The Intrusion Pathway: How Non-Evidential Context Degrades Language Model Performance

Anonymous ACL submission

Abstract

Language models lose performance when prompts contain non-evidential context — text that carries no competing answer and no semantic overlap with the query, including irrelevant statements from unrelated domains and random off-distribution tokens. Why it happens mechanistically is not known. We trace the corruption to a four-stage **Intrusion Pathway**. Early-layer parroting heads deposit prefix content as a diffuse, attention-weighted superposition into the last-token residual. Mid-block relational readers gate on the answer-slot frame and stay closed under non-evidential context, leaving the request representation intact. The damage is then written by a distinct set of late-layer attention heads — **Intrusion heads** — whose OV writes a broad category-boost across the gold-attribute semantic cluster regardless of what the QK pulls into the value. Under non-evidential context this fan-out floods the END residual with semantic neighbours and suppresses the gold via softmax redistribution. The settings where models are deployed at scale — retrieval, tool use, long context — are precisely the settings where non-evidential context dominates. The Intrusion Pathway is the mechanism behind that failure; head-level and path-level interventions remove the damage and can even leave the model above the unmodified baseline.

1 Introduction

A current-day frontier language model or agent rarely sees a bare question. A user's prompt arrives bundled with retrieved passages from a vector store (Lewis et al., 2020; Karpukhin et al., 2020; Izacard et al., 2022), tool-call results from a browsing or search API (Schick et al., 2023; Nakano et al., 2021; Qin et al., 2024), conversation snippets, and traces from agentic sub-loops (Yao et al., 2023; Shinn et al., 2023; Wang et al., 2023a). The model is expected to weld this heterogeneous bundle into a single coherent answer, and the bundle is rarely clean: context closes real factual gaps that parametric memory cannot reach on its own (Mallen et al., 2023; Kandpal et al., 2023), but the same channel just as routinely carries pieces that share none of the answer's content. The pieces of this mixture fall into a now-standard five-way partition—*golden*, *counterfactual*, *related*, *irrelevant*, and *random* context—established across the retrieval-augmentation, knowledge-conflict, and contextual-distraction literatures (Cuconasu et al., 2024; Fang et al., 2024; Niu et al., 2025; Wu et al., 2025a).

For a query q with answer $a^\star$, the usage above partitions any context c into five disjoint classes, each instantiated in Table 1 on the running query *"The capital of Germany is ___"* ($a^\star$ = Berlin). *Golden* context contains $a^\star$ and closes factual gaps the model would otherwise miss (Mallen et al., 2023; Kandpal et al., 2023). *Counterfactual* context asserts a coherent competing answer via the same relation as the query, and is the substrate of the knowledge-conflict literature (Xie et al., 2024; Longpre et al., 2021; Du et al., 2024; Kortukov et al., 2024; Niu et al., 2025; Neeman et al., 2023). *Related* context shares the relational schema of q but asserts about a different entity, functioning as a one-shot demonstration of the task: this is the canonical in-context learning regime (Olsson et al., 2022; Elhage et al., 2021; Hendel et al., 2023; Todd et al., 2024). *Irrelevant* context is a coherent statement from an unrelated domain—no shared entity, relation, or topic with q—and contains one or more salient non-gold (distractor) tokens; it is the dominant failure mode of context-augmented systems, dropping accuracy by 25% to 67% (Cuconasu et al., 2024; Wu et al., 2024; Shi et al., 2023). *Random* context is an arbitrary token or short string with no semantic content, a control for whether the model keys on distribution rather than content (Cuconasu et al., 2024).

