ngoc Bui, Shubham Sharma, Simran Lamba, Saumitra Mishra, and Rex Ying. Cache what lasts: Token retention for memory-bounded kv cache in llms. *arXiv preprint arXiv:2512.03324*, 2025.

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.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2026.

Harry Jake Cunningham and Nicola Muca Cirone. Contribution weights: A geometrical analysis of self-attention transformers. *arXiv preprint arXiv:2606.07604*, 2026.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A simple and effective l_2 norm-based strategy for kv cache compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18476–18499, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Lukas Fesser, Mozes Jacobs, Thomas Fel, Andy Keller, and Sham Kakade. A unifying view of attention sinks: Two algorithms, two solutions. *arXiv preprint arXiv:2606.08105*, 2026.

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 *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*, 2024a.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*, 2024b.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in kv cache reduction: Value also matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21158–21166, 2024c.

Feijiang Han, Xiaodong Yu, Jianheng Tang, Delip Rao, Weihua Du, and Lyle Ungar. Zerotuning: Unlocking the initial token's power to enhance large language models without training. *arXiv preprint arXiv:2505.11739*, 2025.

Zhengfu He, Junxuan Wang, Rui Lin, Xuyang Ge, Wentao Shu, Qiong Tang, Junping Zhang, and Xipeng Qiu. Towards understanding the nature of attention with low-rank sparse decomposition. *arXiv preprint arXiv:2504.20938*, 2025.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. The spacetime of diffusion models: An information geometry perspective. *arXiv preprint arXiv:2505.17517*, 2025.

Hyunwoo Kim, Kisu Lee, Yuna Shin, and Ha Young Kim. Diva: Instruction-aware vision token pruning via dual-probe attention discrepancy. In *European Conference on Computer Vision*, pp. 76–95. Springer, 2026.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

nostalgebraist. Interpreting GPT: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, 2020. URL `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. *arXiv preprint arXiv:2602.15293*, 2026.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

Yuval Ran-Milo. Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks. *arXiv preprint arXiv:2603.11487*, 2026.

Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. Identifying and evaluating inactive heads in pretrained llms. *arXiv preprint arXiv:2504.03889*, 2025.

Piyush Sao. Ghosts of softmax: Complex singularities that limit safe step sizes in cross-entropy. *arXiv preprint arXiv:2603.13552*, 2026.

Seungjun Shin, Jaehoon Oh, and Dokwan Oh. Orthorank: Token selection via sink token orthogonality for efficient llm inference. *arXiv preprint arXiv:2507.03865*, 2025.

Jaewon Sok, Jewon Yeom, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Garbage attention in large language models: Bos sink heads and sink-aware pruning. *arXiv preprint arXiv:2601.06787*, 2026.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia

Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL `https://arxiv.org/abs/2408.00118`.

Sihan Wang and Jiayi Zhao. Fishback: Pullback fisher geometry for optimal activation steering in transformers. *arXiv preprint arXiv:2605.17231*, 2026.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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. 11–20, 2019.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Qingyue Yang, Jie Wang, Xing Li, Yinqi Bai, Xialiang Tong, Huiling Zhen, Jianye Hao, Mingxuan Yuan, and Bin Li. Why attention patterns exist: A unifying temporal perspective analysis. *arXiv preprint arXiv:2601.21709*, 2026.

Suho Yoo, Youngjoon Jang, and Joon Son Chung. On the nature of attention sink that shapes decoding strategy in omni-llms. *arXiv preprint arXiv:2603.14337*, 2026.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A'catch, tag, release'mechanism for embeddings. *Advances in Neural Information Processing Systems*, 38:83140–83181, 2026.

Hanhan Zhou, Shamik Roy, and Rashmi Gangadharaiah. Steering without breaking: Mechanistically informed interventions for discrete diffusion language models. *arXiv preprint arXiv:2605.10971*, 2026.

A Additional Experimental Details

A.1 Initial score-blind prediction protocol

The original score-blind baseline used dose and state descriptors, signed and absolute contribution weights, uniform-vocabulary logit variance, and the top-two direct-logit contrast, before the stronger

Site	Baseline MSE	+Fisher MSE	Improvement [95% CI]	Relative
Qwen L28	37.75	34.82	2.93 [2.06, 3.89]	7.75%
Mistral L27	30.51	26.62	3.89 [2.71, 5.20]	12.76%

Table 5: Original smaller-baseline score-blind prediction at $\alpha = 1$. MSE and absolute improvements are in units of 10^{-6} for $\sqrt{K_{\text{actual}}}$. Both models meet the original rule. Intervals resample the 64 evaluation texts conditional on fitted calibration predictors and do not include calibration-training variance.

probability-shape controls in Section 5.2.1. The original success rule required a positive fixed-fit text-bootstrap interval and at least 5% relative MSE reduction, separately per model. Table 5 preserves those results and their original uncertainty convention; it does not supersede the strengthened-baseline and calibration-refit analyses in the main text.

At the secondary dose $\alpha = 0.25$, relative improvements were 7.87% and 13.09%, respectively. These historical estimates retain their sample, baseline, and fitting protocol; later estimates are not substituted for them.

A.1.1 Strengthened baseline and feature sensitivity

The strengthened baseline includes dose and clean-state descriptors, signed and absolute intra/inter-head contribution weights (Cunningham & Cirone, 2026), uniform-vocabulary logit variance, and a top-two direct-logit contrast. It also includes maximum probability, top-5/top-20 probability mass, probability collision $\sum_i p_i^2$, clean logit gaps and variance, attention entropy, and source-value norm. The extended predictor adds total and unit brightness, rather than replacing the baseline descriptors. Unit brightness alone does not have a positive refit interval in both models.

Finite-shuffle refitting sensitivity. In the strengthened-baseline analysis, the Mistral increment conditional on finite shuffle has the same 2.13% point estimate but intervals of $[-1.88, 3.49]$% and [0.48, 3.73]% across the original and joint-refit analyses. We retain this boundary sensitivity rather than selecting the favorable interval; these analyses are distinct from the original smaller-baseline results above.

A.2 Score-blind single-event selection probe

Predictive MSE and action selection evaluate different properties of a diagnostic. In the secondary lowest-quartile single-event selection probe within text/head/dose-half bins, the $\alpha = 1$ baseline-minus-extended selected-KL improvements are 2.17×10^{-6} $[-0.28, 4.73]\times10^{-6}$ for Qwen and 3.68×10^{-6} $[-0.24, 8.65]\times10^{-6}$ for Mistral. Both intervals cross zero, so this experiment does not establish a cross-model selection advantage, simultaneous-deletion benefit, or compression method. This probe is distinct from the matched-pair triage and action-aligned masking experiments reported elsewhere in the appendix.

A.3 Roles of the additional models

For Llama, the pre-specified candidate rule selects L20, including under the seed-split check. L20 is paired-positive, but its baseline-control and actual-placebo intervals do not isolate a positive direction-specific component. A subsequent exploratory audit of the remaining fixed candidates $\{24, 28, 31\}$ finds stronger evidence at L31: a local KL gap of 5.74×10^{-4} $[4.83, 6.64]\times10^{-4}$ and actual-minus-random contrast of 3.90×10^{-4} $[2.95, 4.88]\times10^{-4}$. We retain L20 and all follow-up candidates rather than treating L31 as a pre-specified replication (Table 12).

Gemma2 L33 tests a different readout condition: event selection incorporates the final-logit soft-cap derivative, multiplying $W_U v$ by $1 - (z/30)^2$ for capped logits z. Its paired local gap is 1.00×10^{-3} $[0.944, 1.07]\times10^{-3}$ (Table 8). This tests survival under bounded nonlinear readout, not the full direction-specific and predictive evidence sequence of the primary models. Stronger-baseline, analytic-control, and outcome-blind site-transfer results are untested here, not implicitly positive or negative.

