Logit-Contribution Scoring Identifies Non-Literal Retrieval Heads[1]

Aryo Pradipta Gema [1] **Beatrice Alex** [2] **Pasquale Minervini** [1 3]

Abstract

In long-context use, large language models frequently synthesize answers from the meaning of a relevant context span rather than literally copy-pasting them. Identifying which attention heads perform this synthesis matters for interpreting long-context model behavior. Yet existing detectors miss these heads by construction: they reward heads whose attended token matches the generated token, a literal-copy criterion that captures *where* a head reads but not *what* it writes through its output-value (OV) circuit, the very mechanism that carries non-literal retrieval. We introduce Logit-Contribution Scoring (LOCOS), a write-aware detector that scores each head by the projection of its OV-circuit output onto the answer-token unembedding direction, contrasting needle and off-needle source positions in a single forward pass. Across three model families (Qwen3, Gemma-3, OLMo-3.1), mean-ablating the top LOCOS heads on the NoLiMa non-literal retrieval benchmark collapses ROUGE-L at lower head counts than prior attention-based detections; on Qwen3-8B, ablating 50 heads drives ROUGE-L from 0.401 to 0.000 while the strongest baseline still retains 0.292. The selected heads are retrieval-specific: parametric recall and arithmetic reasoning stay at baseline under the same ablation. On Qwen3-8B, the same ablation also drops MuSiQue from 0.55 to 0.08 and BABILong from 0.62 to 0.20, while a random-heads control stays within 0.05 of baseline. `locos`

1. Introduction

The ability of large language models (LLMs) to retrieve information from their input context, rather than relying on memorized parametric knowledge, depends on a sparse set of attention heads known as *retrieval heads* (Wu et al., 2025), which build on earlier mechanistic work on induction heads (Elhage et al., 2021; Olsson et al., 2022). However, in practice, context retrieval is rarely literal copy-paste: a user's question may share no lexical overlap with the relevant passage, and the model must identify the relevant snippet, parse its meaning, and synthesize an answer from it (as illustrated in Fig. 1). Yet all existing identification methods, whether via token-matching heuristics (Wu et al., 2025) or weighted attention accumulation (Fu et al., 2025; Lin et al., 2025), evaluate heads on literal copying tasks and share a common observable: each head's attention pattern over source positions. This observable captures *where* a head allocates attention, but does not capture *what information* the head propagates through its output-value (OV) circuit, the very mechanism by which non-literal retrieval is accomplished. Two heads with identical attention patterns but different OV circuits can propagate entirely different information from the same positions. For literal retrieval, where the attended token is the answer token, attention and OV output are trivially aligned, and attention-based scoring works. In non-literal retrieval, a head may attend to "Eiffel Tower" while writing the "Yuki" direction to the residual stream. Attention-based scoring sees the read site ("Eiffel Tower"); the OV circuit determines the write content ("Yuki"), and the two need not agree. This distinction matters in practice, and no prior retrieval-head detector makes this distinction.

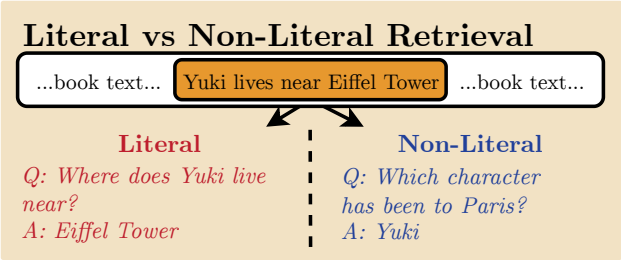

Figure 1. **Non-literal retrieval requires synthesis.** The same context answers two questions differently: a literal question requires reading "Eiffel Tower" directly from the needle, while a non-literal question must produce "Yuki" after synthesizing the context.

Our method, Logit-Contribution Scoring (LOCOS), measures how each attention head contributes to the correct

[1]University of Edinburgh [2]Heriot-Watt University [3]Miniml.AI. Correspondence to: Aryo Pradipta Gema <aryo.gema@ed.ac.uk>, Pasquale Minervini <p.minervini@ed.ac.uk>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

answer token in the unembedding space (See Fig. 2). For each head at each source position, the method computes the scalar projection of the head's weighted OV circuit output onto the correct answer's unembedding vector. Aggregation uses *spatial contrast*: logit contributions from needle positions are compared against length-normalized off-needle contributions within a single decoding step. The method requires only a single forward pass per probing trial.

Ablation experiments on six configurations spanning three model families (*i.e.,* Qwen3 (8B, 14B, 32B) (Team, 2025b), Gemma-3 (12B, 27B) (Team, 2025a), and OLMo-3.1 (32B) (Olmo et al., 2025)) on the NoLiMa non-literal retrieval benchmark heldout set (Modarressi et al., 2025) validate the method: mean-ablating the top-ranked LOCOS heads produces a steeper ROUGE-L degradation curve than all evaluated baselines (§ 4.2). On Qwen3-8B, ablating the top-50 LOCOS heads collapses ROUGE-L from 0.401 to 0.000, while the strongest attention-based baseline retains 0.292 at the same depth. Control experiments confirm retrieval specificity: parametric recall and arithmetic reasoning remain intact under the same ablation (§ 4.6). The same ablation also degrades downstream long-context performance, most strongly on the Qwen3 family (§ 4.8): on Qwen3-8B, ablating the top-50 LOCOS heads drops MuSiQue accuracy from 0.55 to 0.08 and BABILong from 0.62 to 0.20; transfer is most consistent on the Qwen3 family, and the ranking against attention-based baselines is benchmark-dependent on Gemma-3 and OLMo-3.1 (§ 4.8).

2. Background

Notation. Consider a transformer (Vaswani et al., 2017) with L layers and H attention heads per layer, head dimension d_h, and model dimension $d = H \cdot d_h$. Head (l, h) attends to source positions with weights $\alpha_{t,j}^{(l,h)}$, reads value vectors $\mathbf{v}_{t,j}^{(l,h)} \in \mathbb{R}^{d_h}$, and writes to the residual stream via its output projection $W_O^{(l,h)} \in \mathbb{R}^{d \times d_h}$. The per-position output of head (l, h) from source position j is:

$$\mathbf{o}_{t,j}^{(l,h)} \;=\; \alpha_{t,j}^{(l,h)} \cdot W_O^{(l,h)}\, \mathbf{v}_{t,j}^{(l,h)} \;\in\; \mathbb{R}^d\,. \tag{1}$$

The unembedding matrix $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ maps the residual stream to logits; $\mathbf{u}_y \in \mathbb{R}^d$ is its y-th row.

Attention heads as read-and-write circuits. An attention head decomposes into a *QK circuit* that determines $\alpha_{t,j}^{(l,h)}$ (where the head reads) and an *OV circuit* that maps each source value through W_O to the residual stream (what the head writes) (Elhage et al., 2021). The full output of head (l, h) at step t is $\sum_j \mathbf{o}_{t,j}^{(l,h)}$, and the next-token logits come from mapping the final-layer residual stream through W_U. A head is useful for retrieval only when both stages agree: the QK circuit must select the right source positions *and* the OV circuit must write an answer-aligned update.

Induction heads and retrieval heads. Induction heads implement the literal copy pattern $[A][B] \ldots [A] \mapsto [B]$ (Olsson et al., 2022): they place high $\alpha_{t,j}^{(l,h)}$ on a previously matching position and write an output whose projection onto $\mathbf{u}_{y_t}$ is large because the attended token *is* the next token. Retrieval heads, identified on literal needle-in-a-haystack (NIAH) prompts by Wu et al. (2025), generalize this picture to long-context factuality. Their detection procedure rewards heads whose argmax attention falls inside the needle and the attended token matches the generated token, a literal-copy criterion.

Literal versus non-literal retrieval. In a NIAH setup (Kamradt, 2023), a *needle* is a short answer-bearing span inserted into a longer distractor context (the *haystack*); we index it as $[s_\tau, e_\tau)$ for trial τ and refer to all other source positions as *off-needle*. In literal NIAH, the answer token appears in the needle, so large needle attention $\alpha_{t,j}^{(l,h)}$ usually coincides with a large logit-relevant write $\mathbf{u}_{y_t}^\top \mathbf{o}_{t,j}^{(l,h)}$. NoLiMa breaks this equivalence (Modarressi et al., 2025): the answer must be recovered from the meaning of the needle and may share no lexical overlap with it. A head can then retrieve non-literal information by attending to a semantically relevant phrase inside $[s_\tau, e_\tau)$ and writing an answer-aligned direction even when no attended token matches y_t. This is the regime targeted by LOCOS.

3. Logit-Contribution Scoring

We introduce LOCOS, which scores each head by what it *writes* toward the answer rather than where it allocates attention, because non-literal retrieval transforms attended content through the OV circuit before it becomes an answer. We define a three-step procedure:

Per-Position Logit Contribution. Consider a probing trial τ with needle span $[s_\tau, e_\tau)$ embedded in a context of N_τ tokens. Let $\mathcal{A}^\tau$ denote the set of decoding steps at which the model generates a correct answer token (identified by matching against the tokenized gold answer), $y_t \in \mathcal{V}$ the correct token at step t, and N_t the total number of key positions available at step t. The contribution of source position j through head (l, h) to the logit of y_t is:

$$\phi_{t,j}^{(l,h)} \;=\; \mathbf{u}_{y_t}^\top\, \mathbf{o}_{t,j}^{(l,h)} \;=\; \alpha_{t,j}^{(l,h)} \cdot \mathbf{u}_{y_t}^\top W_O^{(l,h)}\, \mathbf{v}_{t,j}^{(l,h)} \;\in\; \mathbb{R}\,. \tag{2}$$

The scalar $\phi_{t,j}^{(l,h)}$ depends on both *where* the head reads (via $\alpha_{t,j}^{(l,h)}$) and *what* it extracts (via $W_O^{(l,h)}\mathbf{v}_{t,j}^{(l,h)}$). A head that attends strongly to a position whose OV output is orthogonal to $\mathbf{u}_{y_t}$ receives $\phi \approx 0$ despite high attention; conversely, a head performing non-literal retrieval (*e.g.,* attending to "Paris" to produce "France") receives large ϕ because its OV circuit transforms the attended representation into an answer-aligned output.

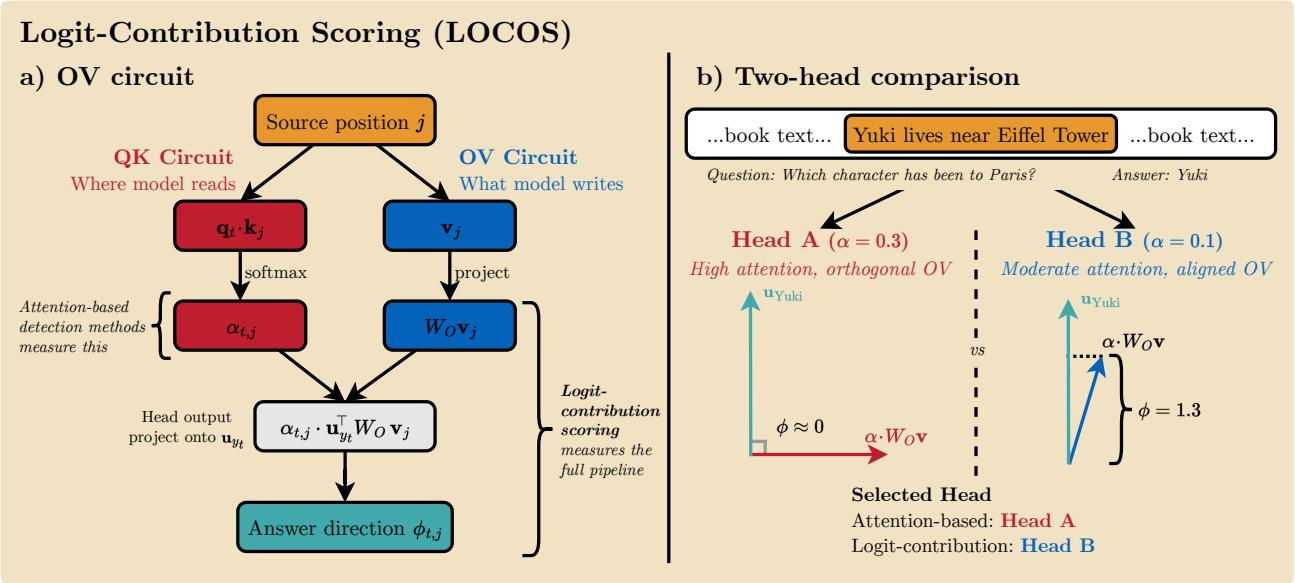

Figure 2. **An attention head has two circuits: where it reads (QK) and what it writes (OV). Logit-Contribution Scoring uses the OV circuit to identify *non-literal retrieval heads*. (a)** Anatomy of a head's per-position output: the QK circuit produces attention weight $\alpha_{t,j}$; the OV circuit produces $W_O \mathbf{v}_j$. Attention-based methods measure only α. Logit-contribution scoring (LOCOS) measures $\phi = \mathbf{u}_{y_t}^\top (\alpha \cdot W_O \mathbf{v}_j)$, capturing the full pipeline. **(b)** Consequence for non-literal retrieval: two heads read from "Eiffel Tower" to answer "Yuki." Head A attends strongly (α=0.30), but its OV output is orthogonal to the answer direction ($\phi \approx 0$). Head B attends moderately (α=0.08) but writes toward the answer (ϕ=1.3). Attention-based methods select Head A; LOCOS selects Head B.

Spatial Contrast. For each head (l, h) at answer step t in trial τ, we define the logit contribution from needle and off-needle positions:

$$\begin{aligned} \Phi_t^{(l,h),+} &= \textstyle\sum_{j=s_\tau}^{e_\tau - 1} \phi_{t,j}^{(l,h)}, \\ \Phi_t^{(l,h),-} &= \tfrac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} \textstyle\sum_{j \notin [s_\tau, e_\tau)} \phi_{t,j}^{(l,h)}. \end{aligned} \tag{3}$$

where the rescaling factor $(e_\tau - s_\tau)/(N_t - (e_\tau - s_\tau))$ makes $\Phi_t^{(l,h),+}$ and $\Phi_t^{(l,h),-}$ comparable: both represent the logit contribution of a region of length $(e_\tau - s_\tau)$. The contrast $\Phi_t^{(l,h),+} - \Phi_t^{(l,h),-}$ is *spatial*: it compares needle and off-needle positions within a single decoding step, rather than the *temporal* contrast (answer vs. non-answer steps) used by attention-based methods. Spatial contrast yields a score from a single answer step, identifies heads that attend to the needle persistently but write answer-relevant content only from needle positions, and cancels uniform contributors such as token-frequency priors whose $\phi_{t,j}^{(l,h)}$ is large but position-independent.

Aggregation. We pool over all answer steps across all trials passing a correctness filter (ROUGE-1 recall $> \rho$, default $\rho = 0.5$), without per-trial normalization. Let $\mathcal{D}_{\text{pass}} \subseteq \{1, \ldots, T\}$ denote the set of passing trials. The final score is:

$$S_{l,h} = \frac{1}{\sum_{\tau \in \mathcal{D}_{\text{pass}}} |\mathcal{A}^\tau|} \sum_{\tau \in \mathcal{D}_{\text{pass}}} \sum_{t \in \mathcal{A}^\tau} \left(\Phi_t^{(l,h),+} - \Phi_t^{(l,h),-} \right). \tag{4}$$

The score $S_{l,h}$ is the mean over all (trial, answer-step) pairs, with each answer step weighted equally. Unlike attention-based methods, we do not clamp $S_{l,h}$ at zero: negative values indicate heads whose logit contribution toward the correct answer originates predominantly from off-needle positions, which points to factors other than needle-specific retrieval (*e.g.,* parametric or contextual associations). A worked example with Per-trial consistency diagnostics is reported in Appx. G. LOCOS recovers attention-based scoring as a special case when the OV circuit contributes no position-dependent signal.[2]

4. Experiments

4.1. Experimental Setup

Probing benchmark. NoLiMa (Modarressi et al., 2025) is a *needle-in-a-haystack* benchmark in which each trial embeds a factual statement in a long context and poses a question whose answer requires understanding the needle's meaning rather than copying a literal token from it. Its non-literal retrieval property expose weaknesses of attention-based scoring. We use the `onehop` reasoning, 10 context lengths $\times$ 10 insertion depths, and 3 characters per entry (context range 1,000–5,000 tokens). We evaluate LOCOS on six models from three families: Qwen3 (8B, 14B, 32B) (Team, 2025b), Gemma-3 (12B, 27B) (Team,

[2] Appx. O works through the reduction.

2025a), and OLMo-3.1 (32B) (Olmo et al., 2025). Trials with ROUGE-1 recall > 0.5 against the gold answer are retained, following Wu et al. (2025).