Three of the five above-mentioned classes have

Class	Example c	Logit diff
Golden	"Germany's capital city is Berlin."	+4.8
Counterfactual	"The capital of Germany is Munich."	−2.6
Related	"The capital of France is Paris."	+1.8
Irr. (single token)	"Computers are useful."	−0.7
Irr. (multi-token)	"Apples are ... from ripening."	−2.2
Random	"Calculator."	−0.4

Table 1: The five-way partition of context, instantiated on *"The capital of Germany is ___"* ($a^\star$ = Berlin). The **Logit difference** column shows the gold token logit change caused in nats by that class. "Irr." abbreviates Irrelevant.

circuit-level accounts. **Counterfactual context** has been localized at every grain—at the FFN level by ROME (Meng et al., 2022), at the activation level by Geva et al.'s (2023) subject-enrichment plus knowledge-hub heads, Jin et al.'s (2024) memory-vs-context split, Ortu et al.'s (2024) identification of L10H7 and L11H10 as suppression-based factual-promoter heads, Yu et al.'s (2023) individual head-level analysis with domain-specific memory heads, Monea et al.'s (2024) last-subject-MLP grounding signature, and Niu et al.'s (2025) contextual entrainment heads.

Golden context is anchored by Wang et al.'s (2023b) IOI circuit and its Name Movers, Negative Name Movers, and S-Inhibition machinery, with McDougall et al.'s (2023) copy-suppression unification, Variengien and Winsor's (2023) request-patching generalization, Lieberum et al.'s (2023) scaling to Chinchilla 70B, and Feng and Steinhardt's (2024) binding-ID substrate.

Related/ICL context starts from induction heads (Olsson et al., 2022; Elhage et al., 2021) with parroting (Goldowsky-Dill et al., 2023), compressed at mid-layer into task vectors (Hendel et al., 2023) and function vectors (Todd et al., 2024), with fully traced circuits for successor heads (Gould et al., 2024) and greater-than (Hanna et al., 2023).

The remaining two classes—**irrelevant** and **random**—have no comparable account. A parallel architectural line studies position-induced attention bias (Chowdhury, 2026), orthogonal to the content-conditional gold-suppression we characterize, and does not address why specific context classes selectively degrade model behavior. The literature has produced behavioral fixes—adversarial noise training (Fang et al., 2024), fine-tuning on mixed-context examples (Yoran et al., 2024), retrieve-and-critique pipelines (Asai et al., 2024), chain-of-note prompting (Yu et al., 2024), K/Q-activation steering toward behaviorally-identified "contextual heads" (Zhu et al., 2025), and benchmark-driven noise characterization (Wu et al., 2025a)—each of which treats non-evidential context as a symptom to be patched, not a mechanism to be characterized. This is the gap the present paper closes.

2 Setup

The Intrusion Pathway is characterized on two circuit benchmarks — one factual-recall, one binding/retrieval — and validated on four downstream tasks. Every analysis in §3 holds the query, gold token, and model fixed, varying only the prefix; this design eliminates per-sample difficulty, tokenization, and parametric-knowledge confounds.

Circuit benchmark. We construct a paired-comparison benchmark spanning two task families: *factual recall* (62,500 tuples drawn from 17 LRE relations covering factual associations, categorical relations, and geographic composition; Hernandez et al., 2024) and *binding/retrieval* (Indirect Object Identification prompts generated combinatorially over the name / place / object pools and template inventory of Wang et al., 2023b). Each sample is instantiated under seven prefix conditions corresponding to the partition of Table 1, with the irrelevant class split into short (∼12-token) and long (∼100-token) variants to separate prefix length from prefix content. Long-irrelevant prefixes are sampled from WikiText-103 (Merity et al., 2016) under a five-level leakage filter that rejects any paragraph containing the gold token, the subject token, same-relation alternatives, or its semantic neighbours (Appendix A.10). Robustness sweeps over length, distractor count, and distractor position, together with per-relation counts, IOI templates, and representative samples, are in Appendix A.