A.4 LAYER LOCALIZATION

Head-level suppression implementation. The head-level localization intervention adds an attention bias of $-$`mask_value` at key position 0 for selected layer/head entries, using an attention-module forward pre-hook before probabilities are computed. Bright and dark groups come from direct-readout audits with matched head counts and mass controls. Suppressing the key renormalizes attention over the remaining keys; this is not the value-only event subtraction used for the primary within-head contrasts. Layer localization is used to decide where to run expensive matched interventions, not to claim that sink heterogeneity occurs only at a single depth. For Qwen3-8B and Mistral-7B, direct readout-Fisher audits were followed by same-layer, same-count, mass-matched head-level suppression; the initial intervention sites were Qwen3-8B layer 26 and Mistral-7B layer 27. Layer sweeps in Mistral showed that nearby late layers can be mixed or wash out, which is why we report layer selection as part of the method rather than hiding it as a fixed hyperparameter. For Llama-3.1-8B, we separated the static audit from the causal layer-selection rule. The static audit is used only to construct per-layer bright, dark, and mass-control groups. Before running the final within-head intervention, we fixed the causal candidate set to layers $\{20, 24, 28, 31\}$, the layers already audited in the preceding Llama sweep, and selected the layer with the largest bright–dark head-level logit-KL gap whose mass-matched bootstrap interval excluded zero. This rule selected layer 20, and we ran the within-head intervention only at that layer. Table 6 reports a held-out version of the same procedure. The localization split uses only dataset seeds 0–1 to apply the pre-specified selection rule; the held-out split then evaluates the selected layer on seeds 2–3 without re-selection. The split audit again selects layer 20 and confirms a positive held-out head-level bright–dark output gap. This check is a procedural guard against choosing and evaluating the Llama layer on the same text split; it is not used to screen any additional within-head layers. We apply the same seed-split reporting audit to the Qwen and Mistral head-level layer sweeps in Table 7. For Mistral, the split rule reselects layer 27 and the held-out gap remains positive. For Qwen, the split rule selects layer 28 as the strongest head-level layer; layer 26 remains positive on held-out seeds and is retained as the original pilot intervention site, while layer 28 is used as the conservative Qwen confirmation after the strict fp32 dose-matched re-slice.

Split	Layer	Text clusters	Bright KL	Dark KL	Bright–dark KL gap
Selection seeds 0–1	L20 selected	256	6.70×10^{-2}	5.59×10^{-3}	6.14×10^{-2} $[5.69, 6.62]\times10^{-2}$
Selection seeds 0–1	L24	256	1.34×10^{-2}	9.54×10^{-3}	3.89×10^{-3} $[3.29, 4.50]\times10^{-3}$
Selection seeds 0–1	L28	256	6.47×10^{-3}	6.74×10^{-3}	-2.62×10^{-4} $[-5.16, -0.12]\times10^{-4}$
Selection seeds 0–1	L31	256	2.00×10^{-3}	6.94×10^{-3}	-4.94×10^{-3} $[-5.45, -4.48]\times10^{-3}$
Held-out seeds 2–3	L20 selected	256	6.32×10^{-2}	5.55×10^{-3}	5.76×10^{-2} $[5.39, 6.16]\times10^{-2}$
Held-out seeds 2–3	L24	256	1.26×10^{-2}	9.09×10^{-3}	3.52×10^{-3} $[3.00, 4.02]\times10^{-3}$
Held-out seeds 2–3	L28	256	6.25×10^{-3}	6.46×10^{-3}	-2.13×10^{-4} $[-4.62, 0.25]\times10^{-4}$
Held-out seeds 2–3	L31	256	1.99×10^{-3}	6.32×10^{-3}	-4.34×10^{-3} $[-4.72, -3.97]\times10^{-3}$

Table 6: Held-out causal layer-localization audit for Llama-3.1-8B. The selection split applies the pre-specified rule over the fixed candidate set $\{20, 24, 28, 31\}$: choose the largest bright–dark head-level logit-KL gap whose text-cluster bootstrap interval is positive. The held-out split evaluates the selected layer on disjoint dataset seeds without re-selection.

Model	Candidate layers	Split-selected layer on seeds 0–1	Held-out gap at selected layer	Held-out gap at original/pilot layer
Mistral-7B	$\{23, 25, 27, 29\}$	L27: 5.58×10^{-3} $[4.87, 6.34]\times10^{-3}$	L27: 5.84×10^{-3} $[5.13, 6.57]\times10^{-3}$	L27: 5.84×10^{-3} $[5.13, 6.57]\times10^{-3}$
Qwen3-8B	$\{20, 22, 24, 26, 28\}$	L28: 1.09×10^{-2} $[1.04, 1.15]\times10^{-2}$	L28: 1.03×10^{-2} $[9.83, 10.70]\times10^{-3}$	L26: 1.72×10^{-3} $[1.30, 2.12]\times10^{-3}$
Llama-3.1-8B	$\{20, 24, 28, 31\}$	L20: 6.14×10^{-2} $[5.69, 6.62]\times10^{-2}$	L20: 5.76×10^{-2} $[5.39, 6.16]\times10^{-2}$	L20: 5.76×10^{-2} $[5.39, 6.16]\times10^{-2}$

Table 7: Seed-split head-level layer-localization audit. The selection split uses only dataset seeds 0–1 and applies the same pre-specified rule within each fixed candidate set: choose the largest positive bright–dark head-level logit-KL gap whose text-cluster bootstrap interval excludes zero. The held-out split evaluates seeds 2–3 without re-selection. The Qwen row is intentionally transparent: layer 28 is the strongest seed-split head-level layer and is used as the conservative Qwen confirmation in the main text, while layer 26 remains positive on the held-out split and is retained as a pilot/supportive site.

A.5 WITHIN-HEAD CLUSTER BOOTSTRAP

Table 8 reports clustered uncertainty for the original within-head comparisons in Figure 2 and the subsequent Qwen L28 confirmation; the historical figure itself does not contain Qwen L28. Rows include the original within-head sites and the seed-split-selected Qwen3-8B layer 28 confirmation; the fp32 strict layer-28 row is the conservative Qwen estimate used for the current main-text direction-specific claim. The bright/dark column reports the matched readout-Fisher separation, while the attention log-ratio verifies that the paired bright events do not simply receive larger first-token mass. The local KL column is the primary causal readout because each intervention removes a single query event; the global KL column is included to show how the same localized effect is diluted when averaged over the full sequence.

Model	Site	Pairs	Clusters	Bright/dark	Attn. log-ratio	Local KL Δ	Global KL Δ
Mistral-7B	L27	2048	512	$79.0\times$ [73.7, 84.5]	-0.039 [-0.050, -0.029]	3.19×10^{-4} [2.70×10^{-4}, 3.75×10^{-4}]	5.02×10^{-6} [4.12×10^{-6}, 5.94×10^{-6}]
Qwen3-8B	L26	2560	512	$446\times$ [360, 562]	-0.070 [-0.079, -0.060]	7.50×10^{-4} [6.86×10^{-4}, 8.14×10^{-4}]	2.23×10^{-5} [1.74×10^{-5}, 2.73×10^{-5}]
Qwen3-8B	L28 split-selected	955	412	$630\times$ [436, 941]	-0.0048 [-0.0081, -0.0016]	1.36×10^{-3} [1.17×10^{-3}, 1.57×10^{-3}]	1.09×10^{-5} [4.80×10^{-6}, 1.70×10^{-5}]
Qwen3-8B	L28 fp32 strict	1148	493	$470\times$ [354, 653]	-0.0038 [-0.0067, -0.0009]	5.60×10^{-4} [4.30×10^{-4}, 7.11×10^{-4}]	1.68×10^{-6} [1.26×10^{-6}, 2.16×10^{-6}]
Llama-3.1-8B	L20	2048	512	$43.8\times$ [41.2, 46.6]	-0.013 [-0.017, -0.009]	2.07×10^{-4} [1.46×10^{-4}, 2.69×10^{-4}]	6.93×10^{-6} [3.96×10^{-6}, 9.85×10^{-6}]
Gemma2-9B	L33 softcap	2048	512	$219\times$ [199, 240]	$+0.010$ [$+0.004$, $+0.015$]	1.00×10^{-3} [9.44×10^{-4}, 1.07×10^{-3}]	5.61×10^{-7} [-2.87×10^{-6}, 3.92×10^{-6}]

Table 8: Text-cluster bootstrap robustness for within-head interventions at $\alpha = 1.0$. Intervals resample text clusters; local KL is the primary readout. The Qwen layer-28 fp32 strict row is a conservative dose-caliper re-slice of the seed-split-selected Qwen site and is the Qwen estimate used for the strict direction-specific claim.

Table 9 reports a leave-one-seed-out stability check over the same within-head CSVs. For each model, we recompute the text-cluster bootstrap after omitting one dataset seed at a time. All leave-one-seed-out intervals remain positive, showing that the within-head effect is not an artifact of a single seed split.