Causal validation via mean-ablation. For each selected head (l, h) and at every decode step t, we replace the post-Q-projection, pre-RoPE query vector with a head-specific calibration vector $\bar{\mathbf{q}}^{(l,h)}$ (computed once over a 50-trial sample; see Appx. B); we then evaluate ROUGE-L on a disjoint held-out set of 800 NoLiMa trials. Mean replacement keeps downstream-layer activations in-distribution (Nanda et al., 2023; Wang et al., 2023). Architecture-specific implementation details are given in Appx. H.

Baselines. We compare LOCOS against: (i) **Random**: uniformly sampled heads, used to check that non-trivial ablation effects require targeted selection. (ii) **Wu/NIAH-scored**: the token-matching retrieval score of Wu et al. (2025) computed on NIAH dataset; and (iii) **Wu/NoLiMa-scored**: the same token-matching criterion computed on NoLiMa probing trials; The suffix on *Wu/X-scored* denotes the dataset used to detect the retrieval heads.

4.2. Ablation Comparison Across Scoring Methods

Fig. 3 compares the four methods under mean-ablation of the top-k heads. On Qwen3-8B, ablating the top-5 LOCOS heads already reduces ROUGE-L to 0.321 (baseline $= 0.401$), while ablating Wu/NIAH heads remains at 0.406. By k=50, LOCOS reaches 0.000, whereas Wu/NIAH-scored still achieves 0.292 and Wu/NoLiMa-scored 0.337. Random-head ablation remains near baseline throughout (0.358–0.402); the degradation is therefore specific to head selection. Wu/NoLiMa-scored-ranked heads ablation *raises* ROUGE-L above baseline at intermediate k (0.428 at k=5), which indicates that its token-matching criterion selects causally irrelevant heads.

Takeaway 1: LOCOS heads produce the steepest ROUGE-L ablation curve on NoLiMa, reaching near-zero output at lower k than all evaluated baselines.

4.3. Isolating the OV Contribution: An Attention-Only Spatial-Contrast Control

LOCOS differs from Wu et al. (2025) in both the per-position observable (OV projection ϕ *v.s.* attention-based token matching) and aggregation (spatial *v.s.* temporal contrast). To isolate the OV contribution, we substitute α for ϕ in Equations (3) and (4), yielding an *attention-only spatial-contrast* score $S^{\text{att}}_{l,h}$ that matches LOCOS's aggregation but removes the OV projection (Proposition 1 and Appx. O.3).

Fig. 4 shows that LOCOS is more damaging on Qwen3-8B, Qwen3-32B, and Gemma-3-12B: at k=50, it reaches 0.000 versus 0.148 on Qwen3-8B and 0.077 versus 0.346 on Qwen3-32B. The scorers are comparable on Qwen3-14B and OLMo-3.1-32B, while Gemma-3-27B inverts at depth: the attention-only control is more damaging at $k \in \{20, 50\}$ (0.118 vs. 0.369 at k=20).[3] Thus, OV projection primarily reduces model-specific failures of head selection rather than uniformly increasing effect size.

Takeaway 2: Under identical spatial-contrast aggregation, an attention-only control is competitive on average but more variable across models. The OV projection enables LOCOS to identify head sets whose group ablation is highly damaging across all six configurations.

4.4. Bottom-k Control: Spatial Source Matters

A potential objection to the ablation result is circularity: the method selects heads whose OV output projects onto $\mathbf{u}_{y_t}$, so ablating them mechanically reduces the y_t logit regardless of whether the heads perform retrieval. If this were the case, *any* set of heads with large answer-aligned logit contribution should be equally causal when ablated, *irrespective of whether that contribution originates from the needle or from unrelated context*. Heads with the most negative spatial contrast scores ($S_{l,h} \ll 0$) (*i.e., bottom-k heads*) provide a direct test: they contribute strongly to the correct-answer logit but predominantly from off-needle positions. The absolute logit contribution of these bottom-k heads is large, so the comparison with top-k heads is not confounded by magnitude. The full score distribution (Appx. I) confirms that all bottom-50 heads have strictly negative $S_{l,h}$ in every model, so the bottom-k experiments exclusively target heads with off-needle-dominant contributions.

As shown in Fig. 5, mean-ablating the top-k LOCOS heads produces steep ROUGE-L degradation, reaching near-zero at k=50 in most models. By contrast, ablating the bottom-k heads leaves ROUGE-L near baseline, tracking the random-head control at all ablation depths.

Takeaway 3: Ablating bottom-k heads (those contributing to the answer logit from off-needle positions) does not degrade NoLiMa ROUGE-L. This indicates that the ablation result is not merely an artifact of removing answer-aligned signal.

4.5. Layer Distribution

Fig. 6 visualizes score distributions across layers for Gemma-3-27B and Qwen3-32B. LOCOS is most concentrated in the Qwen3 family and Gemma-3-27B; at KV-group granularity, Gemma-3-12B and OLMo-3.1-32B span

[3] We analyze the Gemma-3-27B inversion in Appx. N.5, where tuned-lens-corrected projections test whether the direct-path assumption explains this inversion.

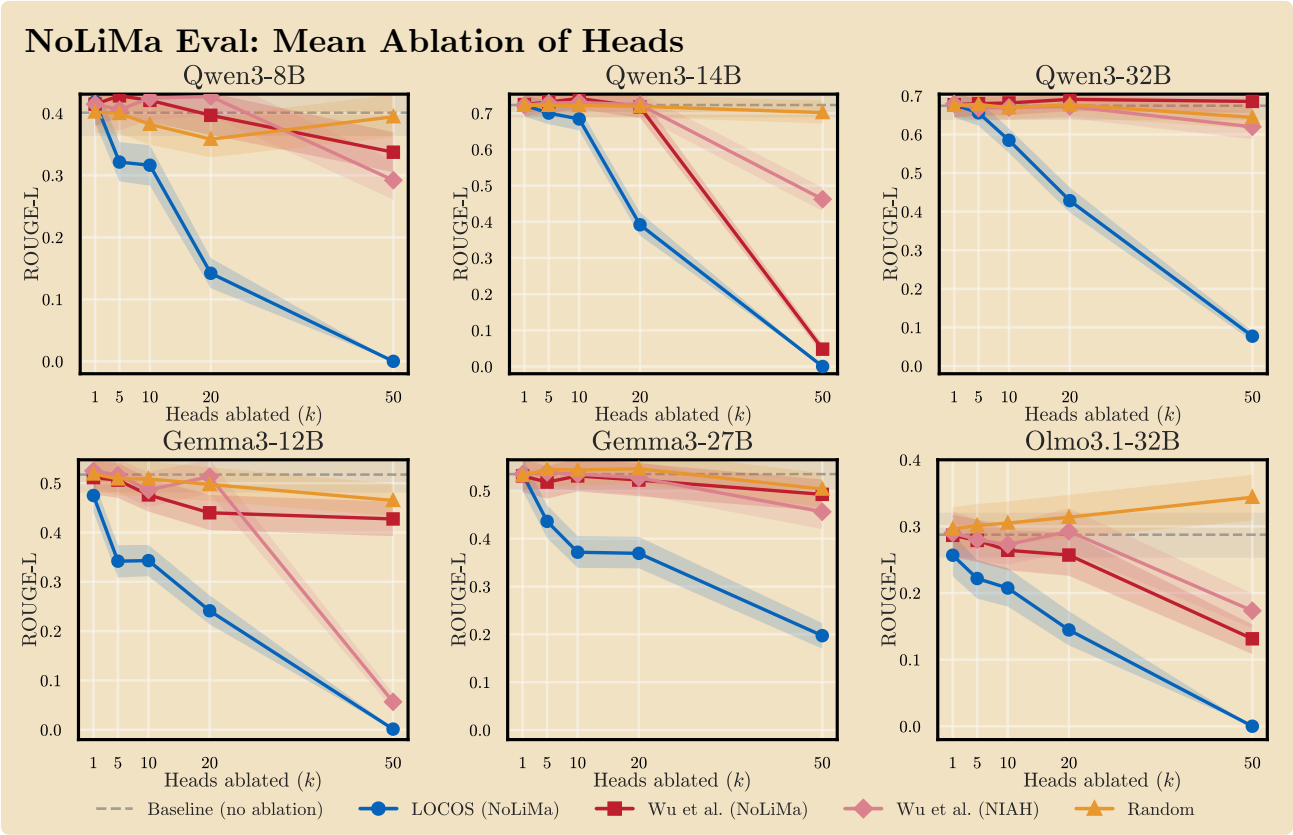

Figure 3. **LOCOS heads produce steeper ROUGE-L degradation under mean-ablation across all six models.** Each panel shows NoLiMa ROUGE-L (800 trials) as a function of the number of ablated heads k for four scoring methods across three model families at two scales each: Qwen3 (8B, 14B, 32B), OLMo-3.1 (32B), and Gemma-3 (12B, 27B). LOCOS (blue) produces the steepest degradation curve in every model, reaching near-zero ROUGE-L by k=50 in five of six configurations and severe degradation (≈ 0.1) in Qwen3-32B.

broader layer ranges (Appx. K). One possible explanation is that LOCOS relies on a *direct-path assumption* that is more accurate near the output, biasing it toward late layers.

A tuned-lens check on Gemma-3-27B preserves the late-layer band, and causal activation patching on Qwen3-8B and Gemma-3-12B also concentrates top-10 heads in upper layers (Appxs. N.4 and N.6). The top-10 sets overlap only marginally between the two scores (2/10 on Qwen3-8B, 3/10 on Gemma-3-12B); the top-k LOCOS set should therefore be read as a collectively causal retrieval circuit, validated by the group ablations of § 4.2, rather than a list of individually load-bearing heads.

Takeaway 4: LOCOS scores concentrate in late layers in the Qwen3 family and Gemma-3-27B; the layer pattern is family-dependent. A tuned-lens variant and a causal-attribution probe confirm the late-layer concentration is not a direct-path artifact on the models examined; the top-k LOCOS set should be read as a collectively causal retrieval circuit instead of a universal architectural pattern across families.

4.6. Retrieval Specificity

A potential concern is that LOCOS identifies generically important heads for model output, not retrieval-specific. We test this by ablating LOCOS heads and measuring performance on tasks that do not require retrieval: (i) **City–country associations**: parametric factual recall (*e.g.,* "Which country does Paris belong to?"). (ii) **PopQA** (Mallen et al., 2023): top 100 popular QA drawn from a knowledge base (*e.g.,* "Who is the mother of Jesus?"). (iii) **Arithmetic**: two–operand addition and subtraction (*e.g.,* "What is 47 + 23?").

To compare specificity across methods, we define the *Dissociation Score* at ablation depth k as $\mathrm{DS}(k) = \Delta R(k) - \Delta P(k)$, where $\Delta R(k) = (R_0 - R(k))/R_0$ and $\Delta P(k) = (P_0 - P(k))/P_0$ are the relative drops in NoLiMa ROUGE-L and aggregate parametric accuracy (mean of city–country, PopQA, and arithmetic), respectively. R_0, P_0 are baselines. DS=1 means retrieval is fully destroyed with zero parametric damage; $k^* = \arg\max_k \mathrm{DS}(k)$ identifies the most-specific ablation depth.

Fig. 7 shows $\mathrm{DS}(k)$ for each scoring method, with LOCOS

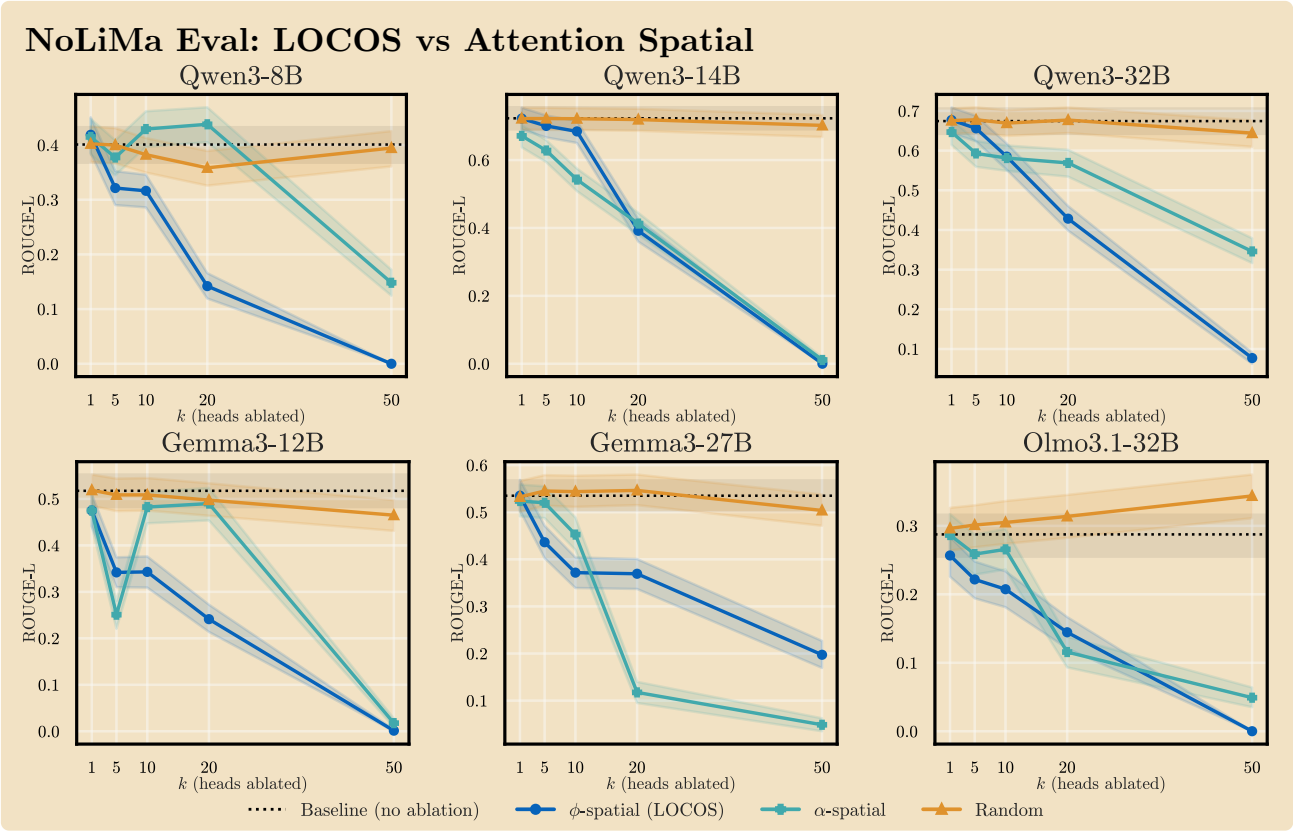

Figure 4. **OV projections improve causal head selection on most models.** Each panel shows NoLiMa ROUGE-L (800 held-out trials) under mean-ablation of the top-k heads ranked by LOCOS (blue) and the attention-only control (cyan). Both scorers use identical spatial-contrast aggregation; only the per-position observable differs. LOCOS is stronger on Qwen3-8B, Qwen3-32B, and Gemma-3-12B, comparable on Qwen3-14B and OLMo-3.1-32B, and weaker at large k on Gemma-3-27B.

achieving the highest peak in every model. Parametric accuracy remains stable under bottom-k ablation (Appx. J), so these heads are not responsible for parametric recall.

Takeaway 5: Retrieval heads identified by LOCOS are retrieval-specific: ablating them severely degrades contextual retrieval while leaving parametric accuracy near baseline, and LOCOS achieves the highest dissociation-score peak in every configuration.

4.7. Literal vs. Non-Literal Retrieval Specificity

§ 4.6 established that ablating LOCOS heads degrades non-literal retrieval far more than parametric tasks. A write-aware detector that measures the full OV pipeline should, in principle, identify retrieval-relevant heads in both literal and non-literal regimes; the novel contribution over attention-based scoring is access to answer-aligned writes that are not predictable from literal token matching. We test this directly by mean-ablating the same top-k LOCOS heads (selected on NoLiMa probing trials) and evaluating on both NoLiMa and NIAH (Kamradt, 2023), treating the two benchmarks as probes for the non-literal and literal retrieval circuits respectively, using the protocol of § 4.1 for both benchmarks.

Fig. 8 shows that the non-literal curve drops faster than the literal curve in all configurations, but the literal curve also drops in five of six models, so the top-k LOCOS set is not strictly non-literal-specific. Both retrieval regimes require heads that read the answer span; they differ only in what the OV circuit writes. The top-k set can mix heads whose OV writes transfer across regimes with heads whose answer-aligned contribution is stronger in NoLiMa; this explains why both curves drop, but NoLiMa drops more. LOCOS therefore identifies retrieval-relevant heads in both regimes; its novelty relative to attention-based scoring lies in identifying write-aware contributions that literal token matching misses.