Models. Circuit analyses use four 7B-scale open-weight base models — Llama-3.1-8B, Mistral-7B-v0.1, Qwen2.5-7B, and Pythia-6.9B — instrumented via TransformerLens (Nanda and Bloom,

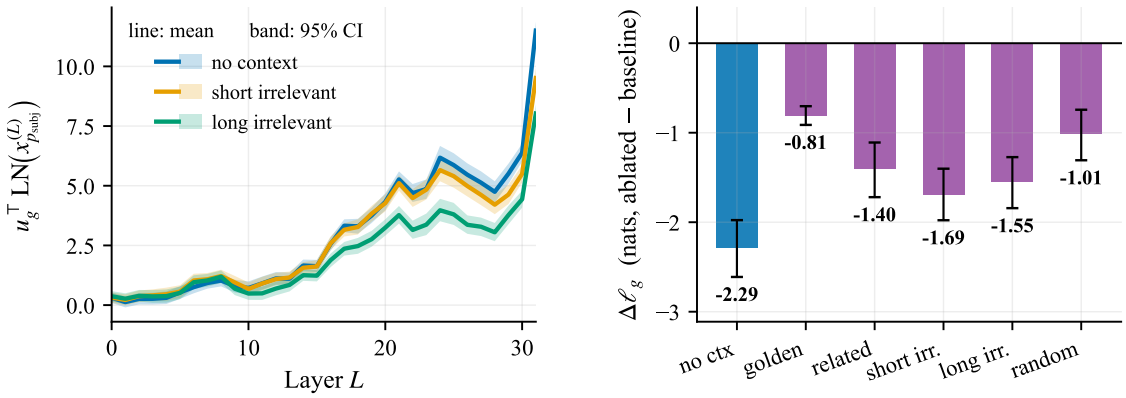

Figure 1: Llama-3.1-8B; bands and bars are 95% bootstrap CIs. (a) Gold-direction projection at the subject tracks across conditions (Pearson $r \geq 0.99$ for initial layers). (b) Cancelling early-band subject-MLPs damages the bare-query baseline comparably to all other prefixes. Remaining models in Appendix B.

2023). The downstream evaluation (§4) additionally reports on larger models.

Notation. Standard transformer notation throughout; context conditions in monospace (`nctx`, `golden`, `counterfactual`, `short_irr`, `long_irr`, `random`). Full glossary: Appendix A.15.

3 The Intrusion Pathway

Table 1 establishes that irrelevant context suppresses gold-token recall, but a bare performance gap says nothing about how the suppression propagates through the forward pass. We therefore trace the computation stage by stage, mapping the mechanistic route by which intrusive information enters, transforms, and ultimately disrupts the output distribution. This tracing serves two purposes. First, it is an interpretability contribution in its own right: understanding the full pathway reveals which components carry the intrusive signal and why certain prefix classes cause more damage than others. Second, it enables path patching — a targeted intervention that severs the information route at its entry point rather than ablating entire components, viable only once the upstream pathway is characterized enough to identify where intrusion branches off from retrieval.

A stage qualifies as a viable damage carrier only if (i) it fires differentially under non-evidential prefixes (`short_irr`, `long_irr`, `random`) relative to `golden` and `nctx`, and (ii) ablating it damages non-evidential performance more than golden-context performance. We trace four stages in order of depth; each following subsection tests both criteria.

3.1 Subject-MLP enrichment is prefix-invariant

The pipeline begins with *subject enrichment*: early-to-mid-layer MLPs at the last subject-token position assemble an "attribute cloud" in the residual stream (Geva et al., 2023; Meng et al., 2022) — the substrate from which downstream heads extract the answer. This enrichment fires identically under every prefix class — including `short_irr`, `long_irr`, and `random` — writing the same attribute cloud regardless of what precedes the subject.

The per-layer logit-lens trajectory of the gold token at the subject position is statistically indistinguishable across all prefix conditions: Pearson $r \geq 0.985$ in every model$\times$condition cell (Table 4; Figure 1a). The gold arrives at the end of the enrichment band intact. Early-layer attention does attend back to prefix positions, but the resulting writes stay outside the attribute subspace of the unembedding.