Model	Site	Full pairs	Full clusters	Full local KL Δ	Leave-one-seed-out local KL range	Leave-one-seed-out CI envelope
Mistral-7B	L27	2048	512	3.18×10^{-4} [2.70, 3.71]$\times 10^{-4}$	$[2.90, 3.46] \times 10^{-4}$	$[2.39, 4.12] \times 10^{-4}$
Qwen3-8B	L26	2560	512	7.51×10^{-4} [6.87, 8.16]$\times 10^{-4}$	$[7.26, 7.70] \times 10^{-4}$	$[6.48, 8.46] \times 10^{-4}$
Qwen3-8B	L28 fp32 strict	1148	493	5.61×10^{-4} [4.28, 7.16]$\times 10^{-4}$	$[5.17, 6.17] \times 10^{-4}$	$[3.73, 8.07] \times 10^{-4}$
Llama-3.1-8B	L20	2048	512	2.06×10^{-4} [1.45, 2.66]$\times 10^{-4}$	$[1.91, 2.16] \times 10^{-4}$	$[1.21, 2.82] \times 10^{-4}$
Gemma2-9B	L33 softcap	2048	512	1.00×10^{-3} [0.94, 1.07]$\times 10^{-3}$	$[0.99, 1.02] \times 10^{-3}$	$[0.92, 1.09] \times 10^{-3}$

Table 9: Leave-one-seed-out robustness for the within-head local KL effect at $\alpha = 1.0$. The range column reports the minimum and maximum leave-one-seed-out mean local KL gaps; the envelope column reports the smallest lower confidence endpoint and the largest upper endpoint across leave-one-seed-out reruns. The Qwen layer-28 fp32 strict row uses the same strict dose-matched subset as the conservative Qwen confirmation in Table 8.

A.6 READOUT-CONTROL AUDITS

Table 10 expands the artifact-control audit summarized in the main text. The direct gap is the matched bright-minus-dark gap in unit readout-Fisher score, computed within the same intervention pairs used for the causal analysis. Logit-ℓ_2 tests whether the effect is just logit-norm magnitude. The three ratio columns compare the direct paired gap with random same-norm, shuffled-vocabulary, and signed-permutation controls. These paired gaps are the relevant control statistic: event-level correlations can reflect generic positional sensitivity, whereas the matched-pair comparison isolates the readout-direction-specific component.

Model	Site	Pairs	Clusters	Direct gap	Logit-ℓ_2 gap	Direct/random	Direct/shuffled	Direct/signed
Mistral-7B	L27	2048	512	5.80×10^{-5} [5.71×10^{-5}, 5.90×10^{-5}]	1.65×10^{-9} [-3.79×10^{-9}, 7.07×10^{-9}]	$7.9\times$	$13.2\times$	$6.6\times$
Qwen3-8B	L26	2560	512	3.20×10^{-3} [3.06×10^{-3}, 3.35×10^{-3}]	5.24×10^{-9} [-6.73×10^{-9}, 1.73×10^{-8}]	$10.4\times$	$12.0\times$	$11.0\times$
Qwen3-8B	L28 split-selected	1714	429	3.25×10^{-3} [3.16×10^{-3}, 3.34×10^{-3}]	2.43×10^{-8} [2.52×10^{-9}, 4.65×10^{-8}]	$10.0\times$	$11.5\times$	$9.4\times$
Llama-3.1-8B	L20	512	128	9.48×10^{-4} [9.12×10^{-4}, 9.84×10^{-4}]	-1.13×10^{-7} [-2.85×10^{-7}, 5.59×10^{-8}]	$10.7\times$	$20.0\times$	$13.6\times$

Table 10: Matched readout-control audits. Direct readout-Fisher paired gaps are compared with logit-norm, random same-norm, shuffled-unembedding, and signed-permutation controls. Ratios use mean paired gaps; intervals resample text clusters.

A.7 BASELINE-CONFOUND AND ACTUAL RANDOM-DIRECTION PLACEBO AUDITS

Table 11 reports two stricter checks that are distinct from the readout-control score audit above. The baseline-control coefficient asks whether brightness still predicts the local intervention gap after controlling for baseline event descriptors such as attention mass, raw residual norm, query position, and entropy. The actual random-direction placebo replaces each selected event direction with a same-norm random residual direction at the same layer and query, then runs the intervention rather than only comparing readout scores. Mistral layer 27 and the Qwen layer-28 fp32 strict site pass both checks; Llama layer 20 remains paired-positive but does not isolate a positive residual brightness-specific component under these stricter audits. For Qwen layer 28, the baseline-control regression uses the finite-covariate subset of the same strict fp32 pairs (1142 of 1148 pairs), while the actual placebo uses all 1148 strict pairs.

Model	Site	Pairs/clusters	Baseline-control coef.	Actual KL gap	Random-direction KL gap	Actual–random	Interpretation
Mistral-7B	L27	2048/512	6.35×10^{-5} [1.53, 11.53]$\times 10^{-5}$	3.18×10^{-4} [2.69, 3.72]$\times 10^{-4}$	1.30×10^{-4} [0.97, 1.64]$\times 10^{-4}$	1.88×10^{-4} [1.36, 2.44]$\times 10^{-4}$	Pass
Qwen3-8B	L28 fp32 strict	1148/493	2.40×10^{-4} [1.20, 3.61]$\times 10^{-4}$	5.60×10^{-4} [4.30, 7.11]$\times 10^{-4}$	1.53×10^{-4} [0.97, 2.23]$\times 10^{-4}$	4.08×10^{-4} [2.85, 5.49]$\times 10^{-4}$	Pass
Llama-3.1-8B	L20	2048/512	2.59×10^{-5} [-2.33, 7.46]$\times 10^{-5}$	2.06×10^{-4} [1.45, 2.66]$\times 10^{-4}$	1.61×10^{-4} [1.02, 2.22]$\times 10^{-4}$	0.45×10^{-4} [-0.16, 1.07]$\times 10^{-4}$	Boundary

Table 11: Confound and actual random-direction placebo checks for the within-head causal effect. The baseline-control coefficient is the residual association of brightness with the local intervention gap after controlling for baseline event descriptors. The actual random-direction columns report finite-dose interventions using same-norm random residual directions, so the actual–random column estimates the direction-specific component beyond generic same-norm perturbability.

A.8 LLAMA FIXED-CANDIDATE LAYER RESCUE AUDIT

Table 12 asks whether the weak direction-specific result at the pre-specified Llama layer 20 reflects a genuine model-family absence or a localization miss. We keep the metric and intervention protocol fixed and audit the remaining Llama candidate layers that had already been included in the causal localization sweep. This is a rescue/falsification audit, not a replacement for the pre-specified layer-20 result: layer 20 remains the procedurally selected site, while layer 31 shows that a stronger within-head direction-specific component exists elsewhere in the same model.

Layer	Pairs	Clusters	Bright/dark	Local KL Δ	Baseline-control coef.	Random KL gap	Actual–random	Reading
L24	1597	511	49× [45, 53]	1.37×10^{-4} [0.84, 1.90]$\times 10^{-4}$	9.21×10^{-5} [2.89, 15.08]$\times 10^{-5}$	0.95×10^{-4} [0.30, 1.55]$\times 10^{-4}$	0.41×10^{-4} [-0.18, 1.03]$\times 10^{-4}$	Placebo weak
L28	1648	511	32× [30, 35]	1.15×10^{-4} [0.56, 1.74]$\times 10^{-4}$	1.46×10^{-4} [0.94, 2.00]$\times 10^{-4}$	0.42×10^{-4} [-0.09, 0.92]$\times 10^{-4}$	0.73×10^{-4} [0.17, 1.28]$\times 10^{-4}$	Pass
L31	762	443	61× [54, 68]	5.74×10^{-4} [4.83, 6.64]$\times 10^{-4}$	2.10×10^{-4} [0.97, 3.34]$\times 10^{-4}$	1.84×10^{-4} [1.18, 2.51]$\times 10^{-4}$	3.90×10^{-4} [2.95, 4.88]$\times 10^{-4}$	Pass, strongest

Table 12: Fixed-candidate Llama rescue/falsification audit under the same strict ± 0.1 dose caliper used for the conservative Qwen layer-28 confirmation. Local KL, baseline-control coefficients, and actual random-direction placebo gaps use text-cluster bootstrap intervals. All leave-one-seed-out local-KL intervals remain positive for L24, L28, and L31; only L28 and L31 also isolate a positive actual-minus-random component, with L31 strongest.