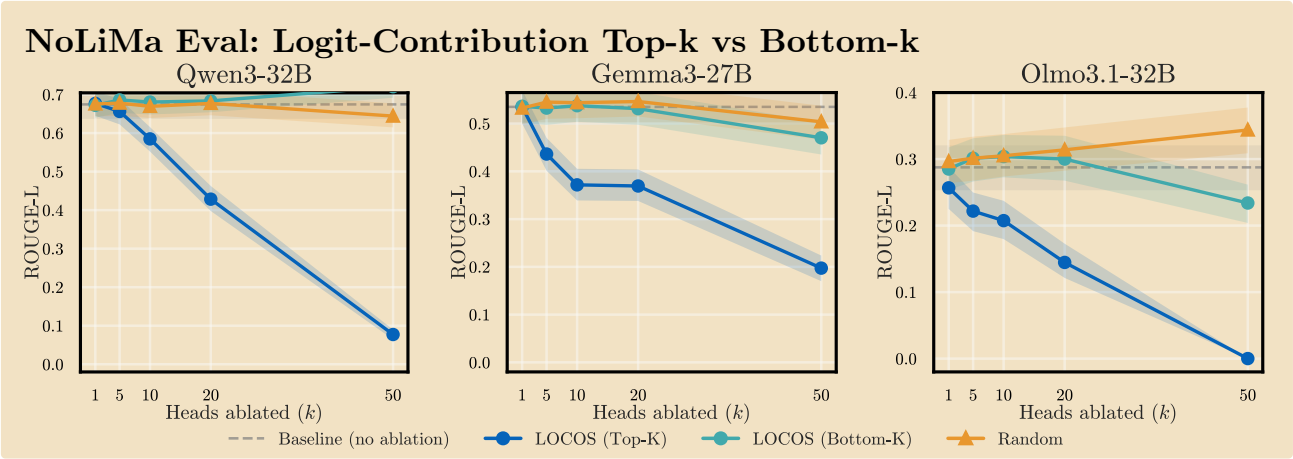

Figure 5. **Bottom-k ablation does not degrade retrieval.** NoLiMa ROUGE-L under top-k (blue), bottom-k (cyan), and random-head (orange) ablation for three representative models; the full six-model version is in Appx. L. Bottom-k heads track the random baseline despite large absolute logit contribution, ruling out the circularity objection.

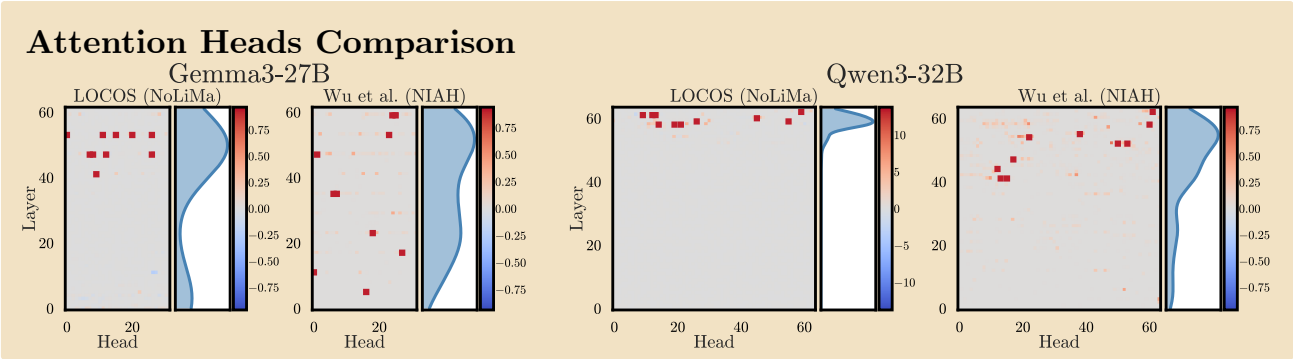

Figure 6. **LOCOS heads are more concentrated in late layers than Wu/NIAH-scored scores.** Layer × Head heatmaps on NoLiMa for Gemma-3-27B and Qwen3-32B; red squares mark top-10 heads. Both scores are high mainly in late layers, but Wu/NIAH-scored also identifies early-to-middle-layer heads.

Takeaway 6: Ablating top-k LOCOS heads degrades both non-literal (NoLiMa) and literal (NIAH) retrieval, with a steeper drop on NoLiMa. The dissociation indicates that LOCOS captures retrieval-relevant contributions beyond those identified by attention-based method, including answer-aligned writes.

4.8. Downstream Long-Context Evaluation

We evaluate the head ablation on two downstream long-context benchmarks that were not used to score heads. **MuSiQue** (Trivedi et al., 2022) is a multi-hop QA benchmark whose questions require composing facts across multiple paragraphs of a long supporting context. **BABILong** (qa2 + qa3 subsets) (Kuratov et al., 2024) requires tracing an object's trajectory through movements interleaved with distractor narrative; literal token copying is insufficient because each candidate location is mentioned many times for many entities (example in Appx. M). We fix k=50 for direct comparison with the headline NoLiMa result and retain the random-heads control and the Wu/NIAH-scored baseline.

Fig. 9 shows that ablating the top-50 LOCOS heads on Qwen3-8B drops MuSiQue accuracy from 0.55 to 0.08 and BABILong from 0.62 to 0.20. Across the twelve model–benchmark cells, LOCOS is the most damaging ablation in six and produces a drop of at least 0.10 below baseline in eight, while the random-heads control stays within 0.05 of baseline in eleven; the degradation is therefore specific to the heads LOCOS identifies. The ranking against attention-based scoring is benchmark-dependent: on MuSiQue, Wu/NIAH-scored ablation is more damaging on Gemma-3 (12B, 27B) and OLMo-3.1-32B, consistent with multi-hop QA recruiting both literal and non-literal retrieval; on BABILong, Wu/NIAH-scored ablation slightly raises accuracy on Qwen3-14B, Gemma-3-27B, and OLMo-3.1-32B, while LOCOS ablation drops accuracy in all models.

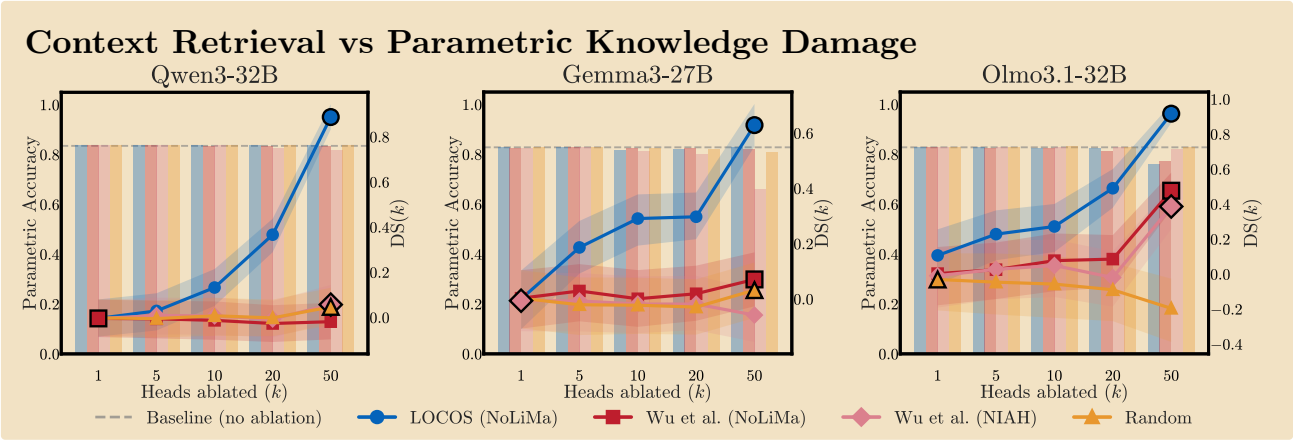

Figure 7. **LOCOS heads exhibit the strongest functional dissociation between retrieval and parametric capabilities.** DS(k) (lines, right axis) and parametric accuracy (bars, left axis) under four scoring methods on three representative models; the full six-model version is in Appx. L. LOCOS (blue) achieves the highest DS in every model configuration; the enlarged marker indicates k^*.

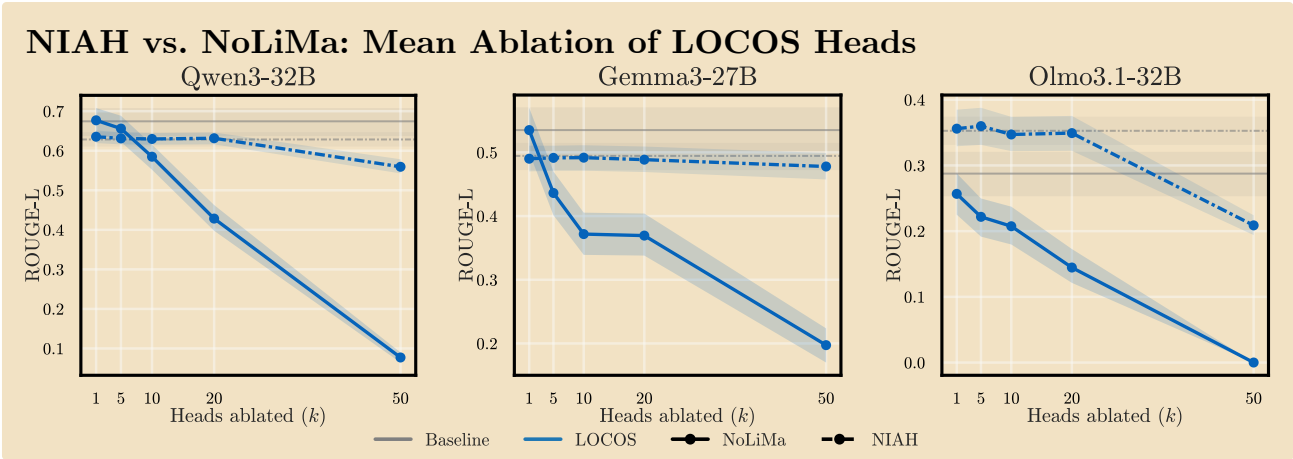

Figure 8. **Ablating LOCOS heads damages non-literal retrieval more than literal retrieval.** ROUGE-L on NoLiMa (solid) and standard NIAH (dashed) under mean-ablation of the same top-k LOCOS heads; baselines are gray. The NoLiMa curve declines more steeply in every configuration, reaching near-zero at k=50 in five of six models.

Takeaway 7: Ablating top-50 LOCOS heads degrades downstream long-context performance, most strongly on the Qwen3 family (Qwen3-8B MuSiQue: 0.55→0.08, BABILong: 0.62→0.20)

5. Related Work

Attention is not explanation. The faithfulness of attention weights as an explanation has been contested (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019), with Bastings & Filippova (2020) arguing that gradient-based saliency methods provide more reliable attribution than attention alone. Our method, LOCOS, measures the head's direct write onto the answer logit rather than its attention distribution, aligning with the *task-grounded faithfulness* criteria (Wiegreffe & Pinter, 2019).

OV/QK circuit analysis and attribution. Our method builds on the QK/OV circuit decomposition of Elhage et al. (2021) and generalizes the literal copying mechanism of induction heads (Olsson et al., 2022) to non-literal retrieval. The logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2025) project residual-stream states onto the unembedding matrix at per-layer granularity; LOCOS applies an analogous projection at per-head, per-position granularity with spatial contrast, building on direct logit attribution. McDougall et al. (2023) characterize copy suppression heads whose OV circuits actively inhibit direct token copying, which may be a complementary mechanism to the non-literal retrieval heads identified here.

Retrieval head identification. Existing methods identify retrieval heads through attention-weight observables. Wu et al. (2025) define retrieval heads via token matching

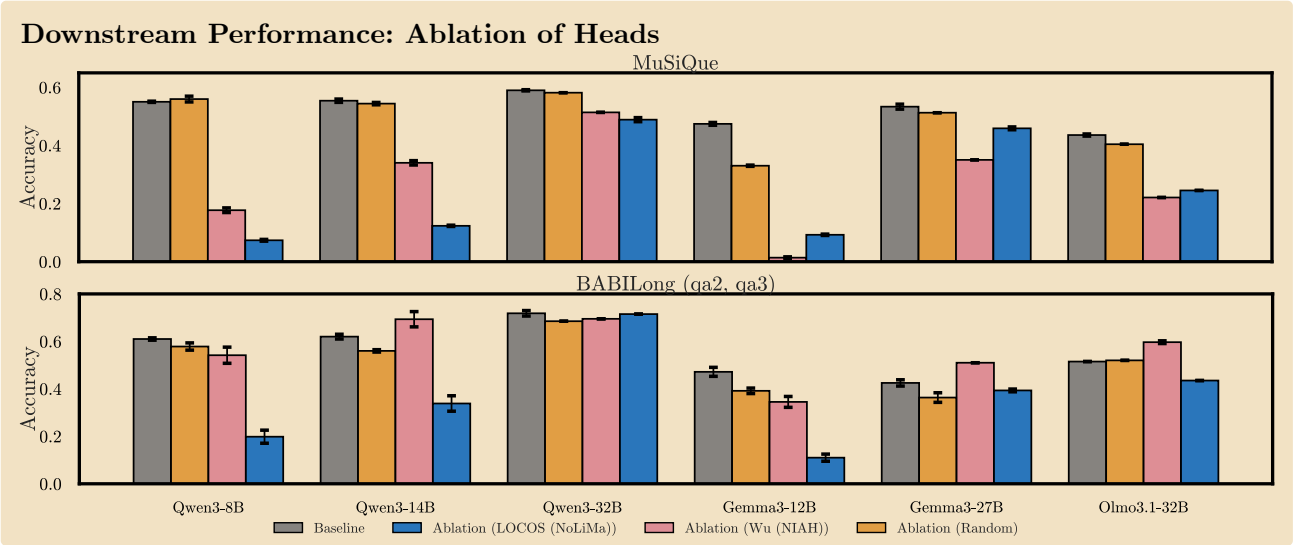

Figure 9. **Mean-ablating top-50 LOCOS heads degrades downstream long-context performance, most strongly on the Qwen3 family.** Accuracy on MuSiQue (top) and BABILong qa2+qa3 (bottom) for six models. Bars show the unablated baseline (gray) and the three ablation conditions: random heads (orange), Wu/NIAH-scored heads (pink), and LOCOS (blue). Error bars are standard deviations across three independent runs. LOCOS produces the largest drop in 6 of 12 model–benchmark cells; Wu/NIAH-scored ablation is more damaging on MuSiQue for Gemma-3 and OLMo-3.1, and slightly raises BABILong accuracy on Qwen3-14B, Gemma-3-27B, and OLMo-3.1-32B.

on needle-in-a-haystack tasks; Fu et al. (2025), Lin et al. (2025), and Xiao et al. (2025) extend attention-based head scoring to KV cache allocation, semantic attention-mass criteria, and retrieval/streaming-head separation. The broader KV cache compression literature (Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024; Cai et al., 2025) motivates per-head retrieval scoring by showing that uniform token eviction misses structured head and layer importance. These methods measure the QK circuit, but not the OV circuit. For literal retrieval, this suffices; for non-literal retrieval, the OV circuit transforms attended content, and attention-based methods miss these heads. Sun et al. (2025) analyze OV circuits for hallucination detection in RAG, decomposing contributions into copying heads (attention) and knowledge FFNs; our method differs by scoring individual heads at per-position granularity with spatial contrast, targeting retrieval head identification.

6. Conclusion

Existing retrieval-head detectors (Wu et al., 2025; Fu et al., 2025; Lin et al., 2025) score attention heads by where they read. This identifies heads that copy literal tokens, but misses cases where heads synthesize answers from the meaning of an attended span. We presented LOCOS (Logit-Contribution Scoring), a detector that scores each head by the projection of its OV-circuit output onto the answer-token unembedding direction, and uses a needle-versus-off-needle spatial contrast to isolate position-specific writes within a single forward pass per probing trial. Across the six configurations evaluated, mean-ablating the top-k heads selected by LOCOS collapses NoLiMa ROUGE-L at lower k than every attention-based baseline we evaluate. The selected heads are retrieval-specific: under the same ablation, parametric recall and arithmetic reasoning remain at baseline. The top-k LOCOS set captures retrieval-relevant heads in both literal and non-literal regimes; its advantage over attention-based scoring is in identifying answer-aligned writes that the token-matching criterion misses (§ 4.7).

Limitations. (i) **Off-needle baseline:** If the context contains distractor information related to the answer, Φ^{-} rises and scores drop for heads performing broad semantic matching rather than targeted needle retrieval—desirable for span-specific retrieval, but may miss heads performing more diffuse contextual integration; (ii) **Architecture coverage**: Our model selection does not include Mixture-of-experts routing, encoder–decoder stacks, and state-space hybrids. The per-head OV decomposition still applies in principle, but the causal head-ablation magnitudes and late-layer concentration we report should not be assumed to transfer without verification.