A section-level ablation confirms this stage is not a viable defense surface either: cancelling early-band subject-MLPs damages `nctx` and non-evidential performance comparably (Figure 1b) — the enrichment is a shared infrastructure, not a class-conditional damage carrier. Six convergent diagnostics in Appendix B confirm Stage 1 invariance across prefix classes: the per-layer u_g-projection, Attributes Rate, median gold rank, attribute-restricted top-30 Jaccard, causal early-band MLP-cancellation drop, and section-level MLP ablation are all statistically indistinguishable.

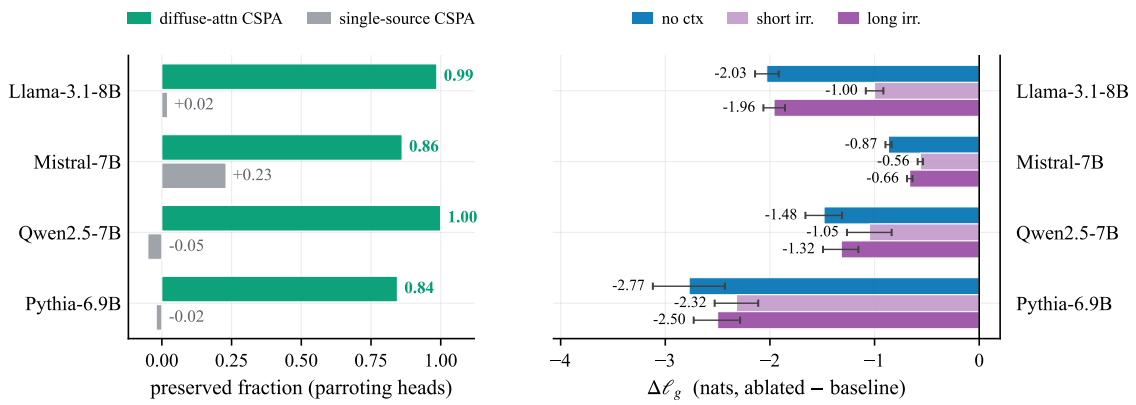

Figure 2: All four 7B baselines. (*left*) Under irrelevant context, single-source CSPA collapses to ≤ 0.23 on the canonical parroting heads, while diffuse-attention CSPA recovers values near 1.0 on the same heads — the OV circuit does identical work; only the target direction is mis-specified. (*right*) Stage-2 section-level ablation: zeroing the parroting head set damages `nctx` as much as or more than irrelevant-context on every model, ruling out class-conditional defense at this stage. Error bars are 95% bootstrap CIs. Full per-condition table in Appendix C.

3.2 Parroting heads diffuse across non-evidential content

At early-to-mid layers, a small set of attention heads deposits in-context content into the end-of-prompt residual via a near-identity OV circuit — a behaviour termed parroting (Goldowsky-Dill et al., 2023; Olsson et al., 2022), with the OV copy mechanism characterised by McDougall et al. (2023). We find the same canonical head set fires across all prefix conditions — but the attention is class-conditional, so only the distribution changes with the prefix type.

Under no context, attention defaults to BOS and query tokens. Under single-distractor irrelevant context (one salient off-topic token), attention concentrates on that distractor. Under multi-token irrelevant context, attention spreads across the k prefix content tokens (Figure 11). In every case the OV circuit of the parroting heads performs the same operation — it writes the attention-weighted sum of source-token embeddings into the END residual:

$$o_h(p_{\text{end}}) = \sum_p A_h[p_{\text{end}}, p] \cdot W_O W_V \, x^{(L)}[p]. \tag{1}$$

Single-source CSPA underestimates diffuse parroting. McDougall et al. (2023) introduce a single-source preserved fraction that projects the head's OV output (Eq. 1) onto one chosen source token's unembedding direction. Under single-distractor irrelevant context, where $A_h[p_{\text{end}}, p_{\text{distractor}}] \approx 0.8$, this projection captures most of the head's work and the preserved fraction is high ($\rho^{\text{single}} \approx 0.8$–$1.0$). Under multi-token irrelevant context, attention distributes across $k = 5$–50 content tokens, each source now contributes $\approx 1/k$ of the attention mass. The projection onto any single source token captures $\approx 1/k$ of the full OV output — for a 10-token irrelevant sentence, the preserved fraction drops from ~ 0.9 to ~ 0.1 despite the OV circuit doing identical work. This collapse is a measurement artifact: the CSPA target direction is mis-specified for multi-token prefixes, not a change in the underlying mechanism (see Appendix C).