A.9 DIRECT-READOUT SMALL-PERTURBATION CHECK

Table 13 checks the implementation-level identity that should hold for an infinitesimal direct-readout perturbation. Here RMSNorm denotes root mean square normalization, which rescales a hidden vector using its root mean square and learned coordinate-wise weights. For each matched event direction v_e, we capture the residual immediately before the final normalization, apply the exact final-RMSNorm counterfactual in logit space, and compare the resulting KL with the quadratic prediction $\frac{1}{2}\alpha^2 \Delta B_{\mathrm{norm}}$ from the final-RMSNorm-Jacobian brightness score. The exact counterfactual and the quadratic prediction agree to within 0.2% even at $\alpha = 0.25$, and essentially coincide at $\alpha = 0.01$. This validates the local readout coordinate used by the diagnostic and rules out an implementation artifact in the final-normalization correction. It does not make the finite-dose within-head intervention tautological: the causal results in the main text delete an intermediate-layer contribution and measure the propagated model response, whereas this audit changes only the final readout after the network has already computed the residual state. On the same Qwen3-8B layer-26 pilot matched pairs used in Table 8, the finite-dose exact final-RMSNorm direct-readout gap at $\alpha = 1.0$ is 2.34×10^{-5} $[2.12, 2.57] \times 10^{-5}$, whereas the propagated local intervention gap is 7.50×10^{-4} $[6.86, 8.12] \times 10^{-4}$; their ratio is 0.031. This compares distinct counterfactuals and is not the fraction of a causal effect

attributable to the direct path. For the soft-capped Gemma2 stress test, the same baseline-aligned exact direct-readout calculation gives a direct KL gap of 8.26×10^{-6} at $\alpha = 1.0$, only 0.82% of the observed 1.00×10^{-3} local intervention gap, so the softcap result is also not explained by the direct logit-lens term alone.

Model and site	Pairs	Clusters	Exact/quadratic, $\alpha = .01$	Exact/quadratic, $\alpha = .25$	Sign consistency, $\alpha = .25$	Direct KL, $\alpha = .25$	Quadratic KL, $\alpha = .25$
Mistral-7B L27	2048	512	$0.99998\times$	$0.99936\times$	100%	1.89×10^{-5}	1.89×10^{-5}
Qwen3-8B L26	2560	512	$1.00005\times$	$1.00131\times$	100%	1.46×10^{-6}	1.45×10^{-6}
Qwen3-8B L28	1714	429	$1.00004\times$	$1.00090\times$	100%	1.24×10^{-5}	1.24×10^{-5}
Llama-3.1-8B L20	2048	512	$1.00001\times$	$1.00020\times$	100%	1.97×10^{-6}	1.97×10^{-6}

Table 13: Final-normalization-aware small-perturbation check for the RMSNorm models. Exact/quadratic is the ratio between the exact baseline-aligned final-readout counterfactual KL gap and the quadratic prediction $\frac{1}{2}\alpha^2 \Delta B_{\text{norm}}$ from the local final-RMSNorm-Jacobian brightness score. Sign consistency reports the fraction of matched pairs for which the exact and quadratic bright-minus-dark gaps have the same sign. The table validates the local direct-readout coordinate; it is not a finite-dose decomposition of the propagated within-head causal effect.

A.10 Final-RMSNorm Jacobian reanalysis

Table 14 addresses whether the direct-readout split is an artifact of ignoring the final RMSNorm before unembedding. For each existing within-head event, we reconstruct v_e, capture the pre-final-normalization residual state at the query position, apply the closed-form local RMSNorm Jacobian $J_{\text{norm}}(h)$, and recompute brightness using $W_U J_{\text{norm}}(h) v_e$. This is still not a full layer-to-output pullback, but it tests the most immediate normalization correction between the residual stream and the unembedding. The correction modestly reduces the bright/dark ratio in Mistral and Llama and leaves Qwen nearly unchanged; in all three RMSNorm models, and in the Qwen layer 28 confirmation, the corrected matched-pair gap remains strongly positive and the corrected score remains highly correlated with the original direct-readout score. Thus the low-cost diagnostic is not relying on a fragile identity-Jacobian approximation at the final normalization layer.

Model and site	Final norm	Pairs	Clusters	Original ratio	RMSNorm-Jacobian ratio	Positive pairs	Rank correlations
Mistral-7B L27	RMSNorm, $\epsilon = 10^{-5}$	2048	512	$79.0\times$	$48.6\times$ [44.9, 52.5]	98.8% [98.3, 99.3]	$\rho(B_{\text{norm}}, K_{\text{actual}}) = 0.44$, $\rho(B, B_{\text{norm}}) = 0.94$
Qwen3-8B L26	RMSNorm, $\epsilon = 10^{-6}$	2560	512	$446\times$	$429\times$ [341, 546]	100% [100, 100]	$\rho(B_{\text{norm}}, K_{\text{actual}}) = 0.54$, $\rho(B, B_{\text{norm}}) = 0.95$
Qwen3-8B L28	RMSNorm, $\epsilon = 10^{-6}$	1714	429	$734\times$	$648\times$ [489, 877]	100% [100, 100]	$\rho(B_{\text{norm}}, K_{\text{actual}}) = 0.62$, $\rho(B, B_{\text{norm}}) = 0.97$
Llama-3.1-8B L20	RMSNorm, $\epsilon = 10^{-5}$	2048	512	$43.8\times$	$34.4\times$ [32.0, 36.9]	99.9% [99.8, 100]	$\rho(B_{\text{norm}}, K_{\text{actual}}) = 0.25$, $\rho(B, B_{\text{norm}}) = 0.85$

Table 14: Final-RMSNorm Jacobian reanalysis of the within-head matched pairs at $\alpha = 1.0$. The original ratio is the bright/dark direct-readout brightness ratio from the matched intervention table, rounded to the same reporting precision. The RMSNorm-Jacobian ratio recomputes the same score after applying the local final-RMSNorm Jacobian to each event direction before unembedding. Intervals resample text clusters. Positive pairs reports the fraction of matched pairs in which the corrected bright event remains brighter than the corrected dark event.

A.11 Matched-sample prevalence and decile audit

Table 15 checks whether the within-head effect is driven only by a few extreme matched pairs. This audit uses the existing intervention CSVs, so the prevalence numbers are conditional on the matched intervention sample rather than on the full candidate event pool. Within that matched sample, the bright/dark contrast is common: at least 96.8% of selected pairs exceed a $10\times$ brightness switch at matched dose, and the top brightness decile has higher mean local KL than the bottom decile in all three RMSNorm models. In the conservative Qwen layer-28 fp32 strict subset, all selected pairs exceed a $10\times$ brightness switch, 66.1% exceed $100\times$, and the top brightness decile has about $99\times$ the local KL of the bottom decile. The trend is strongest in Mistral and Qwen and weaker but still positive in Llama; accordingly, we use this audit to support breadth within the matched sample, not to claim a globally monotone event law over all possible sink events. At the pre-specified Llama layer 20, continuous-prediction strength is weak: its rank correlation and decile ratio are smaller, but the paired local-KL interval excludes zero, the held-out layer-selection audit still selects L20, and the final-RMSNorm correction preserves the bright-over-dark ordering for 99.9% of matched pairs. The exploratory fixed-candidate audit finds stronger within-head evidence at L31. It establishes

that the tested Llama sites differ, but does not by itself identify why L20 is weaker or prove that L31 is the strongest layer outside the tested candidates. Table 16 complements this matched-sample audit by measuring brightness switching before bright/dark pairing, over the full candidate pools within the audited sink heads used by the within-head searches. Because the WikiText segments do not share a fixed first token, this audit samples random pairs within the same text and head, so the first-token value direction is fixed within each sampled comparison. The strict same-text/head random-pair statistic is the primary pre-matching prevalence readout: random same-text, same-head query pairs exceed a $10\times$ switch with probability 24.9–42.5%. This pre-matching result rules out the interpretation that large bright/dark contrasts are solely an artifact of selecting a few extreme matched pairs. It does not claim that attention heads generally exhibit switching; it characterizes prevalence within high-mass sink heads selected by the preceding sink audits.

Model and site	Events	Pairs	$\rho(B, K_{\mathrm{actual}})$	Top/bottom KL decile	Top–bottom KL	Pairs $\geq 10\times$ switch	Positive KL gap pairs
Mistral-7B L27	4096	2048	0.50	$17.1\times$	7.80×10^{-4}	99.5% ($33.5\% \geq 100\times$)	87.6%
Qwen3-8B L26	5106	2546	0.52	$10.4\times$	1.24×10^{-3}	100% ($65.2\% \geq 100\times$)	78.2%
Llama-3.1-8B L20	4096	2048	0.20	$1.77\times$	3.22×10^{-4}	96.8% ($17.0\% \geq 100\times$)	60.0%

Table 15: Matched-sample prevalence and brightness-decile audit at $\alpha = 1.0$. Events are the matched bright/dark intervention events, not the full candidate event pool. The decile columns compare the top and bottom event deciles after ranking the matched sample by readout-Fisher brightness. Pair-switch columns report how often the selected bright event is at least $10\times$ or $100\times$ brighter than its matched dark event while using the same head and first-token write direction.