Acknowledgements

APG was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. BA is a Fellow at and has been supported by the Generative AI Lab (GAIL) at the University of Edinburgh. PM was supported by the Engi-

neering and Physical Sciences Research Council (EPSRC) through the AI Hub in Generative Models (grant number EP/Y028805/1). This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh. We also thank Neel Rajani, Raj Bhalwankar, Matteo Attimonelli, and Federico Tiblias for their helpful comments and suggestions.

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023*, pp. 4895–4901. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.298. URL https://doi.org/10.18653/v1/2023.emnlp-main.298.

Bastings, J. and Filippova, K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–155, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://aclanthology.org/2020.blackboxnlp-1.14/.

Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2025.

Cai, Z., Zhang, Y., Gao, B., Liu, Y., Liu, T., Lu, K., Xiong, W., Dong, Y., Hu, J., and Xiao, W. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2025. URL https://arxiv.org/abs/2406.02069.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Fu, Y., Cai, Z., Asi, A., Xiong, W., Dong, Y., and Xiao, W. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FJFVmeXusW.

Gema, A. P., Jin, C., Abdulaal, A., Diethe, T., Teare, P. A., Alex, B., Minervini, P., and Saseendran, A. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10003–10039, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.531. URL https://aclanthology.org/2025.findings-emnlp.531/.

Jain, S. and Wallace, B. C. Attention is not Explanation. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Kamradt, G. LLMTest_NeedleInAHaystack: Pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023. GitHub repository.

Kuratov, Y., Bulatov, A., Anokhin, P., Rodkin, I., Sorokin, D. I., Sorokin, A., and Burtsev, M. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=u7m2CG84BQ.

Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., and Chen, D. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*, pp. 22947–22970. Curran Associates, Inc., 2024.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Lin, X., Wang, J., Kondrateva, O., Shi, Y., Li, B., and Zhang, G. L. Compresskv: Semantic retrieval heads know what tokens are not important before generation. *arXiv preprint arXiv:2508.02401*, 2025.

Ma, Y. and Okazaki, N. From interpretability to performance: Optimizing retrieval heads for long-context language models. *ArXiv*, abs/2601.11020, 2026. URL https://api.semanticscholar.org/CorpusID:284860903.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

McDougall, C., Conmy, A., Rushing, C., McGrath, T., and Nanda, N. Copy suppression: Comprehensively understanding an attention head, 2023. URL https://arxiv.org/abs/2310.04625.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Modarressi, A., Deilamsalehy, H., Dernoncourt, F., Bui, T., Rossi, R. A., Yoon, S., and Schuetze, H. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0OshX1hiSa.

Nanda, N. and Bloom, J. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

nostalgebraist. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong blog post.

Olmo, T., Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., Morrison, J., Poznanski, J., Lo, K., Soldaini, L., Jordan, M., Chen, M., Noukhovitch, M., Lambert, N., Walsh, P., Dasigi, P., Berry, R., Malik, S., Shah, S., Geng, S., Arora, S., Gupta, S., Anderson, T., Xiao, T., Murray, T., Romero, T., Graf, V., Asai, A., Bhagia, A., Wettig, A., Liu, A., Rangapur, A., Anastasiades, C., Huang, C., Schwenk, D., Trivedi, H., Magnusson, I., Lochner, J., Liu, J., Miranda, L. J. V., Sap, M., Morgan, M., Schmitz, M., Guerquin, M., Wilson, M., Huff, R., Bras, R. L., Xin, R., Shao, R., Skjonsberg, S., Shen, S. Z., Li, S. S., Wilde, T., Pyatkin, V., Merrill, W., Chang, Y., Gu, Y., Zeng, Z., Sabharwal, A., Zettlemoyer, L., Koh, P. W., Farhadi, A., Smith, N. A., and Hajishirzi, H. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Serrano, S. and Smith, N. A. Is attention interpretable? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Sun, Z., Zang, X., Zheng, K., Xu, J., Zhang, X., Yu, W., Song, Y., and Li, H. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ztzZDzgfrh.

Team, G. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025a. URL https://goo.gle/Gemma3Report.

Team, Q. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. ♭ MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *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, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Xiao, G., Tang, J., Zuo, J., Guo, J., Yang, S., Tang, H., Fu, Y., and Han, S. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QHROe7Skvh.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C. W., Wang, Z., and Chen, B. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, pp. 34661–34710. Curran Associates, Inc., 2023.

Appendix

Table of Contents

Reproducibility.

Appendix A - Datasets, Models, and Licenses . . . 14
Appendix B - Experimental Setup Details . . . 14
Appendix C - Compute Resources . . . 15
Appendix D - Code and Data Availability . . . 15
Appendix E - Broader Impacts . . . 15
Appendix F - Declaration of LLM Usage . . . 16

Method walk-through.

Appendix G - Worked Example . . . 17
Appendix H - Architecture Adaptations . . . 17

Additional empirical analyses.

Appendix I - Logit-Contribution Score Distribution . . . 18
Appendix J - Bottom-k Dissociation Score . . . 20
Appendix K - KV-Group $\times$ Layer View . . . 20
Appendix L - Six-Model Versions of Main-Text Ablation Figures . . . 22
Appendix M - Downstream Benchmark Example . . . 22

Theoretical analysis.

Appendix N - Direct-Path Robustness via Tuned Lens . . . 23
Appendix O - Relationship to Attention-Based Scoring . . . 27

Outlook.

Appendix P - Future Work . . . 31

A. Datasets, Models, and Licenses

Tab. 1 lists every external asset used in this work, together with the version we used and its license. All assets are used within the terms of their respective licenses; we do not redistribute model weights or dataset content.

Table 1. **External assets used in this work.** Model checkpoints are accessed via the HuggingFace Hub; dataset and library versions are the ones used in our experiments.

Asset	Identifier / version	Use	License
Models			
Qwen3-8B	`Qwen/Qwen3-8B`	detection, ablation	Apache-2.0
Qwen3-14B	`Qwen/Qwen3-14B`	detection, ablation	Apache-2.0
Qwen3-32B	`Qwen/Qwen3-32B`	detection, ablation	Apache-2.0
Gemma-3-12B	`google/gemma-3-12b-it`	detection, ablation	Gemma Terms of Use[†]
Gemma-3-27B	`google/gemma-3-27b-it`	detection, ablation	Gemma Terms of Use[†]
OLMo-3.1-32B	`allenai/OLMo-3.1-32B-Instruct`	detection, ablation	Apache-2.0
Datasets			
NoLiMa (Modarressi et al., 2025)	v1 (`onehop` reasoning type)	probing benchmark	CC-BY-4.0
NIAH (Kamradt, 2023)	standard configuration	baseline scoring (Wu/NIAH)	MIT
PopQA (Mallen et al., 2023)	v1	parametric specificity control	MIT
City–country associations	authors' construction	parametric specificity control	released with this paper
Arithmetic (two-operand $\pm$)	authors' construction	parametric specificity control	released with this paper
Software			
vLLM	v0.10.0	model loading, weight sharding	Apache-2.0
PyTorch	2.4	numerical backend	BSD-3-Clause
Transformers (HuggingFace)	4.45	tokenizers, configs	Apache-2.0
`rouge-score` (Lin, 2004)	0.1.2	ROUGE-L evaluation	Apache-2.0
TransformerLens (Nanda & Bloom, 2022)	2.7	inspiration for direct logit attribution	MIT

[†]Gemma Terms of Use permit research and commercial use subject to a prohibited-use policy and an attribution requirement.

B. Experimental Setup Details

This appendix consolidates the procedural details that allow the experiments in § 4 to be reproduced. Tab. 2 summarizes the hyperparameters used throughout; the paragraphs below give the rationale and the few cases where defaults were varied.

Table 2. **Hyperparameters used throughout the paper.**

Component	Value	Note
Probing		
NoLiMa reasoning type	`onehop`	following Modarressi et al. (2025)
Context lengths	10 settings, 1k–5k tokens	standard NoLiMa protocol
Insertion depths	10 per length	standard NoLiMa protocol
Characters per entry	3	standard NoLiMa protocol
Trial filter	ROUGE-1 recall > 0.5	following Wu et al. (2025)
Decoding (probing and evaluation)		
Strategy	greedy (temperature $= 0$)	deterministic
Max new tokens	50	sufficient for NoLiMa answers (2–5 tokens)
Aggregation		
Bootstrap resamples (B)	1,000	for 95% CI on $S_{l,h}$
Confidence interval	2.5^{th}–97.5^{th} percentile	non-parametric
Ablation		
Calibration trials	50 (sampled from passing pool)	for mean-activation
Held-out evaluation trials	800	disjoint from probing pool
Ablation depths k	$\{0, 5, 10, 20, 30, 40, 50\}$	for top-k and bottom-k

Calibration for mean-ablation. Mean activations for the ablation intervention are computed from 50 NoLiMa trials sampled uniformly from the pool of ROUGE-passing trials. For each head (l, h), the calibration vector $\bar{\mathbf{q}}^{(l,h)} \in \mathbb{R}^{d_h}$ is a *mean of trial-means*: we first compute the token-average of the post-Q-projection query $\mathbf{q}_t^{(l,h)} = W_Q^{(l,h)} \mathbf{x}_t^{(l)}$ within each calibration trial, then average these per-trial vectors across the 50 trials. Each trial therefore contributes a single equal-weight entry irrespective of its sequence length; a flat mean over all (token, trial) pairs would up-weight longer trials. During ablation, the hook intercepts the post-Q-projection vector *before* the rotary position embedding is applied and replaces it with $\bar{\mathbf{q}}^{(l,h)}$ at every token position; RoPE then rotates $\bar{\mathbf{q}}^{(l,h)}$ position-by-position in the usual way. The value projection $W_V^{(l,h)}$ and the output projection $W_O^{(l,h)}$ are unchanged. Because $\bar{\mathbf{q}}^{(l,h)}$ is a single content-independent vector, the resulting

attention logits depend only on the keys (and on the position-dependent RoPE rotation of $\bar{\mathbf{q}}^{(l,h)}$), so the head's attention distribution becomes content-independent (approximately uniform across positions).

ROUGE-L evaluation. ROUGE-L is computed at the summary level (longest common subsequence between the generated text and gold answer, normalized by gold length) using the `rouge-score` library (Lin, 2004). Generation uses greedy decoding (temperature $= 0$) with a maximum of 50 new tokens.

Bootstrap confidence intervals. For each head's global score $S_{l,h}$, we resample the set of ROUGE-passing trials $B = 1{,}000$ times (with replacement), recompute $S_{l,h}$ for each bootstrap sample, and report the 2.5th and 97.5th percentiles as the 95% confidence interval.

Wu et al. scoring on NoLiMa and NIAH. The Wu et al. retrieval score assigns credit at decode step t if (i) the head's argmax attention position falls within the needle span, and (ii) the token at that position matches the generated token. On NoLiMa, this criterion undercounts retrieval heads because of token-identity mismatch: the question is asked non-literally, so the attended needle token rarely matches the generated answer token. The top head's mean score drops from 0.97 (NIAH) to 0.03 (NoLiMa), a $30\times$ gap driven by the scoring criterion rather than by differences in retrieval behavior.

C. Compute Resources

Hardware. All experiments were run on $2\times$ NVIDIA H100 80GB. Tensor parallelism was used for the larger checkpoints: $\text{TP} = 2$ for the 14B/12B/27B/32B models, $\text{TP} = 1$ for Qwen3-8B.

Per-experiment cost. Detection requires one forward pass per probing trial, generating tokens autoregressively while recording attention weights and the value cache (the primary memory overhead). Ablation evaluation is similarly one forward pass per trial per ablation depth k, plus the 50-trial calibration pass.

D. Code and Data Availability

The detection, ablation, and evaluation code is released at `https://anonymous.4open.science/r/locos-8B85` The repository contains:

- Detection scripts producing the per-head $S_{l,h}$ scores for each model in § 4.1.
- The vLLM-based ablation driver described in Appx. H, including the monkey-patches for each supported attention class.
- Evaluation scripts for NoLiMa ROUGE-L (§ 4.2), the parametric controls (§ 4.6), and the bottom-k analyses (§ 4.4).
- A `README` with exact commands, environment specification (`environment.yml`, pinned package versions matching Tab. 1), and the random seeds used for all sampling steps.

NoLiMa probing inputs are generated from the public NoLiMa release via the included generation script; we do not redistribute the underlying haystack content. The author-released parametric control sets (city–country associations, two-operand arithmetic) are included as JSON files in the repository.

E. Broader Impacts

LOCOS is a diagnostic tool: it identifies attention heads that contribute to non-literal retrieval, without modifying model weights or behavior. Its primary intended uses are interpretability research and downstream applications that benefit from head-level retrieval signal, such as KV cache compression (Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024; Fu et al., 2025; Lin et al., 2025; Xiao et al., 2025) and retrieval-aware decoding (Gema et al., 2025). We do not release a new model, and the method does not enable new generative capabilities.

Potential positive impacts. A more accurate map of which heads carry non-literal retrieval may enable (i) more aggressive KV cache compression at fixed quality, lowering inference cost and energy use; (ii) more targeted interventions for hallucination mitigation in retrieval-augmented generation (Gema et al., 2025); and (iii) cleaner mechanistic accounts of long-context behavior in production-scale models.

Potential negative impacts and mitigations. The method exposes which heads are causally critical for retrieval, which in principle could be used to construct adversarial inputs that suppress retrieval (a denial-of-capability attack). We judge this risk to be low because: (i) the same information is recoverable from any open-weight model with a few hours of compute, so the marginal uplift from publication is small; (ii) the method requires white-box access to attention weights and value

caches, which is not available through standard inference APIs; and (iii) the attack surface (degrading long-context retrieval) is qualitatively similar to risks already present in any KV cache compression method. We therefore do not impose access restrictions on the released code.

Fairness and demographic considerations. Our evaluation uses only the English NoLiMa benchmark and English parametric controls; we do not evaluate retrieval-head identification across languages, dialects, or demographic axes, and we do not claim that the identified heads generalize to non-English text (§ 6). Practitioners deploying LOCOS-derived KV-compression policies on multilingual systems should re-validate on their target languages.

F. Declaration of LLM Usage

We used LLMs as writing assistants for paper polishing and for routine refactoring of plotting scripts. Specifically: (i) prose passes for clarity and concision, particularly on appendix sections, (ii) LaTeX-formatting suggestions, and (iii) minor refactoring of plotting scripts. The LLM was not used to design experiments, derive the theoretical results, or generate numerical results. All experimental claims and numerical values reported in the paper are produced by the analysis pipeline released with the code.

G. Worked Example

The following example uses head $(16, 1)$ from Qwen3-8B on a NoLiMa trial where the answer is "Yuki" and the needle span $[200, 212)$ (12 tokens) is embedded in a context of $N = 3{,}000$ tokens. Numerical values below are illustrative, chosen to show the typical scale of α, φ, Φ^+, and Φ^- for a strongly-retrieving head; they are not drawn from a specific logged trial.

Step 1: Per-position logit contribution (Equation (2)). At decode step $t{=}1$ producing "Yuki" (y_1), consider needle position $j = 205$ containing the token "Kiasma". The head's attention weight is $\alpha_{1,205}^{(16,1)} = 0.08$. After the OV circuit, the per-position output $\mathbf{o}_{1,205}^{(16,1)}$ has a large component aligned with $\mathbf{u}_{\text{Yuki}}$:

$$\phi_{1,205}^{(16,1)} = \mathbf{u}_{\text{Yuki}}^{\top} \mathbf{o}_{1,205}^{(16,1)} = 1.3\,.$$

Despite moderate attention ($\alpha = 0.08$), this position contributes substantially to the "Yuki" logit because the OV circuit transforms the "Kiasma" representation into an output aligned with the answer direction. This is exactly the non-literal retrieval that attention-based scoring cannot detect.

Step 2: Spatial contrast (Equation (3)). With $N_1 = 3{,}000$ key positions at step $t{=}1$:

$$\Phi^+ = \textstyle\sum_{j=200}^{211} \phi_{1,j}^{(16,1)} = 4.7 \quad \text{(total logit contribution from needle)}$$
$$\Phi^- = \tfrac{12}{2988} \textstyle\sum_{j \notin [200,212)} \phi_{1,j}^{(16,1)} = \tfrac{12}{2988} \times 58.3 = 0.23 \quad \text{(length-normalized off-needle)}$$

The 12 needle tokens contribute 4.7 to the "Yuki" logit; a comparable 12-token span from the remaining context contributes only 0.23.

Step 3: Aggregation (Equation (4)). After 200 ROUGE-passing trials contributing a total of 247 answer steps, head $(16, 1)$ has pooled score $S_{16,1} = 3.8$ with 95% bootstrap CI $[3.1, 4.5]$ and per-trial consistency 0.94 (positive in 188/200 trials). A head with $S = 0.2$, CI $[-0.1, 0.5]$, and consistency 0.53 is not reliably retrieval-active.