Model and site	Eligible events	Heads	Same-text/head 90/10 ratio	Heads $\geq 10\times$	Random same-text/head pairs $\geq 10\times$
Qwen3-8B L26	488845	5	$1242\times$ [1153, 1277]	100%	42.5%
Qwen3-8B L28	327148	4	$1215\times$ [1024, 1300]	100%	42.2%
Mistral-7B L27	393214	4	$131\times$ [93.5, 182]	100%	33.5%
Llama-3.1-8B L20	393216	4	$56.8\times$ [49.8, 60.5]	100%	27.5%
Gemma2-9B L33	393216	4	$31.2\times$ [29.6, 36.5]	100%	24.9%

Table 16: Full candidate-pool brightness-switching audit within audited sink heads, before bright/dark matching. Both the 90/10 column and the final random-pair column are computed within same-text/head scopes, fixing the first-token write direction. The final column estimates how often two eligible events with the same first-token write direction differ by at least $10\times$ in brightness-over-norm.

A.12 Boundary quantification

Direction-level projections are supplementary operators: low-rank bright/dark projections can separate effects, but high-rank projections mix them. They are not a canonical basis for the event-level diagnostic.

Table 17 retains the supplementary representation, cache, and long-context checks; they are not additional premises of the two central findings. The representational rows compare the final-layer effective-rank contrast against mass-matched controls: the large propagated geometry change is driven primarily by bright suppression, so it should not be described as dark ablation directly causing representational collapse. The KV-native rows summarize same-count, non-overlapping cache eviction; negative dark–bright values mean bright KV heads perturb rolling logits more than dark KV heads, which argues against treating readout-dark sink heads as cache-critical anchors.

Boundary audit	Scope	Quantity	Estimate	Reading
Representational decomposition	Mistral L27, final layer, no first token	Bright–control effective-rank Δ	$+4.32$ [$+4.11$, $+4.54$]	Bright suppression strongly de-mixes final states
Representational decomposition	Mistral L27, final layer, no first token	Dark–control effective-rank Δ	$+0.066$ [$+0.034$, $+0.098$]	Dark suppression has only a small direct rank effect
Representational decomposition	Mistral L27, final layer, no first token	Dark–bright effective-rank Δ	-4.07 [-4.29, -3.86]	Final contrast is propagated geometry, not dark collapse alone
KV-native eviction	Mistral L27, prefix 512, all positions	Abs-logit dark–bright	-0.063 [-0.064, -0.062]	Bright KV eviction changes rolling logits more
KV-native eviction	Mistral L27, prefix 512, all positions	KL dark–bright	-1.8×10^{-3} [-1.9, -1.7]$\times 10^{-3}$	Dark KV heads are not more cache-critical by output sensitivity

Table 17: Quantified boundary checks. The table gives representative numbers for the two boundary claims: readout-dark components are not globally inert, but the available representational and KV-native checks do not support stronger claims that they directly cause collapse or act as cache-critical anchors.

A.13 DECISION-RELEVANCE EVICTION PROBE

Tables 18 and 19 report scoped decision-relevance probes rather than KV-compression benchmarks. We zero only first-token values at a fixed budget of two KV heads, then measure rolling teacher-forced KL over a 128-token continuation after a 512-token prefix. The low-score runs use seeds $0, 1, 2, 3$; the high-score sanity runs use seeds $1000, 2000, 3000$. Intervals resample text clusters within each run, so the high/low ratios are descriptive point-estimate comparisons rather than paired confidence intervals. The result is directionally useful but deliberately narrow: high dose-aware readout-Fisher/value eviction is much more disruptive than low dose-aware eviction, while total readout-Fisher and value norm select the same or nearly the same coarse KV heads, unit-readout-Fisher is worse, and oracle versus amortized total readout-Fisher does not separate.

Model and layer	Score	Low-score KL	High-score KL	High/low	Selection unit	Reading
Qwen3-8B L26	Total readout-Fisher	8.07×10^{-4} $[7.91, 8.25]\times10^{-4}$	2.83×10^{-3} $[2.70, 2.96]\times10^{-3}$	$3.5\times$	KV head	Dose-aware low-score eviction is safer
Qwen3-8B L26	Value ℓ_2	8.07×10^{-4} $[7.91, 8.25]\times10^{-4}$	2.85×10^{-3} $[2.71, 2.99]\times10^{-3}$	$3.5\times$	KV head	Same coarse selection as total readout-Fisher in the low run
Mistral-7B L27	Total readout-Fisher	1.71×10^{-4} $[1.68, 1.75]\times10^{-4}$	7.29×10^{-4} $[6.97, 7.64]\times10^{-4}$	$4.3\times$	KV head	Dose-aware low-score eviction is safer
Mistral-7B L27	Value ℓ_2	1.65×10^{-4} $[1.62, 1.68]\times10^{-4}$	7.29×10^{-4} $[6.97, 7.64]\times10^{-4}$	$4.4\times$	KV head	Value dose is not separated from total readout-Fisher at this granularity

Table 18: Scoped zero-value KV eviction probe. Low-scoring dose-aware readout-Fisher/value heads are less disruptive than high-scoring heads under a fixed two-KV-head budget, but this should not be read as a compression method or as evidence that readout-Fisher beats value norm. The coarse KV-head intervention is a decision-relevance sanity check; it does not support a static pruning recipe, unit-readout-Fisher selection, or an oracle-over-amortized claim.

Table 19 gives the finer matched-dose version of the same question using the existing within-head intervention pairs. Each pair fixes the text, layer, head, and first-token write direction family, and matches attention mass, raw residual dose, entropy, and query position. The low-brightness policy suppresses the matched readout-dark event, the high-brightness policy suppresses the matched readout-bright event, and the random baseline is the within-pair random-choice expectation. Because this expectation equals the average of the high- and low-policy outcomes, low beating random is algebraically implied by a positive high–low gap, not an independent application success or a held-out policy comparison. We audit this probe with three pre-specified checks: high–low local KL must have a positive text-cluster interval, low-brightness suppression must beat the matched-random expectation, and aggregate attention-mass and raw-ℓ_2 log-ratios must stay within ±0.1. All four model/layer runs pass these checks. This re-analysis expresses the existing paired effect contrast as a within-pair choice; it is not an independent demonstration of actionable superiority.

Model and layer	Low-bright KL	High-bright KL	Random KL	High/low	High–low KL	Dose check
Mistral-7B L27	3.23×10^{-4}	6.41×10^{-4}	4.82×10^{-4}	$1.99\times$	3.18×10^{-4} $[2.70, 3.71]\times10^{-4}$	raw-ℓ_2 log-ratio -0.039
Qwen3-8B L26	5.45×10^{-4}	1.30×10^{-3}	9.21×10^{-4}	$2.38\times$	7.51×10^{-4} $[6.87, 8.16]\times10^{-4}$	raw-ℓ_2 log-ratio -0.070
Llama-3.1-8B L20	6.29×10^{-4}	8.35×10^{-4}	7.32×10^{-4}	$1.33\times$	2.06×10^{-4} $[1.45, 2.66]\times10^{-4}$	raw-ℓ_2 log-ratio -0.013
Gemma2-9B L33 softcap	4.92×10^{-4}	1.50×10^{-3}	9.94×10^{-4}	$3.04\times$	1.00×10^{-3} $[0.94, 1.07]\times10^{-3}$	raw-ℓ_2 log-ratio $+0.010$

Table 19: Within-pair choice restatement of the historical matched deletion contrasts. The random-choice expectation is the arithmetic mean of high- and low-brightness outcomes, so low beating random follows algebraically from a positive high–low gap and is not independent evidence of policy superiority. Reported attention-mass and residual-norm log-ratios describe aggregate dose balance, not a universal per-pair caliper. The Qwen row is the historical L26 panel, not the primary L28 confirmation.

A.14 FIXED-PLAN CNN/DAILYMAIL REPLICATION

Protocol. We use only article text from `abisee/cnn_dailymail`, configuration 3.0.0, test split, dataset revision `96df5e686bee6baa90b8bee7c28b81fa3fa6223d`; reference summaries are not used. The original model revisions, Qwen L28 heads $\{0, 9, 19, 29\}$ and Mistral L27 heads $\{8, 10, 11, 31\}$, and analysis settings are fixed before measuring the new outcomes. Per model, 64 calibration and 64 evaluation articles each supply eight score-blind queries per head, with maximum length 512 and minimum query position 32. Four additional disjoint smoke-test articles check reconstruction and no-op validity, not scientific effect size. Normalized-content and token-prefix duplicates are excluded within the run and against the supplied prior manifest; this is not pretraining decontamination. Inference uses eager fp32 with TF32 disabled, at $\alpha \in \{0.25, 1\}$, producing 4,096 unique events and 8,192 intervention rows per model. Both completed runs report zero maximum

Setting	Pairs / texts	Coverage	$\alpha = 0.25$	$\alpha = 1$
Qwen L28	182 / 63	17.77%	0.102 [0.053, 0.165]	1.730 [0.863, 2.863]
Mistral L27	282 / 64	27.54%	0.060 [0.022, 0.113]	1.036 [0.425, 1.889]

Table 20: CNN/DailyMail matched bright-minus-dark local KL, in units of 10^{-4}, with 95% text-bootstrap intervals. Coverage is the fraction of 2,048 evaluation events per dose retained in pairs. Both models have 64 eligible evaluation texts.