H. Architecture Adaptations

The logit-contribution equation (Equation (2)) is architecture-invariant: any transformer decomposing into attention weights α, value vectors $\mathbf{v}$, an output projection W_O, and an unembedding matrix W_U provides the four inputs the method requires. The adaptations needed for different architectures are mechanical, not algorithmic.

Grouped-query attention (GQA). In GQA models (Qwen3, Gemma-3, OLMo-3.1), each KV group serves $G = H_Q/H_{KV}$ query heads. The value cache is stored at H_{KV} granularity and expanded to H_Q heads before the per-position projection:

$$\tilde{\mathbf{V}}_h^{(l)} = \mathbf{V}_{\lfloor h \cdot H_{KV}/H_Q \rfloor}^{(l)} \quad \text{for } h = 0, \ldots, H_Q - 1\,.$$

Query heads within the same group share value vectors but differ in $W_O^{(l,h)}$, so their logit-contribution scores can differ.

Vision–language models (Gemma-3). The decoder layers and unembedding matrix reside at a nested path (`language_model.model.layers` and `language_model.lm_head`). Per-head QK normalization affects attention distributions but does not change the value vectors or output projections. Detection operates on text-only self-attention; extending to multimodal inputs is left for future work.

Implementing head ablation in vLLM. Our ablation intervention requires a masked pass that zeros the queries of a designated head subset while sharing the same prefix history. vLLM's default generation path is built around paged attention and a continuous-batch scheduler, neither of which exposes the per-head query manipulation the method needs. We therefore use vLLM specifically for model loading, tensor-parallel weight sharding, and tokenization, and replace its generation loop with a manual autoregressive driver. At load time, we monkey-patch each supported attention class (`Gemma3Attention`, `Qwen3Attention`, `Olmo2Attention`) so that, when the driver is active, the layer routes through `F.scaled_dot_product_attention` backed by two independent sequential KV caches—one per pass. Each decode step runs the forward with retrieval-head queries zeroed. For tensor-parallel execution (TP > 1), the patching and KV bookkeeping run inside every worker, and the two passes are dispatched via vLLM's `collective_rpc`; global head indices are remapped to local shard indices per rank. The trade-off is an $O(n^2)$ KV cache and single-request decoding rather than continuous batching; the benefit is that any model loadable by vLLM under one of the supported attention classes is supported without modifying vLLM internals.

I. Logit-Contribution Score Distribution

Fig. 10 shows the distribution of LOCOS scores $S_{l,h}$ across all attention heads for each of the six models, with the top-50 and bottom-50 heads highlighted. In every model, the score distribution is heavily right-skewed: a small number of heads have large positive scores, the vast majority cluster near zero, and the bottom-50 heads all have strictly negative scores. This confirms that the bottom-k ablation experiments in § 4.4—which ablate up to k=50 heads ranked from most negative—exclusively target heads with $S_{l,h} < 0$, *i.e.,* heads whose logit contribution toward the correct answer originates predominantly from off-needle positions. The clear separation between positive and negative tails across all six models supports the validity of the spatial contrast as a discriminative criterion.

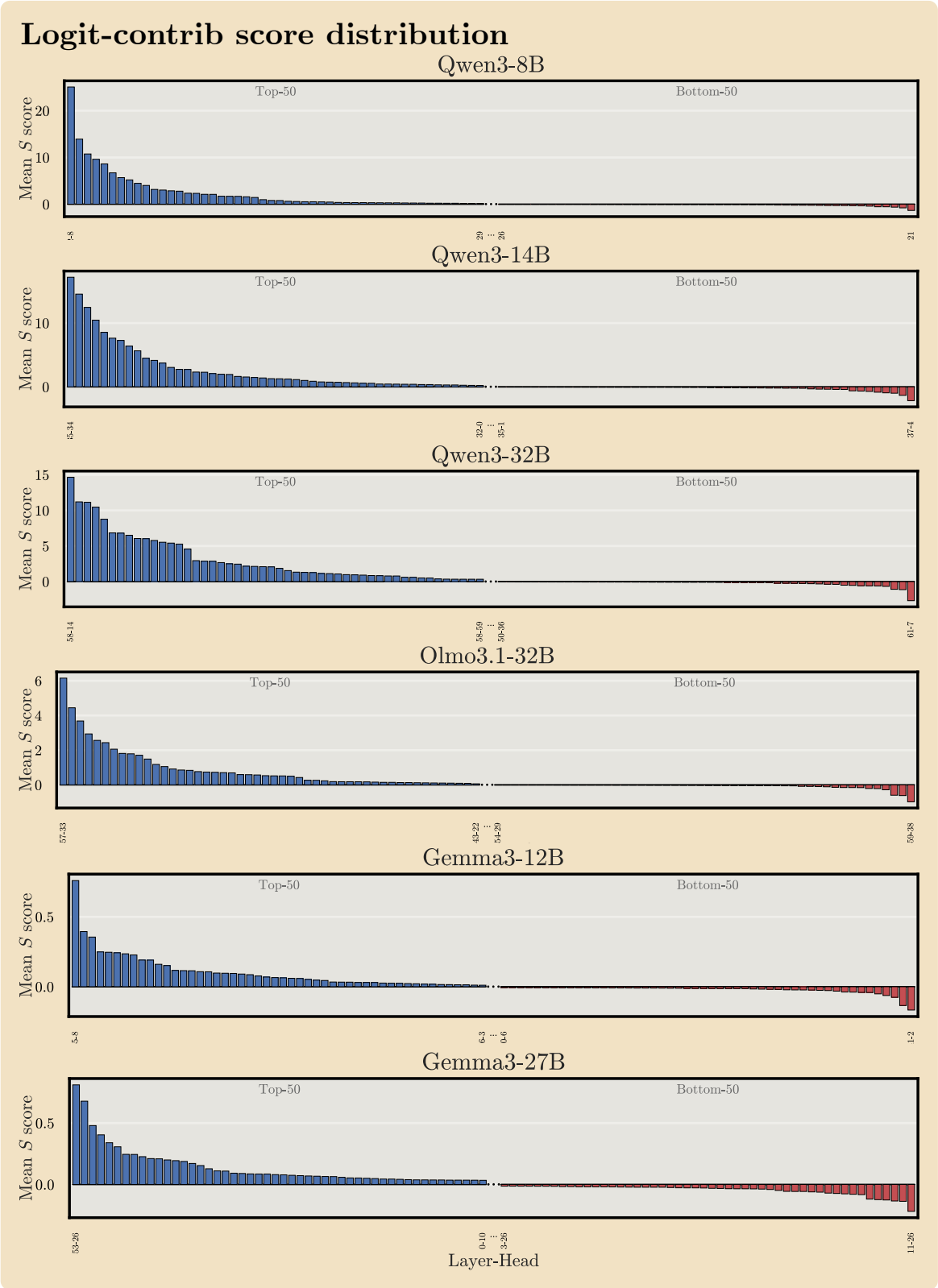

Figure 10. **Distribution of LOCOS scores across all heads for each model.** Heads are sorted by $S_{l,h}$; the top-50 (blue, left) and bottom-50 (red, right) are highlighted. In every model, the bottom-50 heads have strictly negative scores, confirming that the bottom-k experiments (§ 4.4) exclusively ablate heads whose answer-aligned logit contribution originates from off-needle positions.

J. Bottom-k Dissociation Score

Fig. 11 shows the dissociation score under bottom-k ablation alongside parametric accuracy. While top-k ablation produces high dissociation scores—retrieval degrades steeply with parametric accuracy intact—bottom-k ablation yields near-zero dissociation at all depths, confirming that neither retrieval nor parametric performance is affected.

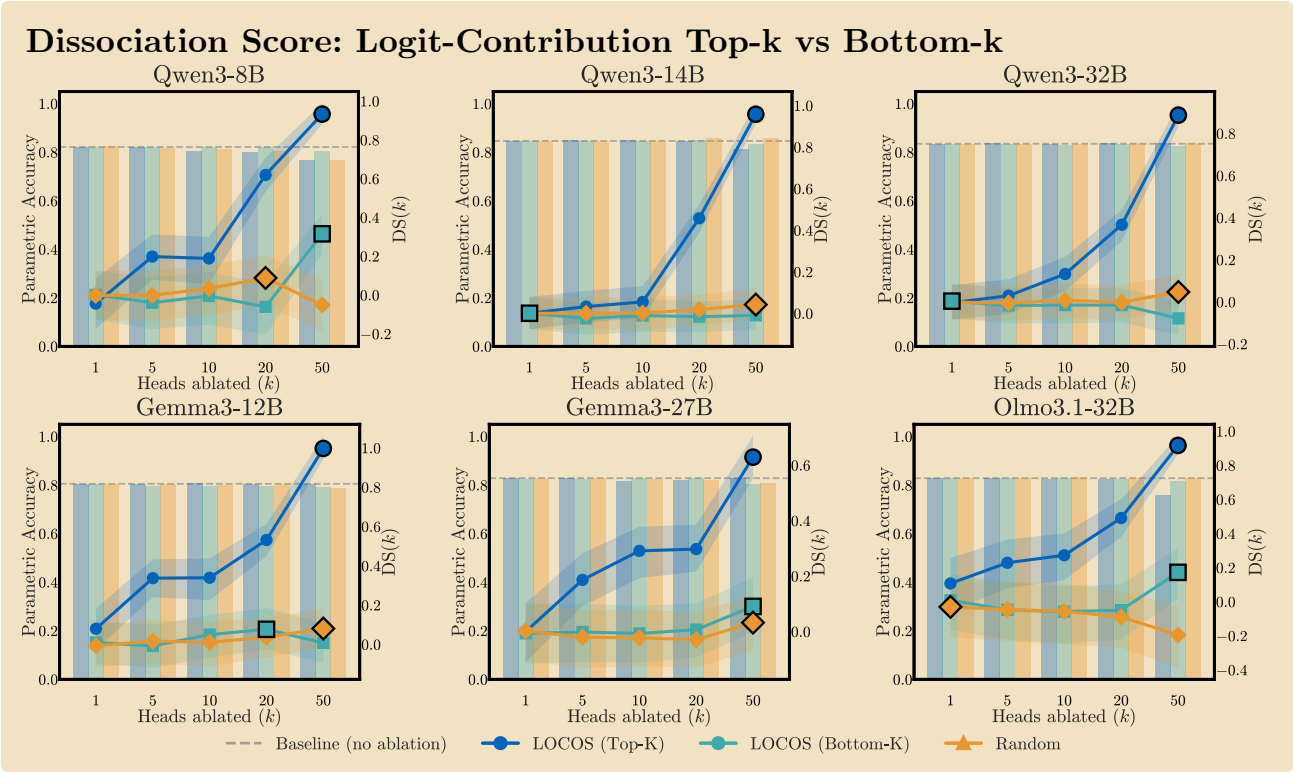

Figure 11. **Bottom-k ablation produces near-zero dissociation.** Dissociation score DS(k) and parametric accuracy as a function of ablation depth k for bottom-k heads across six models. Unlike top-k ablation (Fig. 7), bottom-k ablation leaves both retrieval and parametric performance near baseline.

K. KV-Group $\times$ Layer View of Logit-Contribution Scores

In GQA models (Ainslie et al., 2023), query heads sharing a KV group consume identical key and value vectors, so their per-head LOCOS scores are correlated. The cache unit in GQA is the KV group, not the individual query head, so KV cache footprint is also reasoned about at (layer, KV-group) granularity. This appendix asks one spatial question that the head-level analyses in §§ 4.2 and 4.4 do not: at which layers do the highest-scoring KV groups appear?

Observations. Tab. 3 summarizes the top-10 (layer, KV-group) cells across all six GQA models. Two patterns emerge. First, layer concentration is family-dependent. The Qwen3 family places every top-10 cell in the upper quartile of layers across all three scales (29–34 of 36, 30–38 of 40, and 54–62 of 64), and Gemma-3-27B is concentrated in the upper third (41–53 of 62). Gemma-3-12B and OLMo-3.1-32B do not show comparable upper-layer concentration: their top-10 cells span 0–47 of 48 and 27–57 of 64 respectively. Second, KV-group spread is roughly at chance. The number of distinct KV groups containing top-10 cells (6–8) is at or just above the uniform-allocation expectation (≈ 5.9 for 8-group models, ≈ 7.6 for 16-group models), so the KV-group dimension does not concentrate under LOCOS in any of the six models. The evidence is six models from three families on one detector and one benchmark; the family-dependent layer pattern in particular warrants replication on additional benchmarks before generalization.

The layer pattern is detector- and metric-conditional. The late-layer concentration in Fig. 12 is not evidence that retrieval circuitry lives only in late layers. Two alternative explanations remain unaddressed by the figure. First, the per-position score (Equation (2)) is a direct-path projection onto the answer-token unembedding. Late-layer heads sit closer to the unembedding with fewer downstream non-linearities to redirect their output, so their direct-path projections are systematically larger as a

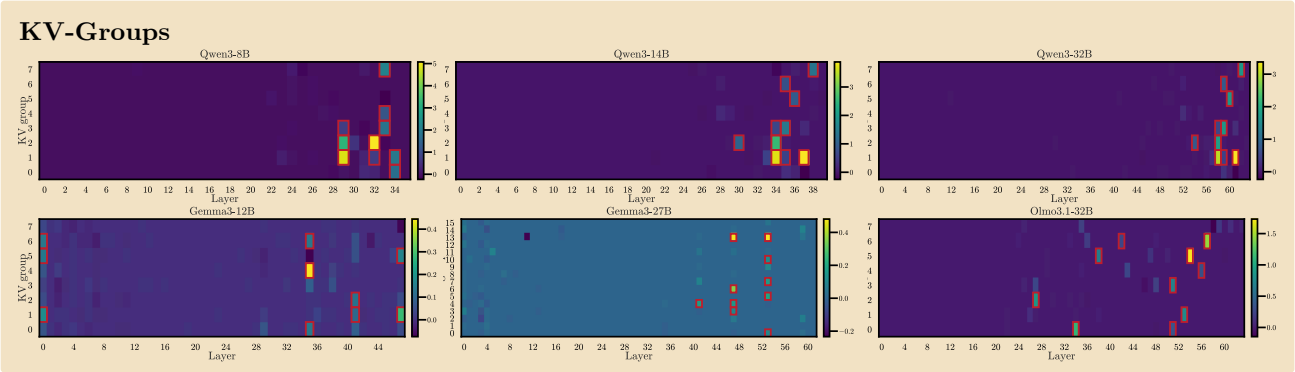

Figure 12. **Top-10 LOCOS cells concentrate in late layers in the Qwen3 family on NoLiMa, but span broader layer ranges in Gemma-3-12B and OLMo-3.1-32B.** Per-(layer, KV-group) mean LOCOS score on NoLiMa for Qwen3-8B, Qwen3-14B, Qwen3-32B, OLMo-3.1-32B, Gemma-3-12B, and Gemma-3-27B. Layer is on the x-axis, KV group on the y-axis, color encodes the mean score across passing trials. Red boxes mark the top-10 (layer, KV-group) cells per model. Per-panel color scales differ; the layer pattern is read from the spatial location of red boxes, not from absolute magnitude. Per-model layer spans and KV-group counts are reported in Tab. 3.

Table 3. **Spatial summary of the top-10 (layer, KV-group) LOCOS cells on NoLiMa.** Listed at KV-group granularity, since query heads in a group share keys and values. The rightmost column reports the number of distinct KV groups containing at least one top-10 cell; under uniform allocation of 10 cells the expected count is ≈ 5.9 for 8-group models and ≈ 7.6 for 16-group models.

Model	Layers $\times$ KV-groups	Top-10 layer span	Top-10 KV groups	Distinct KV groups in top-10
Qwen3-8B	36×8	layers 29–34	$\{0, 1, 2, 3, 4, 7\}$	$6/8$
Qwen3-14B	40×8	layers 30–38	$\{1, 2, 3, 5, 6, 7\}$	$6/8$
Qwen3-32B	64×8	layers 54–62	$\{1, 2, 3, 5, 6, 7\}$	$6/8$
OLMo-3.1-32B	64×8	layers 27–57	$\{0, 1, 2, 3, 4, 5, 6\}$	$7/8$
Gemma-3-12B	48×8	layers 0–47	$\{0, 1, 2, 4, 5, 6\}$	$6/8$
Gemma-3-27B	62×16	layers 41–53	$\{0, 3, 4, 5, 6, 7, 10, 13\}$	$8/16$

property of the metric, irrespective of whether they implement retrieval. The tuned-lens variant in Fig. 16 addresses this concern at head granularity for Gemma-3-27B; we have not re-run it at KV-group granularity for the two Qwen3 models. Second, the same KV-group $\times$ layer view computed from Wu (NIAH) attention scores places top cells across a wider layer range, so a "where retrieval lives" claim made from Fig. 12 alone would not survive a different detector. Fig. 12 is therefore descriptive of LOCOS scores at KV-group granularity, not a localization of retrieval circuitry. Whether these cells implement retrieval rather than some other late-layer behavior is settled by the head-level NoLiMa ablation in § 4.2 and the specificity controls in § 4.6, not by this figure.

Implication for KV cache compression. The unit of KV cache memory in GQA is the (layer, KV-group) cell, so concentration at this granularity is what saves cache memory; head-level concentration does not, because query heads in a group share keys and values. A compression policy could keep full-resolution K/V for the small set of top-scoring cells (10 of 288 in Qwen3-8B and 10 of 320 in Qwen3-14B) and apply heavier quantisation, eviction, or window truncation to the rest. The lower three quarters of layers contain no top-10 cell in either model. This gives a selection criterion at finer granularity than the per-layer eviction policies of Zhang et al. (2023), Li et al. (2024), and Xiao et al. (2024); whether it produces a better quality–budget trade-off is the experiment we do not run. Xiao et al. (2025) make the closely related argument that a small set of retrieval heads should retain full KV cache while the rest use streaming attention; the KV-group view sharpens this argument for GQA models, where retention saves cache only at the group level.

Caveat for compression. The late-layer pattern is detector- and metric-conditional, so a policy derived from LOCOS alone risks over-compressing mid-layer KV that other detectors flag as relevant. A practical heuristic should pool signal from multiple retrieval-head detectors rather than rely on a single score. We do not run a KV-compression experiment in this paper, and promoting the implication above to a claim would require, at minimum, top-k versus random KV-group eviction on long-context perplexity, NIAH, and NoLiMa, with a held-out suite (*e.g.,* parametric recall, arithmetic) to check that the policy leaves non-retrieval capabilities intact.

L. Six-Model Versions of Main-Text Ablation Figures

The ablation figures in §§ 4.4, 4.6 and 4.7 show three representative models (one per family). This appendix reports the full six-model versions. The conclusions stated in the main text are based on all six configurations.

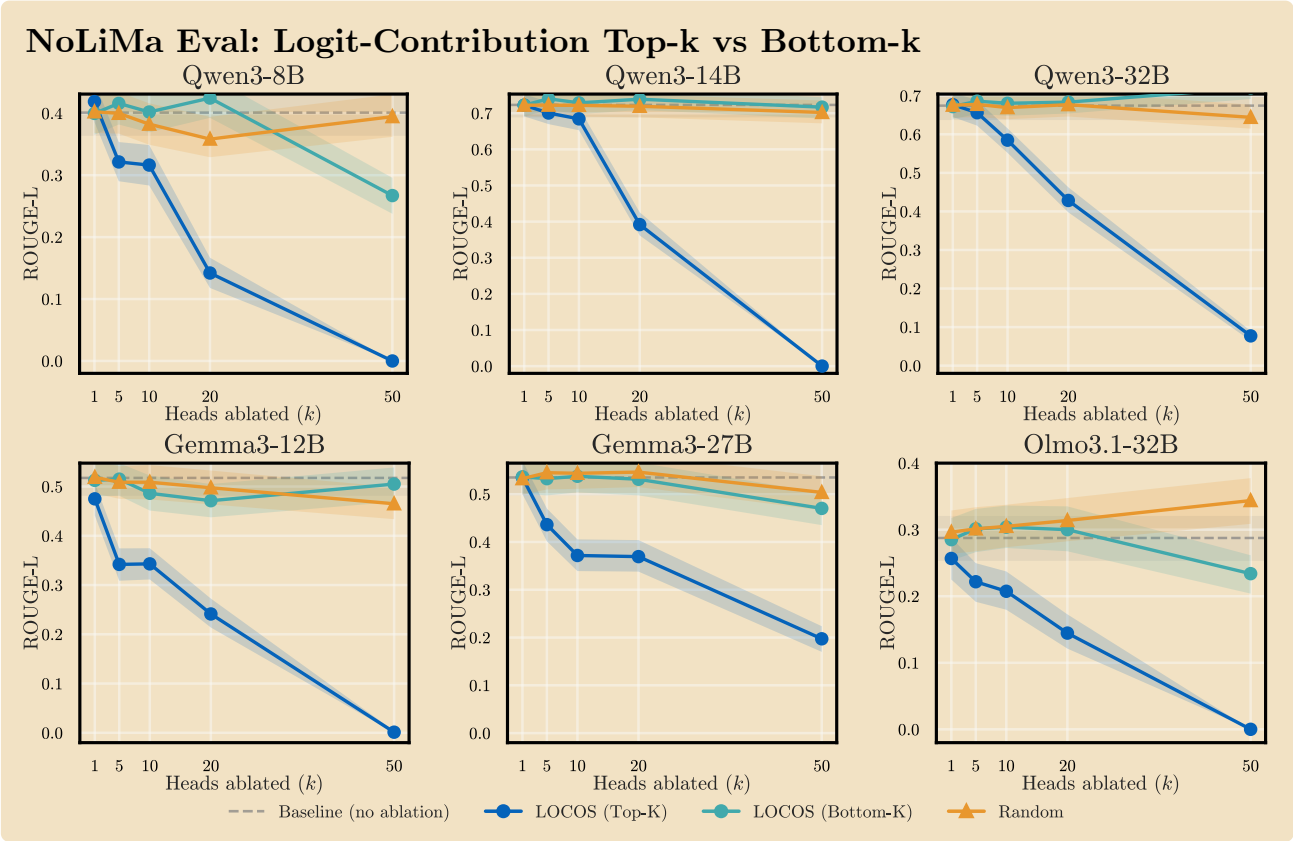

Figure 13. **Bottom-k ablation does not degrade retrieval (six-model version of Fig. 5).** Each panel shows NoLiMa ROUGE-L as a function of ablation depth k for top-k (blue), bottom-k (purple), and random heads (red) across all six models.

M. Downstream Benchmark Example

BABILong qa2 (truncated):
Context: Mary journeyed to the bathroom. Sandra went to the garden. [. . .] Daniel journeyed to the bedroom. [. . .] Daniel took the football there. [. . .] Daniel dropped the football. [. . .] Daniel grabbed the football there. [. . .] Daniel went to the kitchen.
Question: Where is the football?
Answer: kitchen

The example above illustrates why literal token copying is insufficient for BABILong: the correct location ("kitchen") is the *most recent* of many positions where the entity and object appear, requiring the model to trace a trajectory across interleaved narrative rather than match a unique token.

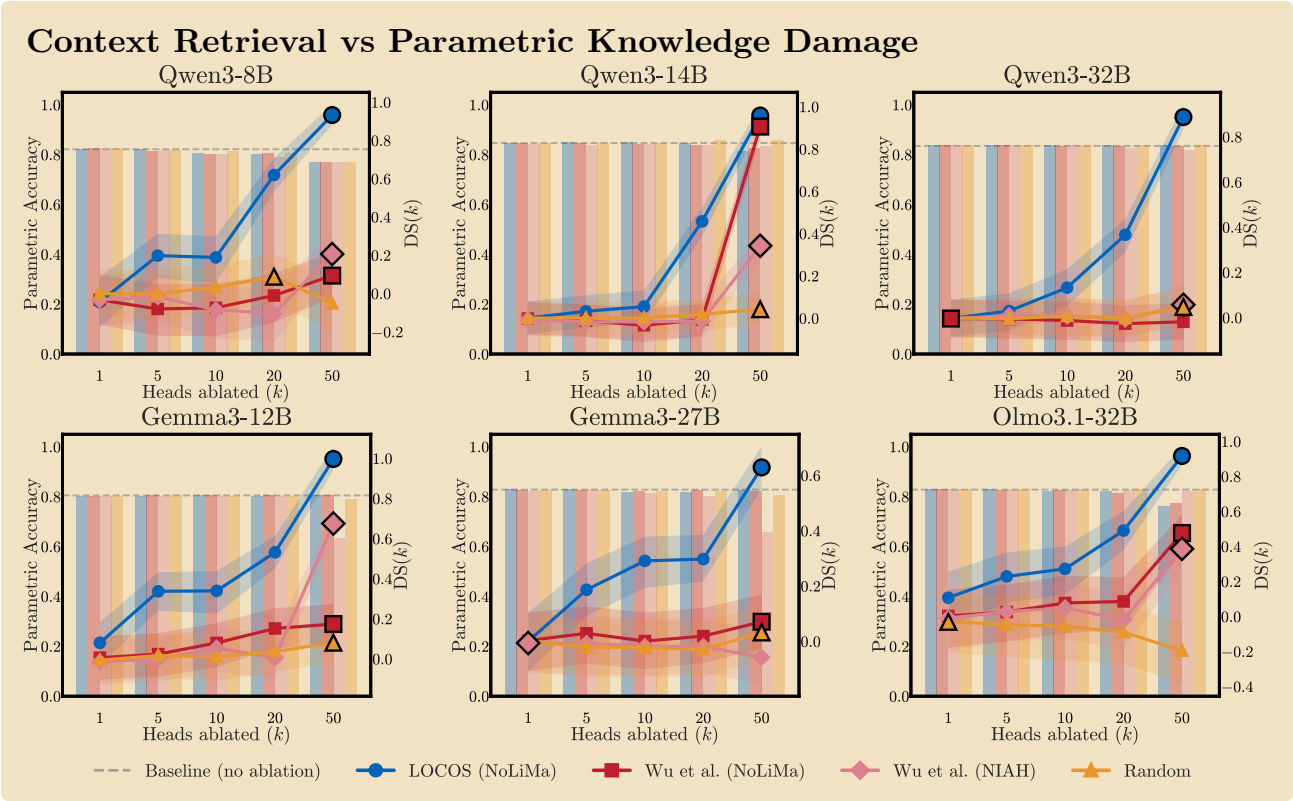

Figure 14. **Functional dissociation between retrieval and parametric capabilities across all six models (six-model version of Fig. 7).** Each panel shows DS(k) (lines, right axis) and parametric accuracy (bars, left axis) as a function of ablation depth k for four scoring methods. LOCOS (blue) achieves the highest DS in every model configuration; the enlarged marker indicates k^*.

N. Direct-Path Robustness via Tuned Lens

This appendix derives the direct-path identity that motivates the per-position score $\phi_{t,j}^{(l,h)}$ in Equation (2), characterizes what the projection misses, and shows how a tuned-lens variant addresses the bias that affects early-layer heads.

N.1. The direct-path identity

Let $\mathbf{x}_t^{(l)} \in \mathbb{R}^d$ denote the residual stream at layer l and decode step t. A standard transformer block updates the residual stream additively,

$$\mathbf{x}_t^{(l)} \;=\; \mathbf{x}_t^{(l-1)} \;+\; \sum_{h=1}^{H} \mathbf{a}_t^{(l,h)} \;+\; \mathbf{m}_t^{(l)}, \tag{5}$$

where $\mathbf{a}_t^{(l,h)} = \sum_j \mathbf{o}_{t,j}^{(l,h)}$ is the head-(l,h) output (Equation (1)) and $\mathbf{m}_t^{(l)}$ is the MLP output of layer l. We write $\sigma_l(\cdot)$ for the layer-norm rescaling that precedes the unembedding (*e.g.,* the final RMSNorm). The next-token logits at step t are

$$\boldsymbol{\ell}_t \;=\; W_U\, \sigma_L\big(\mathbf{x}_t^{(L)}\big) \;\in\; \mathbb{R}^{|\mathcal{V}|}. \tag{6}$$

Unrolling Equation (5) from layer 1 to L and substituting into Equation (6),

$$\boldsymbol{\ell}_t \;=\; W_U\, \sigma_L\Big(\mathbf{x}_t^{(0)} + \sum_{l=1}^{L}\Big[\sum_{h=1}^{H}\sum_{j=1}^{N_t} \mathbf{o}_{t,j}^{(l,h)} + \mathbf{m}_t^{(l)}\Big]\Big). \tag{7}$$

Equation (7) is exact. Because σ_L is non-linear in general, the contribution of any single $\mathbf{o}_{t,j}^{(l,h)}$ to $\boldsymbol{\ell}_t$ is not additively separable.

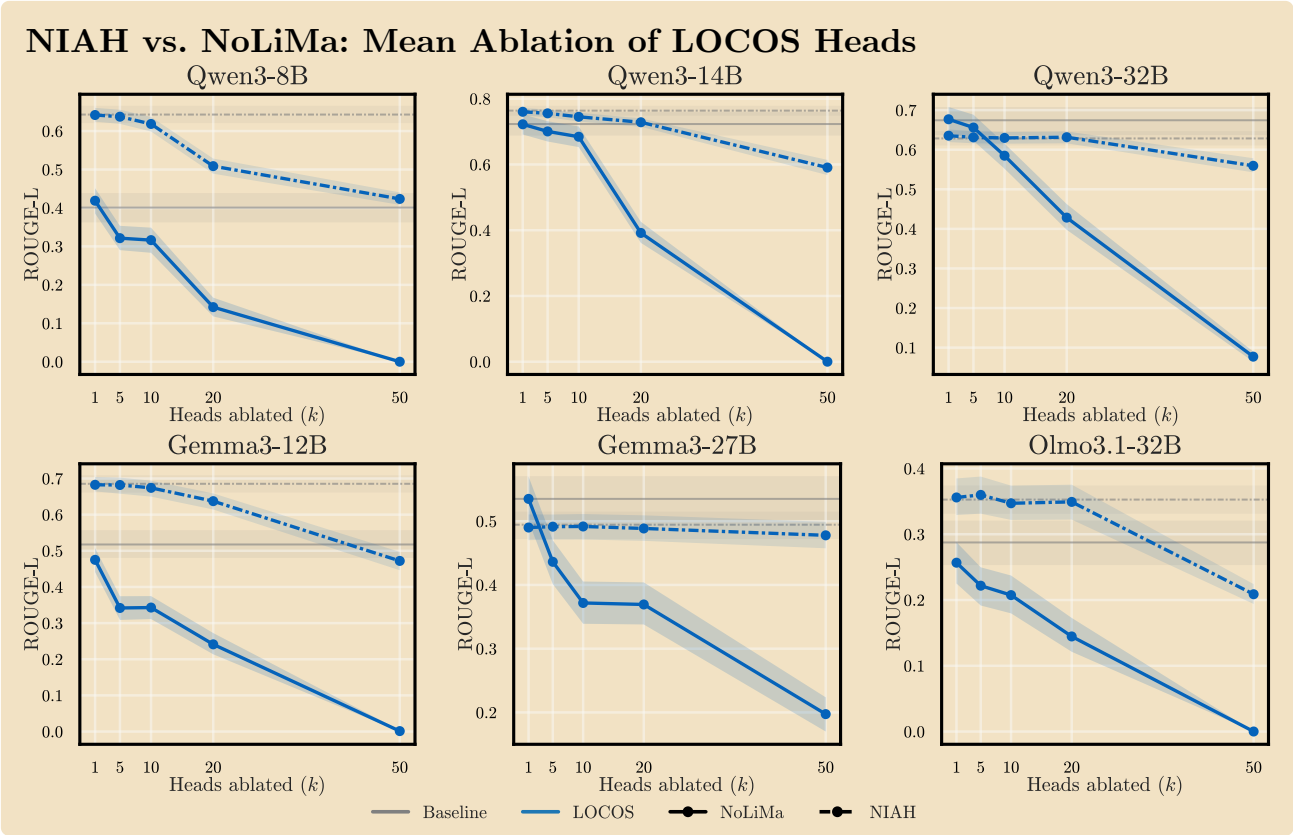

Figure 15. **Non-literal vs. literal retrieval damage across all six models (six-model version of Fig. 8).** Each panel shows ROUGE-L on NoLiMa (solid blue) and standard NIAH (dashed blue) under mean-ablation of the same top-k LOCOS heads, with the NoLiMa and NIAH baselines marked by solid and dashed gray lines. The NoLiMa curve declines more steeply than the NIAH curve in every configuration.

The direct-path approximation. Linearizing σ_L at $\mathbf{x}_t^{(L)}$ gives $\sigma_L(\mathbf{x}_t^{(L)} + \mathbf{o}_{t,j}^{(l,h)}) - \sigma_L(\mathbf{x}_t^{(L)}) \approx J_{\sigma_L}(\mathbf{x}_t^{(L)})\,\mathbf{o}_{t,j}^{(l,h)}$, so the contribution of $\mathbf{o}_{t,j}^{(l,h)}$ to the logit of token y is approximately $\mathbf{u}_y^\top J_{\sigma_L}(\mathbf{x}_t^{(L)})\mathbf{o}_{t,j}^{(l,h)}$. The *direct-path score* replaces J_{σ_L} with the identity, yielding

$$\phi_{t,j}^{(l,h)} \stackrel{\text{def}}{=} \mathbf{u}_{y_t}^\top \mathbf{o}_{t,j}^{(l,h)}. \tag{8}$$

This is the score used throughout the main text (Equation (2)). The omitted Jacobian acts close to a per-step rescaling for the RMSNorm models we evaluate (Qwen3, Gemma-3, OLMo-3 are all RMSNorm); we treat the substitution as an approximation, not an identity, and validate it empirically via the tuned-lens variant in Appxs. N.3 and N.4.