Setting	Predictor	Add Fisher	Add S	Add Fisher given S
Qwen L28	Ridge	5.58 [3.70, 9.65]	3.42 [2.60, 7.19]	2.13 [−0.73, 5.18]
Mistral L27	Ridge	5.78 [2.36, 12.12]	4.09 [1.98, 9.24]	2.18 [0.23, 5.39]
Qwen L28	HGB	0.17 [−6.07, 13.83]	−1.83 [−7.68, 10.17]	0.04 [−7.34, 12.55]
Mistral L27	HGB	−1.34 [−9.53, 5.00]	−0.77 [−7.58, 4.17]	−1.88 [−7.07, 8.16]

Table 21: CNN/DailyMail pooled relative MSE reduction (%) with calibration-refit 95% intervals. The first two columns add features to the same strong baseline; the last compares baseline+S against baseline+S+Fisher. Positive values indicate lower held-out error.

reconstruction and no-op errors, and clean-feature enrichment preserves the original intervention outcomes. Predictors are independently calibrated on the news articles; no frozen-predictor or generation-task transfer is claimed.

Secondary matched contrasts. Within each evaluation text/head, the two highest and two lowest unit-brightness queries from the eight-query pool are matched without reuse or access to deletion outcomes. Both attention-mass and residual-norm absolute log-ratios must be at most 0.1; entropy and query fraction enter matching distance but have no hard balance guarantee. The same pairs are evaluated at both doses. Table 20 reports text-equal means with 5,000 whole-text bootstrap resamples, conditional on retained pairs. The positive contrasts support replication in the retained sparse pool, not population-wide prevalence or a new direction-specific placebo result.

Predictive contrasts and aggregation. Table 21 reports the primary $\sqrt{K_{\text{actual}}}$ target at $\alpha = 1$, with 200 joint calibration/evaluation text-bootstrap refits. Each contrast uses its own reference risk; percentages must not be subtracted as a mechanism decomposition. Intervals are marginal rather than simultaneous across comparisons. Both Fisher and S improve pooled ridge prediction, but the conditional Fisher gain is resolved only for Mistral in this corpus. This reversal of which model has a positive conditional interval relative to WikiText-2 does not itself establish a between-model or between-corpus difference. Head-equal macro relative Fisher-after-S gains are -3.99% [-14.52, 7.31] for Qwen and 2.34% [-1.04, 7.68] for Mistral; neither excludes zero. Qwen's macro baseline-to-Fisher and baseline-to-S intervals also cross zero. Pooled results therefore do not imply uniform benefits across heads, and the HGB results do not establish predictor-independent gains.

A.15 Analytic permutation control across sites

The analytic control uses the existing enriched event records, with vocabulary sizes 151,936 for Qwen and 32,768 for Mistral from the original model configurations. It adds no new model interventions. The four source datasets retain their calibration/evaluation splits; the eight-site data are identical to those underlying Table 4. Ridge and HGB each use 200 calibration/evaluation article-bootstrap refits, sharing each article draw across feature sets, sites, and held-layer folds. Original-site summaries pool text-equal risk, whereas the eight-site summaries equally average within-site relative reductions.

Table 22 does not establish a stable conditional advantage in either regime. Analytic S itself improves Mistral new-text ridge prediction by 4.94% [0.52, 12.60]; its other eight-site regime/predictor intervals do not establish a positive gain. This bounded result complements, rather than replaces, the original-site evidence in Table 3.

Refitting sensitivity and provenance. Positive features are log-transformed before fitting; the original baseline contains $\log(\sum_i p_i^2)$, not $\log(1 - \sum_i p_i^2)$. Dispersion alone has unresolved ridge increments, whereas baseline+S improves over baseline+dispersion by 4.52% [2.92, 7.04] and 7.77% [5.22, 10.35] in Qwen and Mistral, respectively. Nevertheless, $\log S$ is a sum of log variance and

Model	Predictor	New text	New text + held layer
Qwen	Ridge	0.23 [−2.60, 7.09]	−0.27 [−5.07, 6.89]
Mistral	Ridge	4.43 [−1.16, 8.78]	−1.71 [−6.31, 14.92]
Qwen	HGB	4.52 [−15.89, 9.10]	−4.42 [−17.96, 9.87]
Mistral	HGB	2.98 [−4.18, 7.43]	8.03 [−8.48, 9.93]

Table 22: Eight-site relative MSE reduction (%) from adding real Fisher conditional on analytic S, with calibration-refit 95% intervals. All intervals cross zero; their widths preclude interpreting this as equivalence or proof of no alignment-specific value.

Site	Norm-policy KL	Fisher-policy KL	Gain [95% CI]
Qwen L28	104.61	105.68	−1.07 [−4.48, 1.88]
Mistral L27	22.71	30.88	−8.17 [−19.57, −0.32]

Table 23: Action-aligned pilot. All values are in units of 10^{-11}. Gain is norm-policy minus Fisher-policy KL, with 5,000 text-bootstrap draws. The table concerns this local candidate set, not deployment performance.

log dispersion, and ridge's standardization, individual squared terms, and penalty change with the supplied representation. The comparison therefore does not isolate a physical interaction mechanism or distinguish basis expansion from regularization effects. The signed deviation $B - S$ is descriptive, not a separately manipulated causal variable; we do not fit it through the positive-feature log transform or report an "explained percentage" of mechanism.

The analytic audit uses cross-validation/refit seed 20260924, versus 20260923 in the initial eight-site audit. This changes calibration folds and potentially hyperparameter choices, so fitted point estimates can change as well as interval endpoints. For example, Mistral's new-text ridge baseline-to-real-Fisher comparison is 6.77% [0.45, 13.55] in the initial audit and 5.27% [−0.31, 12.13] in the analytic audit on the same events. We retain both analyses; the later result is not an independent replication, and neither interval is selected to replace the other. The comparisons with analytic S always use the same-round baseline, seed, and resampling scheme.

A.16 Action-aligned single-query masking pilot

This follow-up aligns the score with a different action: masking a past key for one KV group at one receiving query, with attention renormalization. It is neither deletion of an unrescaled natural value contribution nor persistent cache eviction. Each model uses 16 texts, 128 decision contexts, and 512 measured candidate actions: the first key plus three sampled non-first keys per context. We compare final-normalization-aware action Fisher with projected action norm; positive gain means lower actual KL under the Fisher policy.

Table 23 provides no evidence of an advantage over norm: Qwen is unresolved, and Mistral's interval is negative in this pilot. Both policies avoid the first-token action in all 128 contexts; first-token actions account for approximately 99.97% and 99.99% of the total candidate KL in Qwen and Mistral. Consequently, beating random choice mainly reflects avoiding a high-damage candidate, not a geometry-specific advantage. Norm is already close to the measured candidate oracle, leaving very little absolute headroom. The very small policy differences have not been independently validated against fp32 forward rounding at their own scale; no-op and mask/additive consistency checks are insufficient to establish deployment-level harm or benefit. We retain the result as a scoped decision boundary, without changing the policy, score, or candidate set after observing it.

A.17 Natural first-token and non-first-token contribution controls

This experiment fixes the primary sites and samples four queries per text/head without consulting scores or intervention outcomes. For each model, 64 calibration and 64 evaluation articles define 1,024 candidate pairs in each split. At each query, the control is the preceding non-first key with closest natural residual norm, subject to the fixed absolute log-norm caliper of 0.1. No vector is rescaled and unmatched candidates are retained in coverage counts. The two groups use separate

Site	α	First gain	Non-first gain	Difference [95% CI]
Qwen L28	1	7.93%	18.34%	-10.41 [-17.08, -3.06]
Mistral L27	1	19.18%	26.24%	-7.06 [-15.76, 6.65]
Qwen L28	.25	7.98%	18.29%	-10.31 [-17.11, -2.83]
Mistral L27	.25	22.54%	26.49%	-3.95 [-15.34, 7.23]

Table 24: Natural-key control on dose overlap. Gain is the relative reduction in held-out, text-equal prediction MSE after adding Fisher; the difference is first minus non-first in percentage points. The primary dose is $\alpha = 1$; .25 is secondary. Intervals jointly bootstrap matched evaluation texts 5,000 times, conditional on the separate calibration fits. These are not estimates of the causal effect of first-key identity.

calibration fits in the original ridge protocol of Appendix A.1, with key distance added to both the baseline and Fisher-augmented feature sets; they do not use the later strengthened probability-shape baseline. Table 24 reports relative MSE reductions for $\sqrt{K_{\text{actual}}}$ and their paired-text difference, not differences in deletion KL magnitude.

Coverage and balance. Qwen retains 154 calibration pairs across 57 texts and 144 evaluation pairs across 58 texts; Mistral retains 102 across 51 and 79 across 46, respectively. Evaluation coverage is 14.06% and 7.71%; among candidates with first-token attention at least 0.1, coverage is 135/1,014 (13.31%) and 76/1,013 (7.50%). Maximum absolute matched dose log-ratios are 0.09760 and 0.09991. Mean attention is not matched: first/non-first means are $0.4825/0.03878$ for Qwen and $0.5054/0.09923$ for Mistral. Mistral's evaluation pairs are also uneven across heads: 3, 7, 17, and 52 for heads 8, 10, 11, and 31. These restrictions prevent extrapolation to unmatched events or uniform head-level effects.

Prediction contrasts and interpretation. At $\alpha = 1$, the first/non-first gain intervals are $[1.15, 14.55]\%/[8.22, 27.28]\%$ for Qwen and $[8.18, 33.90]\%/[16.93, 30.50]\%$ for Mistral. The absolute MSE-improvement differences are -5.12×10^{-7} [-1.025×10^{-6}, -3.65×10^{-8}] and -1.54×10^{-6} [-4.219×10^{-6}, 5.20×10^{-8}], respectively. Thus Qwen's smaller first-token increment is present for both absolute and relative predictive improvements; Mistral does not resolve either difference. All fitted penalties select 100, the upper end of the pre-specified ridge grid, further motivating a fixed-predictor-family interpretation rather than an information-theoretic claim. Non-first keys can themselves be sinks, source content and direction differ, and the comparison does not identify why either group benefits from brightness. It also uses different samples and features from the original score-blind experiment; its first-token gains do not replace Table 5.

A.18 Historical panels and supplementary coordinate audits

Figure 2 retains the original four-model panel for provenance. Its Qwen site is L26, not the current primary L28 site; original-run magnitudes must not be substituted for the strict fp32 estimates in Table 1. The Llama panel is L20, not the exploratory L31 follow-up.

Figure 3 displays the same original sites as aggregate brightness–dose coordinates. Near-zero aggregate dose log-ratios do not imply that every constituent pair satisfies a strict caliper. Large brightness ratios in score-selected pairs are descriptive, not independent evidence of population-wide causal-effect separation.

Native value geometry. In the Qwen native audit over 3,072 layer/head/text sites, median unit-brightness percentile relative to random directions is 0.406. High-mass dark sites have median percentile 0.0469 ($n = 233$), while 505 sites exceed the 0.90 percentile. In the associated μ/u decomposition, the aggregate $\log B(\hat{\mu})/B(\hat{u})$ is near zero and the fraction $\hat{\mu}$-brighter is about 0.53; the component-scaled median ratio is approximately $\exp(-2) = 0.13$. These quantities characterize a selected decomposition, not the predictive performance of actual contribution weights. The latter comparison is tested separately in Section 5.2.1.

Operational proxies. Figure 4 compares researcher-defined high-attention, low-transport-dose and high-transport-dose proxies. Their different suppression phenotypes show that an operational selection rule changes the sampled population. They do not identify the mechanisms or reproduce the tasks studied by earlier rest-state and transport accounts.

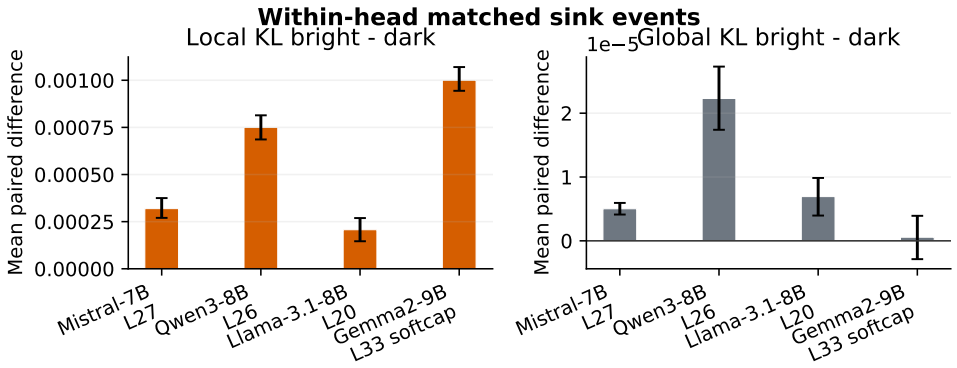

Figure 2: **Historical within-head matched-event panel.** Original comparisons at Mistral L27, Qwen L26, Llama L20, and Gemma2 L33. This panel summarizes score-selected pairs and their local deletion gaps; it is not a plot of the subsequent strict fp32 Qwen L28 confirmation or its actual-intervention placebo.

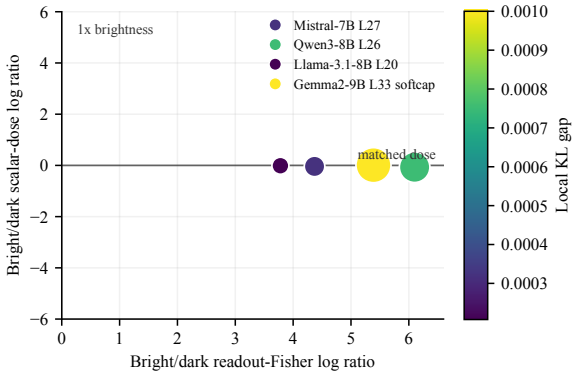

Figure 3: **Historical brightness–dose summary.** The original Mistral L27, Qwen L26, Llama L20, and Gemma2 L33 comparisons at $\alpha = 1$. Axes report aggregate bright/dark brightness and scalar-dose log-ratios; marker color/size encodes the local KL gap. Qwen L28 is not represented.

A.19 Numerical coverage of a downstream-response probe

We conduct a supplementary numerical audit, separate from the finite-dose interventions and held-out prediction results. For each primary model, a fixed hash-based subset of the score-blind evaluation pool contains 16 texts and 128 events, without consulting brightness or intervention outcomes during subset selection. Let $z_e(t)$ be the affected query's logits after adding tv_e at the attention output and running the remaining network. We approximate the downstream directional response by $d_\epsilon = [z_e(\epsilon) - z_e(-\epsilon)]/(2\epsilon)$. All forward passes use fp32; logit differences and Fisher-weighted calculations use float64. This does not recover information already lost in fp32 forward computation.

For a logit vector u, define $\|u\|_p^2 = \sum_i p_i(u_i - E_p[u])^2$ using the clean predictive distribution. An event is numerically consistent for a step pair when its response is nondegenerate and $\|d_\epsilon - d_{\epsilon/2}\|_p / \|d_{\epsilon/2}\|_p \leq 0.1$. After the initial small-step check showed incomplete coverage, a fixed-sample numerical follow-up evaluated two larger-step pairs, retaining all events and the original results (Table 25).

The recorded audit rule required at least 90% coverage; neither model reaches this aggregate criterion, so the audit retains an inconclusive status rather than being relabeled after inspecting results. The larger-step checks improve coverage, but finite-scale agreement does not establish an exact $\alpha \to 0$ Jacobian or identify a propagation mechanism. Accordingly, we do not use the resulting direction cosines or energy ratios as confirmatory evidence for downstream rotation, amplification, or causal

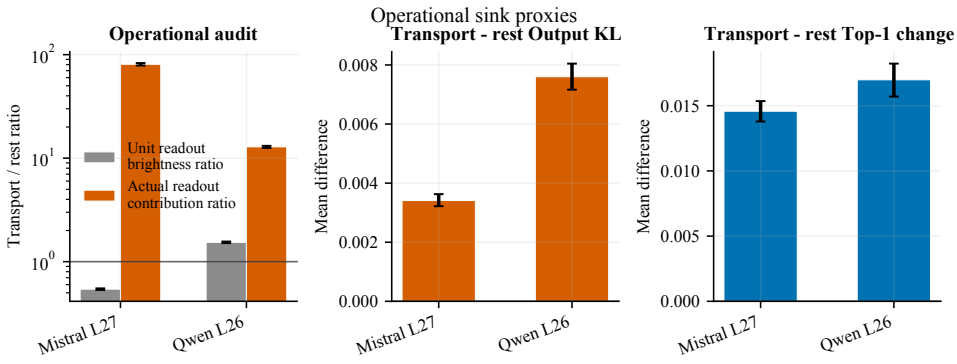

Figure 4: **Supplementary operational-proxy audit.** Transport-style proxies have larger Fisher-weighted contribution and suppression effects than rest/key-bias proxies. The Qwen contrast includes dose and unit brightness, whereas the Mistral contrast is primarily dose-driven. This is a comparison of constructed proxies, not a causal unification of prior sink mechanisms.