N.2. What the direct path misses

The error in the direct-path score (Equation (8)) relative to the exact contribution decomposes into three sources, in increasing order of severity for early-layer heads:

1. **LayerNorm rescaling.** The final σ_L rescales each direction by $\gamma/\rho(\mathbf{x}_t^{(L)})$. This is approximately constant across heads at a given step but varies across steps; it does not change rankings within a step but inflates magnitudes.
2. **Downstream attention re-mixing.** A head (l, h) at layer $l < L$ writes into the residual stream that is read by all heads at layers $l+1, \ldots, L$. If a downstream head (l', h') with $l' > l$ amplifies or cancels the answer-aligned component of $\mathbf{o}_{t,j}^{(l,h)}$, the direct-path score under-counts (or, in the cancellation case, can sign-flip relative to) the true causal contribution.
3. **MLP composition.** MLP sublayers between layer l and L apply non-linear transformations (gated activations, GELU/SwiGLU). A retrieval head whose output only becomes answer-aligned after a downstream MLP read–write pair is invisible to any linear probe of $\mathbf{x}_t^{(l)}$.

Sources (1) and (2) are linear in $\mathbf{o}_{t,j}^{(l,h)}$ and are therefore correctable in principle by a learned linear lens. Source (3) is non-linear and is fundamentally outside the reach of any linear probe.

N.3. Tuned-lens formalism

The tuned lens of Belrose et al. (2025) learns, for each layer l, an affine map $T_l : \mathbb{R}^d \to \mathbb{R}^d$ such that $W_U\, T_l(\mathbf{x}_t^{(l)})$ approximates the model's true next-token logits $\boldsymbol{\ell}_t$. Concretely, $T_l(\mathbf{x}) = A_l\mathbf{x} + b_l$, with (A_l, b_l) trained by minimizing the KL divergence between $\mathrm{softmax}(W_U T_l(\mathbf{x}_t^{(l)}))$ and the true output distribution on a held-out corpus. The tuned lens absorbs sources (1) and (2) of Appx. N.2 into the learned linear map; source (3) remains uncaptured.

Lens-corrected score. Substituting the tuned lens for the identity in Equation (8) yields a layer-aware score

$$\phi_{t,j}^{(l,h),\mathrm{TL}} \;\stackrel{\mathrm{def}}{=}\; \mathbf{u}_{y_t}^{\top} A_l\, \mathbf{o}_{t,j}^{(l,h)} \;+\; \text{constant in } j, \tag{9}$$

where the bias b_l contributes a j-independent term that cancels under the spatial contrast (Equation (3)). Writing $\tilde{\mathbf{p}}^{(l,h)} = (W_O^{(l,h)})^{\top} A_l^{\top}\mathbf{u}_{y_t}$, the lens-corrected per-position score takes the same factored form as the direct-path score:

$$\phi_{t,j}^{(l,h),\mathrm{TL}} \;=\; \alpha_{t,j}^{(l,h)} \cdot \big(\mathbf{v}_{t,j}^{(l,h)}\big)^{\top} \tilde{\mathbf{p}}^{(l,h)} \;+\; \text{const.},$$

so the same precomputation-and-batched-inner-product implementation applies; only the projection vector $\tilde{\mathbf{p}}$ changes.

What the lens does not fix. Equation (9) replaces the direct-path Jacobian by a layer-aware linear approximation, but it cannot capture any non-linear composition (source (3)). A head whose write only becomes answer-aligned after passing through a downstream MLP scores low under both the direct-path and tuned-lens variants. Distinguishing a genuinely causally inert head from one whose contribution is hidden by downstream non-linearity requires a non-linear probe, of which the canonical example is causal activation patching (Wang et al., 2023; Meng et al., 2022).

N.4. Empirical comparison

Fig. 16 compares layer×head score heatmaps for Gemma-3-27B under Equation (8) and Equation (9). The agreement on the late-layer band ($l \in [35, 60]$) suggests that the late-layer concentration reflects retrieval structure rather than a systematic underestimation of early-layer heads by the direct-path projection. Two layer-11 heads (26 and 27) surface only under the tuned lens; we treat these as candidate early-layer retrieval heads that the direct-path score under-ranks but the linear correction recovers.

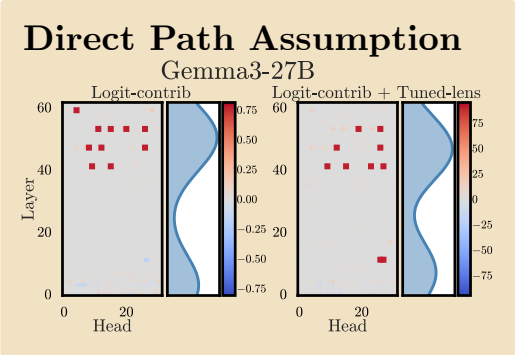

Figure 16. **Late-layer concentration persists under tuned-lens projection.** Heatmaps for Gemma-3-27B: direct-path LOCOS (left) vs. tuned-lens variant (right). Both methods concentrate high-scoring heads in layers 35–60; the layer-marginal distributions peak in the same band. The tuned-lens variant surfaces two additional heads at layer 11 (heads 26 and 27) that do not appear in the direct-path top-k set, but does not produce a broader redistribution toward earlier layers. The score magnitudes differ (± 75 vs. ± 0.75) because the learned affine map amplifies contributions by accounting for downstream transformations.

N.5. Gemma-3-27B inversion

Replacing the direct-path projection with the tuned-lens readout largely closes the gap with the attention-only control at large k on Gemma-3-27B (Fig. 17). At small k, however, direct LOCOS selects more damaging heads than either α-spatial scoring or the tuned-lens variant. Thus, the Gemma-3-27B anomaly is not cleanly explained as a direct-path readout artifact: correcting the readout helps only at the deepest sweep and worsens the top-ranked heads. This suggests that, for this model, spatial attention contrast captures a causal signal that the evaluated linear write-projections do not rank consistently (Appx. N.2). For Gemma-3-27B, attention placement under spatial contrast carries more causal signal than any linear write-projection we evaluate.

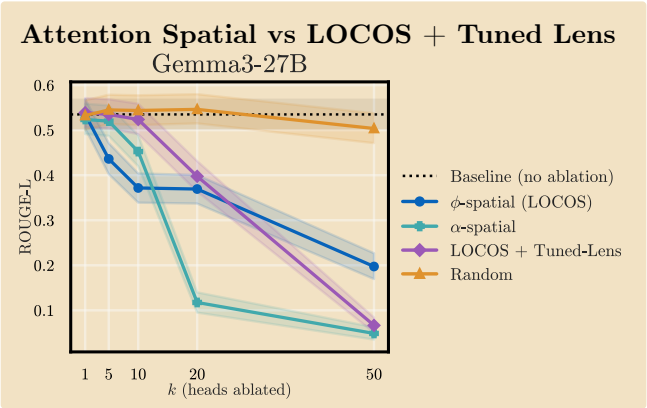

Figure 17. **Tuned-lens correction only partly resolves the Gemma-3-27B inversion.** NoLiMa ROUGE-L under mean-ablation of top-k heads ranked by direct LOCOS, the attention-only spatial-contrast control, and the tuned-lens-corrected LOCOS variant on Gemma-3-27B. The tuned-lens variant closes much of the gap with attention-only scoring at large k, but direct LOCOS selects the most damaging heads at small k.

N.6. Beyond linear probes

A linear probe (direct path or tuned lens) cannot detect a head whose contribution to y_t materializes only through MLP composition. Per-head causal activation patching is the canonical gradient-free, non-linear-aware test: for each head (l, h), run a clean forward pass on the needle prompt and a corrupt forward pass with the needle removed or scrambled, then patch only the (l, h) activation from the clean run into the corrupt run and measure the change in logit difference at the answer position,

$$\mathrm{CA}(l,h) \;=\; \big[\ell^{\mathrm{patch}}_{y^*} - \ell^{\mathrm{patch}}_{y^{\mathrm{cf}}}\big] - \big[\ell^{\mathrm{corr}}_{y^*} - \ell^{\mathrm{corr}}_{y^{\mathrm{cf}}}\big], \tag{10}$$

where y^* is the gold first answer token and $y^{\mathrm{cf}} = \arg\max_{y \neq y^*} \ell^{\mathrm{corr}}_y$ is the strongest non-gold competitor under the corrupt baseline. We refer to this score as *causal attribution* (CA).

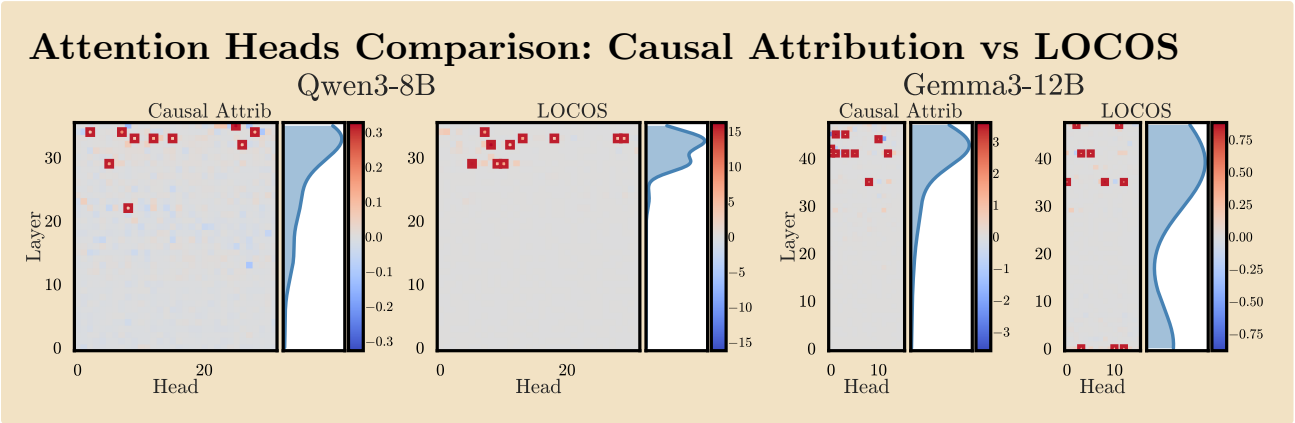

Figure 18. **Causal attribution vs. LOCOS top-10 heads on Qwen3-8B and Gemma-3-12B.** Per-(layer, head) score heatmaps with red boxes marking each method's top-10 cells; layer-marginal kernel densities on the right of each panel. Both methods concentrate top-10 heads in the upper layers in both models, but the top-10 sets overlap only marginally (2/10 for Qwen3-8B, 3/10 for Gemma-3-12B). On Gemma-3-12B, LOCOS surfaces several layer-0 heads that causal attribution does not flag.

Empirical comparison on two models. Fig. 18 reports causal attribution and LOCOS side by side for Qwen3-8B and Gemma-3-12B. The shared finding is that the late-layer concentration is not a direct-path artifact: causal attribution—which is non-linear-aware and gradient-free—also places its top-10 heads predominantly in the upper layers in both models (layers 22–35 for Qwen3-8B, concentrated in 33–35; layers 35–45 for Gemma-3-12B, concentrated near 41). LOCOS's top-10 sets sit in similar bands (29–34 for Qwen3-8B; layers spanning 0, 35, 41, 47 for Gemma-3-12B). Top-10 overlap is small (2/10 and 3/10), which is consistent with the two scores capturing different aspects of retrieval-related computation rather than one being a strict refinement of the other.

Where the methods diverge, and why. The notable disagreement on Gemma-3-12B is that LOCOS surfaces heads at layer 0 that causal attribution does not. Two explanations are consistent with the figure and we cannot distinguish them here. First, layer-0 heads may be a direct-path artifact: their write enters the residual stream before any downstream MLP can either amplify or cancel the answer-aligned component, so the linear projection onto $\mathbf{u}_{y_t}$ records a magnitude that the model itself does not realize at the output (sources (2)–(3) of Appx. N.2). Second, layer-0 heads may participate in distributed circuits where ablating any single head leaves performance intact (because parallel paths compensate), so causal attribution—a single-head intervention—under-counts them while LOCOS correctly registers each head's contribution to the linear readout. The bottom-k control (§ 4.4) and the top-k ablation curves (Fig. 7) bound the issue at the level of method validation: LOCOS-selected top heads, including any artifactual ones, are causally critical *collectively* under group ablation even if individual heads are not under single-head patching.

Scope. The comparison is two models from two families on one benchmark with one alternative detector. Causal attribution is also expensive: it requires one additional forward pass per head per trial (a clean–corrupt patching pair scaled by the number of heads, $H_Q \cdot L$ per trial), so the cost grows linearly with the head count and is roughly two orders of magnitude above a single LOCOS pass for the models we evaluate. We therefore do not run it on the other four models or on NIAH/parametric controls; running it across the full evaluation matrix is left to future work. The takeaway we draw is narrow: the late-layer concentration of LOCOS top heads is corroborated by a non-linear-aware detector on the two models for which we have both scores, and the residual disagreement (notably the Gemma-3-12B layer-0 heads) is a candidate site for future single-head causal validation rather than a refutation of the present paper's group-ablation results.

O. Relationship to Attention-Based Scoring

This appendix gives a self-contained derivation of how LOCOS relates to the attention-pattern observable used by prior detection methods (Wu et al., 2025; Fu et al., 2025; Lin et al., 2025). Appx. O.1 fixes notation; Appx. O.2 decomposes the per-position OV output into answer-aligned and answer-orthogonal components; Appx. O.3 states and proves the reduction of LOCOS to attention-based scoring under a position-independent OV output; Appxs. O.4 and O.5 cast the Wu and HeadKV/CompressKV scores as special cases under additional assumptions; Appx. O.6 contrasts spatial and temporal aggregation; Appx. O.7 gives the gradient interpretation; and Appx. O.8 consolidates the results.

O.1. Setup

We work with the standard pre-softmax decomposition of a transformer attention head (Elhage et al., 2021). At decode step t, head (l, h) produces query $\mathbf{q}_t^{(l,h)}$ and, for each source position $j \in \{1, \ldots, N_t\}$, key $\mathbf{k}_{t,j}^{(l,h)}$ and value $\mathbf{v}_{t,j}^{(l,h)} \in \mathbb{R}^{d_h}$. The QK circuit produces

$$\alpha_{t,j}^{(l,h)} \;=\; \frac{\exp\big(\mathbf{q}_t^{(l,h)\top}\mathbf{k}_{t,j}^{(l,h)}/\sqrt{d_h}\big)}{\sum_{j'=1}^{N_t}\exp\big(\mathbf{q}_t^{(l,h)\top}\mathbf{k}_{t,j'}^{(l,h)}/\sqrt{d_h}\big)}, \tag{11}$$

and the OV circuit writes $\mathbf{o}_{t,j}^{(l,h)} = \alpha_{t,j}^{(l,h)} \cdot W_O^{(l,h)}\mathbf{v}_{t,j}^{(l,h)}$ to the residual stream (Equation (1)). The full head output is $\mathbf{a}_t^{(l,h)} = \sum_{j=1}^{N_t}\mathbf{o}_{t,j}^{(l,h)}$.

For brevity, we drop the (l, h) superscripts in this appendix when no ambiguity arises and write α_j, $\mathbf{v}_j$, $\mathbf{w}_j \stackrel{\text{def}}{=} W_O\mathbf{v}_j$, and $\mathbf{u} \stackrel{\text{def}}{=} \mathbf{u}_{y_t}$. With this notation the per-position score Equation (2) is $\phi_j = \alpha_j \cdot \mathbf{u}^\top\mathbf{w}_j$.

O.2. Parallel–orthogonal decomposition of the OV output

Decompose each unweighted OV output $\mathbf{w}_j \in \mathbb{R}^d$ along $\mathbf{u}$ and orthogonal to it:

$$\mathbf{w}_j \;=\; c_j \cdot \mathbf{u} \;+\; \mathbf{w}_j^\perp, \qquad c_j \;=\; \frac{\mathbf{u}^\top\mathbf{w}_j}{\|\mathbf{u}\|^2}, \qquad \mathbf{u}^\top\mathbf{w}_j^\perp = 0. \tag{12}$$

The scalar c_j is the answer-aligned write magnitude at source position j; the residual $\mathbf{w}_j^\perp$ writes into directions orthogonal to the answer token. Substituting Equation (12) into the per-position score gives the identity

$$\phi_j \;=\; \alpha_j \cdot \|\mathbf{u}\|^2 \cdot c_j, \tag{13}$$

which factorises cleanly into *where* the head reads (α_j) and *what answer-aligned content* it writes from that position (c_j). The factor $\|\mathbf{u}\|^2$ is constant across heads and positions and cancels under any cross-head ranking, so we may treat $\phi_j \propto \alpha_j c_j$ for the purposes of head selection.