Site	.02/.01	.10/.05	.05/.025	Both larger pairs
Qwen L28	82/128	127/128	114/128	114/128 (89.06%)
Mistral L27	96/128	126/128	114/128	114/128 (89.06%)

Table 25: Numerical consistency coverage for fixed natural-event directional responses. Columns report events satisfying the 10% relative-change criterion, not predictive accuracy or the probability that a mechanism is correct. The final column requires the same event to satisfy both larger-step checks.

mediation. These numerical limitations do not change the separately measured finite-dose outcomes in the main experiments.

B Extended Related Work

Attention sinks and sink functions. Prior work has described attention sinks as key-bias or rest-state phenomena, semantic tagging and release mechanisms, geometry-based selection signals, architectural consequences of softmax normalization, compression-valley correlates, and input-conditioned active–dormant head behavior (Xiao et al., 2024; Gu et al., 2025; Zhang et al., 2026; Shin et al., 2025; Qiu et al., 2026; Guo et al., 2024b; Ran-Milo, 2026; Queipo-de Llano et al., 2025). Compression-valley analyses connect sinks to macro-scale residual compression and depth dynamics (Queipo-de Llano et al., 2025), while Omni-LLM analyses show that sink value representations can carry global signals despite high sink attention (Yoo et al., 2026). These emergence and information-content questions differ from our controlled-dose deletion estimand; our experiments do not reproduce their mechanisms or resolve their task-level conclusions.

Conditional sink computations and multimodal interventions. Fesser et al. (2026) distinguish adaptive no-op from broadcast computations using synthetic tasks and pretrained vision transformers; adaptive no-op already allows receiving-token dependence. Brightness instead measures a residual write's direct output sensitivity, not whether it implements either computation. Choi et al. (2026) study global information in multimodal sink representations and task-dependent effects of key scaling. Their information and task-utility measurements differ from our single-event output-KL estimand, and key scaling also changes attention routing. DIVA uses contextual attention probes for instruction-aware vision-token pruning (Kim et al., 2026), a policy evaluation distinct from our diagnostic.

TAPPA analyzes query self-similarity across decoding steps in the query–key (QK) circuit with rotary positional embeddings (RoPE) to explain temporal attention patterns (Yang et al., 2026). Our question concerns the received value's output effect after controlling direction and dose, rather than which past position attracts attention.

Contribution and value-aware attention analyses. Several methods refine raw attention weights with value norms, key/value norm heuristics, output-alignment terms, sink-token hidden-state geometry, or sparse decompositions of attention superposition (Guo et al., 2024c; Devoto et al., 2024; Cunningham & Cirone, 2026; Shin et al., 2025; He et al., 2025). Contribution-weight analyses are especially relevant because they incorporate value alignment with head or attention output, not merely attention mass or value norm (Cunningham & Cirone, 2026). Their signed and absolute intra- and inter-head scores are therefore substantive competing predictors. Section 5.2.1 compares these features against actual deletion outcomes, rather than relying on descriptive differences between coordinate systems. Fisher's primary-site ridge gains are conditional on the predictor and sampling frame, not evidence that contribution weights are intrinsically incorrect or a reproduction of their full token-removal benchmark.

No-op, inactive-head, and pruning views. Several studies treat sink-dominated heads as redundant or approximately no-op components and use sink statistics to identify heads that can be pruned or ablated with limited benchmark loss (Sok et al., 2026; Sandoval-Segura et al., 2025). Our readout-dark label concerns event-level readout sensitivity, whereas pruning requires a separate policy and utility evaluation. Other concurrent work treats the initial sink token as an actionable control surface for attention-logit tuning (Han et al., 2025), or learns KV-retention scores that recover sink-token heuristics among other long-memory patterns (Bui et al., 2025). Our eviction and event-selection probes test this distinction without establishing a competing compression method.

Readout-lens and logit-space diagnostics. Our direct readout-Fisher score is a second-order, distributional cousin of logit-lens and direct-logit-attribution tools (nostalgebraist, 2020; Elhage et al., 2021; Belrose et al., 2023). It also connects to recent logit-direction sensitivity analyses that study the spread of directional logit derivatives under cross-entropy (Sao, 2026). Unlike direct logit attribution to a target token, our score measures the Fisher-weighted variance of the entire direct logit perturbation at the current predictive distribution.

Fisher and non-Euclidean activation geometries. Park et al. (2023) identify a non-Euclidean inner product for concept representations, connecting their geometry to probing and steering. Categorical output geometry and Fisher-based intervention design provide related but distinct perspectives (Park et al., 2026; Wang & Zhao, 2026; Zhou et al., 2026). Information-geometric diffusion work shows how choosing the distribution family indexed by the latent representation can determine whether the induced Fisher–Rao geometry is meaningful or collapsed (Karczewski et al., 2025). We do not propose a competing intervention geometry. Instead, we use the direct output Fisher as a diagnostic coordinate for natural residual contributions.

Faithfulness and causal interpretation. Our boundary results align with warnings that output-preserving localization, subspace interventions, benchmarked mechanism recovery, and mechanism localization are not equivalent (Makelov et al., 2023; Mueller et al., 2025; Guo et al., 2024a). The same component can be quiet under one readout and consequential under another downstream measure.

C Detailed Limitations

Intervention effects are not a mechanism attribution. The matched intervention changes a contribution, not brightness independently of context; both the predictive distribution and downstream Jacobian vary across queries. Neither predictive improvement nor direct/actual KL ratios identify the paths producing the deletion effect. RMSNorm and small-perturbation checks validate the readout calculation, not a propagation mechanism. The supplementary finite-difference audit has joint numerical-consistency coverage of 114/128 events (89.06%) in each primary model, not an exact Jacobian measurement (Appendix A.19).

The sampling frame limits generality. Score-based pairing does not estimate causal-effect prevalence, and score-blind sampling at the original sites retains their selection history. The prospective eight-site audit tests a fixed depth grid, not the population of all heads or layers; its exclusions cover the supplied historical manifests rather than every earlier experiment. Llama L31 is exploratory, not an independent pre-specified replacement for weak L20. Nor is contextual sensitivity established as sink-exclusive. The natural-key comparison finds positive predictive increments for matched non-first-token contributions, a smaller first-token increment in Qwen, and an unresolved difference

in Mistral. Its low evaluation coverage (14.06% for Qwen, 7.71% for Mistral), unequal head representation, and differing source positions limit extrapolation to unmatched events or sink identity itself. All experiments use teacher-forced text rather than generation-time deployment.

Incremental prediction is conditional, not universal superiority. The main prediction tables include calibration refitting, unlike the original fixed-fit intervals retained in the appendix. Gains remain predictor-dependent: the analytic control establishes ridge representation utility, not alignment irrelevance or an information-theoretic decomposition. Its original post-hoc analysis reuses evaluated data and changes the calibration-CV seed; changes in fitted point estimates and interval endpoints must not be treated as independent replications or selectively substituted for earlier results (Appendix A.15). The separate fixed-plan CNN/DailyMail experiment uses new articles but retains the selected sites and refits predictors; its pooled ridge gains are not uniformly supported by head-equal or HGB contrasts (Appendix A.14). The 200-refit tail quantiles have Monte Carlo uncertainty, particularly when interval endpoints are close to zero. Some clean-state controls use observed next-token information, so the regression comparison is an offline diagnostic, not a deployable online selection policy. Original matched panels and precision-controlled reruns must also be distinguished: Qwen L26 has substantial dtype-dependent effect magnitudes, and the current primary Qwen evidence comes from the strict fp32 L28 audit.

Sensitivity is not utility. Output KL measures distributional change, not whether that change helps a task. The primary-dose score-blind selection-benefit intervals cross zero (Appendix A.2), and coarse KV eviction does not distinguish total brightness from value norm when their selected heads coincide. The action-aligned pilot also fails to establish an advantage over norm; its easy low-damage candidates and extremely small policy differences limit both utility and numerical interpretations (Appendix A.16). These results do not establish a pruning, cache-compression, or steering advantage. Query-conditionality does not rule out useful static policies, but our diagnostic alone does not certify one.