Lemma 1 (Sufficient statistic)**.** *For any score that aggregates per-position OV writes against the answer direction* $\mathbf{u}$*, the pair* (α_j, c_j) *is sufficient: any answer-orthogonal component* $\mathbf{w}_j^{\perp}$ *contributes zero.*

Proof. Immediate from $\mathbf{u}^\top \mathbf{w}_j^{\perp} = 0$ and the linearity of $\mathbf{u}^\top \mathbf{o}_j = \alpha_j \mathbf{u}^\top \mathbf{w}_j$. □

Lemma 1 shows that any modification of ϕ that is linear in the residual stream and projects onto $\mathbf{u}$—for instance the tuned-lens variant of Appx. N—inherits the same factorisation, with c_j replaced by a lens-corrected scalar.

O.3. Reduction to attention-based scoring

Proposition 1 (Reduction under position-independent OV write)**.** *Suppose head* (l, h) *writes a position-independent answer-aligned magnitude at step* t*, i.e., there exists a scalar* c_t *such that*

$$c_j \;=\; c_t \quad \textit{for every } j \in \{1, \ldots, N_t\}. \tag{14}$$

Then the spatial-contrast score (Equation (3)) reduces to

$$\Phi_t^+ - \Phi_t^- \;=\; \|\mathbf{u}\|^2 \cdot c_t \cdot \Big(M_t^+ \;-\; \tfrac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} \cdot M_t^- \Big), \tag{15}$$

where $M_t^+ = \sum_{j \in [s_\tau, e_\tau)} \alpha_j$ *and* $M_t^- = \sum_{j \notin [s_\tau, e_\tau)} \alpha_j$ *are the needle and off-needle attention masses respectively. Up to the per-step scalar* $\|\mathbf{u}\|^2 c_t$*,* LOCOS *coincides with the length-normalized attention-mass contrast between needle and off-needle positions.*

Proof. Under Equation (14), Equation (13) reads $\phi_j = \|\mathbf{u}\|^2 c_t \cdot \alpha_j$ with the scalar $\|\mathbf{u}\|^2 c_t$ independent of j. Substituting into the definitions of $\Phi_t^{\pm}$ from Equation (3),

$$\Phi_t^+ = \|\mathbf{u}\|^2 c_t \sum_{j \in [s_\tau, e_\tau)} \alpha_j \;=\; \|\mathbf{u}\|^2 c_t \, M_t^+,$$
$$\Phi_t^- = \tfrac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} \|\mathbf{u}\|^2 c_t \sum_{j \notin [s_\tau, e_\tau)} \alpha_j \;=\; \tfrac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} \|\mathbf{u}\|^2 c_t \, M_t^-.$$

Subtracting yields Equation (15). □

Corollary 1 (Sign of the reduction)**.** *Under Proposition 1,* $\mathrm{sign}(\Phi_t^+ - \Phi_t^-) = \mathrm{sign}(c_t) \cdot \mathrm{sign}\big(M_t^+ - \frac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} M_t^-\big)$*. A head that reads from the needle (*M_t^+ *exceeds the length-normalized off-needle mass) and writes a positive answer-aligned direction (*$c_t > 0$*) is correctly assigned a positive* LOCOS *score; a head that suppresses the answer logit (*$c_t < 0$*) receives a negative score.*

Where the reduction holds. Equation (14) requires the head to write the same answer-aligned magnitude c_t at every source position. This is the defining property of a literal-copy head: an induction head $[A][B] \ldots [A] \mapsto [B]$ writes the "advance the previous token by one step" direction independently of which $[A]$ position it attended to. In our notation, $\mathbf{w}_j$ depends on j only through $\mathbf{v}_j$, and for a literal-copy head $\mathbf{u}^\top W_O \mathbf{v}_j$ takes the same value whenever $\mathbf{v}_j$ encodes the answer token.

Where the reduction fails. For a non-literal retrieval head, c_j depends on whether j lies inside the needle (where the value vector encodes the semantic concept that the head must transform into the answer direction) or outside (where the value encodes unrelated context). Equation (14) breaks, and Proposition 1 no longer holds: LOCOS sees this position dependence through the per-position c_j in Equation (13), while attention-based scoring observes only α_j and is invariant to c_j.

O.4. Wu's token-matching score as a special case

The score of Wu et al. (2025) assigns credit at decode step t to head (l, h) on a literal-NIAH trial when (i) the head's argmax attention position falls within the needle span and (ii) the token at that position matches the generated token. Formally, with $j^* = \arg\max_j \alpha_{t,j}^{(l,h)}$ and x_j the input token at position j,

$$\mathrm{Wu}_t^{(l,h)} \;=\; \mathbf{1}\big[\, j^* \in [s_\tau, e_\tau) \,\big] \cdot \mathbf{1}\big[\, x_{j^*} = y_t \,\big]. \tag{16}$$

Proposition 2 (Wu's score as a hard-thresholded special case)**.** *Suppose three conditions hold for head (l, h) at step t on a literal-NIAH trial:*

(W1) *The head is a literal-copy head: $W_O \mathbf{v}_j \approx \mathbf{u}_{x_j}$, the unembedding row of the attended token, so $\mathbf{u}_{y_t}^\top W_O \mathbf{v}_j \propto \mathbf{u}_{y_t}^\top \mathbf{u}_{x_j}$.*
(W2) *The unembedding rows are approximately orthogonal across distinct tokens: $\mathbf{u}_{y_t}^\top \mathbf{u}_{x_j} \approx \|\mathbf{u}_{y_t}\|^2 \cdot \mathbf{1}[x_j = y_t]$.*
(W3) *Attention is concentrated: $\alpha_{j^*} \gg \alpha_j$ for $j \neq j^*$, so the spatial-contrast score is dominated by the argmax position.*

Then $\mathrm{Wu}_t^{(l,h)} = 1 \iff \Phi_t^+ - \Phi_t^- > 0$ to leading order, and the rankings induced by Wu's score and LOCOS *coincide on the trial.*

Proof. Under (W1) and (W2), $c_j \approx \mathbf{1}[x_j = y_t]$. The needle contribution becomes $\Phi_t^+ \approx \|\mathbf{u}\|^2 \sum_{j \in [s_\tau, e_\tau),\, x_j = y_t} \alpha_j$ and the off-needle contribution $\Phi_t^- \approx \frac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} \|\mathbf{u}\|^2 \sum_{j \notin [s_\tau, e_\tau),\, x_j = y_t} \alpha_j$. On a literal-NIAH trial the answer token y_t appears predominantly at one position inside the needle, so Φ_t^- is negligible. Under (W3), $\Phi_t^+ \approx \|\mathbf{u}\|^2 \alpha_{j^*} \cdot \mathbf{1}[j^* \in [s_\tau, e_\tau)] \cdot \mathbf{1}[x_{j^*} = y_t] = \|\mathbf{u}\|^2 \alpha_{j^*} \cdot \mathrm{Wu}_t^{(l,h)}$, which is positive iff $\mathrm{Wu}_t^{(l,h)} = 1$. □

Proposition 2 clarifies why Wu's score is a strong baseline on literal NIAH: under (W1)–(W3), it is a hard-thresholded version of LOCOS. It also clarifies the failure mode on NoLiMa: assumption (W1) collapses, since a non-literal retrieval head transforms the value vector through a learned $W_O \mathbf{v}_j$ that does not factor through $\mathbf{u}_{x_j}$. The Wu indicator drops to near zero (the attended token is not the answer token) while c_j remains large, producing the $0.97 \to 0.03$ gap reported in Appx. B.

O.5. HeadKV/CompressKV as weighted attention accumulation

Fu et al. (2025) (HeadKV) and Lin et al. (2025) (CompressKV) score heads by accumulating attention mass on retrieval-relevant positions, optionally weighted. A schematic form, abstracting over implementation details, is

$$\mathrm{HeadKV}^{(l,h)} \;\propto\; \sum_{t \in \mathcal{A}^\tau} \sum_{j \in \mathcal{R}_t} w_j \cdot \alpha_{t,j}^{(l,h)}, \tag{17}$$

where $\mathcal{R}_t$ is the set of retrieval-relevant source positions at step t and $w_j \geq 0$ are non-negative weights.

Corollary 2 (HeadKV/CompressKV are non-negative reweightings of attention mass)**.** *Equation (17) is non-decreasing in each $\alpha_{t,j}^{(l,h)}$ for $j \in \mathcal{R}_t$ and depends on the OV circuit only through the choice of $\mathcal{R}_t$ and w_j. In particular, two heads with identical attention pattern $\alpha_{t,\cdot}^{(l,h)}$ but different output projections $W_O^{(l,h)}$ receive the same HeadKV/CompressKV score;* LOCOS *distinguishes them via c_j.*

Proof. The first claim is immediate. The second follows because Equation (17) is a function of α alone and the $\mathcal{R}_t, w_j$ choices are tied to the attention pattern, while Equation (13) is a function of both α and c_j. □

Corollary 2 formalises the central conceptual claim of the paper: any score that observes only the attention pattern, however weighted, is blind to differences in $W_O \mathbf{v}_j$ and therefore cannot distinguish the two heads sketched in Fig. 2(b).

O.6. Spatial vs. temporal contrast

Attention-based methods commonly compute a *temporal* contrast: attention mass accumulated at answer-token decode steps, optionally minus mass at non-answer steps (Wu et al., 2025). LOCOS uses a *spatial* contrast (Equation (3)). The two are not equivalent.

Proposition 3 (Spatial contrast dominates under decode-step-stationary attention)**.** *Suppose head (l, h) exhibits a constant attention pattern across decode steps, $\alpha_{t,j}^{(l,h)} = \alpha_j^{(l,h)}$ for all $t \in \mathcal{A}^\tau \cup \mathcal{A}^{\tau,\text{neg}}$, where $\mathcal{A}^{\tau,\text{neg}}$ is a set of non-answer steps and $|\mathcal{A}^\tau| = |\mathcal{A}^{\tau,\text{neg}}|$. Then the temporal contrast at any single source position is zero, while the spatial contrast is generally non-zero whenever the head allocates more mass to the needle than the length-normalized off-needle expectation.*

Proof. The temporal contrast at position j is $\sum_{t\in\mathcal{A}^\tau} \alpha_{t,j}^{(l,h)} - \sum_{t\in\mathcal{A}^{\tau,\text{neg}}} \alpha_{t,j}^{(l,h)} = (|\mathcal{A}^\tau| - |\mathcal{A}^{\tau,\text{neg}}|) \cdot \alpha_j^{(l,h)} = 0$ by the cardinality assumption. The spatial contrast at any single step t is $M_t^+ - \frac{e_\tau - s_\tau}{N_t - (e_\tau - s_\tau)} M_t^-$, which is non-zero iff the per-position attention mass exceeds the off-needle average. □

Proposition 3 captures the second motivation for the spatial contrast in §3: a head that persistently reads the needle—e.g., as "context" rather than "answer"—receives no credit under temporal contrast but is correctly identified by spatial contrast when its OV write is answer-aligned at needle positions.

O.7. Gradient interpretation

We give the standard derivation of ϕ as a leading-order gradient. Let $\mathcal{L}_t = -\log p(y_t \mid \text{context})$ denote the cross-entropy at step t, with $p(\cdot \mid \text{context}) = \text{softmax}(\boldsymbol{\ell}_t)$. Under the direct-path approximation $\boldsymbol{\ell}_t \approx W_U \mathbf{x}_t^{(L)}$ (Appx. N.1), the chain rule gives

$$\nabla_{\mathbf{o}_{t,j}^{(l,h)}} \mathcal{L}_t \;=\; W_U^\top \nabla_{\boldsymbol{\ell}_t} \mathcal{L}_t \;=\; W_U^\top \big(\mathbf{p}_t - \mathbf{e}_{y_t}\big) \;=\; -(1 - p(y_t))\, \mathbf{u}_{y_t} \;+\; \sum_{v \neq y_t} p(v)\, \mathbf{u}_v, \tag{18}$$

where $\mathbf{p}_t$ is the predicted distribution and $\mathbf{e}_{y_t}$ is the one-hot target. The dominant term is $-(1 - p(y_t))\mathbf{u}_{y_t}$, a positive scalar multiple of $\mathbf{u}_{y_t}$ whenever $p(y_t) < 1$. Therefore the projection $\mathbf{u}_{y_t}^\top \mathbf{o}_{t,j}^{(l,h)} = \phi_{t,j}^{(l,h)}$ captures the leading-order direction of steepest descent of $\mathcal{L}_t$ in the $\mathbf{o}_{t,j}^{(l,h)}$ subspace.

Corollary 3 (Gradient-faithful sign)**.** *Under the direct-path approximation and $p(y_t) < 1$, increasing $\phi_{t,j}^{(l,h)}$ decreases $\mathcal{L}_t$ to first order; ablating a head with large $\phi_{t,j}^{(l,h)}$ should therefore raise the cross-entropy at y_t by an amount proportional to $\phi_{t,j}^{(l,h)}$.*

The corollary motivates the use of ϕ for head selection: heads with large positive ϕ are heads whose ablation should hurt the answer logit most, which is exactly the head set we want LOCOS to identify. Attention-based scoring, by contrast, projects the per-position output onto a uniform direction in j (the all-ones combiner) and is gradient-free with respect to the answer token—which is why it generalizes poorly when the answer-aligned direction varies across attended positions.

O.8. Summary

Tab. 4 consolidates the structural differences derived above.

Table 4. **Attention-based scoring is a special case of LOCOS**, recovered when the OV circuit contributes no position-dependent answer-aligned signal (Proposition 1).

Property	**Attention-based**	**Logit contribution (LOCOS)**
Per-position observable	$\alpha_{t,j}^{(l,h)}$	$\alpha_{t,j}^{(l,h)} \cdot \mathbf{u}_{y_t}^\top W_O^{(l,h)} \mathbf{v}_{t,j}^{(l,h)}$
Includes OV circuit	no	yes
Distinguishes heads with equal α	no (Cor. 2)	yes
Contrast axis	temporal (answer vs. non-answer steps)	spatial (needle vs. off-needle positions)
Requires non-answer steps	yes	no (Prop. 3)
Score sign	non-negative (clamped)	unclamped (negative ⇒ off-needle-dominant; Cor. 1)
Reduction	—	equals attention mass when $W_O \mathbf{v}_j$ is position-independent
Wu/NIAH score	native	special case (Prop. 2, conditions W1–W3)
HeadKV/CompressKV	native	subsumed (Cor. 2)
Gradient interpretation	projects onto $\mathbf{1}$ in j	projects onto $\mathbf{u}_{y_t}$ in $\mathbb{R}^d$ (Eq. 18)

The reductions above predict three empirical patterns:

- On literal NIAH (assumptions W1–W3 hold), Wu/NIAH and LOCOS should rank heads similarly.

- On NoLiMa (assumption W1 fails), the rankings should diverge, with LOCOS assigning credit to heads whose $W_O \mathbf{v}_j$ is answer-aligned even when the attended token is not the answer token.
- Under causal validation, the divergent rankings should manifest as different ablation curves: heads identified only by LOCOS should be more causally critical on NoLiMa than heads identified only by Wu.

The Wu/NIAH-vs-NoLiMa gap reported in Appx. B (top head: 0.97 on NIAH, 0.03 on NoLiMa, no causal effect under ablation) and the ablation-curve separation in Fig. 3 are consistent with all three predictions. The bottom-k control in § 4.4 additionally rules out the trivial alternative that any answer-aligned signal would suffice for causal effect: heads with large $|c_j|$ but off-needle-dominant attention mass leave NoLiMa ROUGE-L near baseline under ablation, exactly as Corollary 1 predicts.

P. Future Work

Head-aware KV cache compression methods (Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024; Fu et al., 2025; Lin et al., 2025) allocate per-head budgets from attention-based retrieval scores. If the heads that matter for non-literal retrieval write rather than attend, then attention-based budgets will under-allocate cache to the heads doing the work and over-allocate it to heads whose attention is bookkeeping. Substituting LOCOS as the scoring function is a direct test of this prediction at fixed cache budget, and Appx. K sketches a per-(layer, KV-group) variant for grouped-query attention. Context-faithful decoding methods (Gema et al., 2025; Ma & Okazaki, 2026) contrast a base model with one whose attention-identified retrieval heads are masked; masking LOCOS heads instead changes which circuit is suppressed and is a candidate for a stronger contrastive signal.