An attention-weighted target recovers the signal. We replace the single-source target with an attention-weighted target direction that tracks where the head is actually looking:

$$u_*^{\text{attn}} = \sum_{p \in \mathcal{P}_{\text{content}}} A_h[p_{\text{end}}, p] \, u_{t_p}. \tag{2}$$

The diffuse preserved fraction computed against this target recovers values in the range 0.56–1.00 under multi-token irrelevant context, compared to 0.02–0.20 under the single-source measurement on the same data. A shuffled-attention control — keeping the head's actual OV output but randomly permuting its attention pattern over key positions — drops the preserved fraction by 0.16–0.55 with non-overlapping 95% CIs in all 12 model×condition cells (Appendix C), confirming the metric measures attention-correlated copying rather than a generic OV positive-projection artifact.

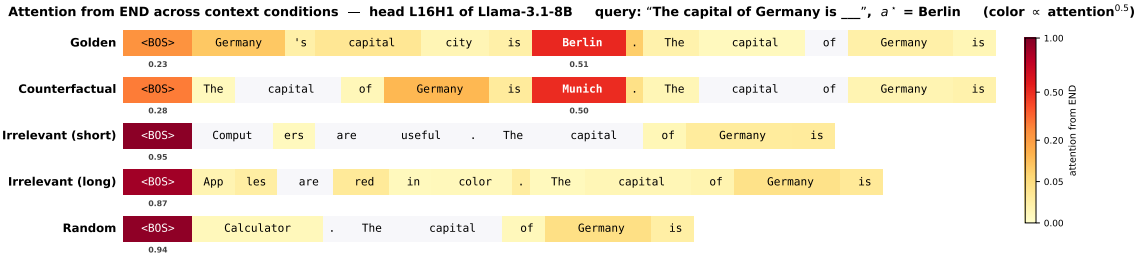

Figure 3: Attention from END for relational reader L16H1 of Llama-3.1-8B across context conditions. Under *golden* and *counterfactual* prefixes the QK gate opens onto the answer-slot token; under *irrelevant* and *random* prefixes no answer-slot frame exists, the gate stays closed, and the head writes nothing into the END residual — preserving mid-block invariance. Remaining models in Appendix D.

Implication for the late block. The **diffused parroting** result specifies what enters the late block under irrelevant context. The END residual carries not a single distractor-token direction (as it would under random or single-distractor context) but an attention-weighted superposition $\sum_p \alpha_p \, u_{t_p}$ of source-token embeddings from across the irrelevant prefix. This superposition is per-sample variable — it depends on which prefix tokens the parroting heads happened to attend to — and it is the input on which late-block extraction heads (§3.4) will operate. The mechanistic implication is that the late-block damage should also be per-sample variable and diffuse rather than concentrated in one token direction. We confirm this in §3.4.

Ablation confirms the defense impossibility. Ablating the canonical parroting set damages golden context as much as or more than irrelevant context on three of four models (Figure 2, right) — the same heads that diffusely deposit irrelevant content do their primary work boosting in-context gold when it is present. Removing them to suppress irrelevant-context damage would simultaneously destroy the model's ability to use helpful context. Seven convergent diagnostics — the full head identification protocol, per-condition re-identification, rank-stability checks, single-source vs. diffuse-CSPA comparisons, the shuffled-attention control, region-wise attention breakdowns, and the section-level ablation — all corroborate Stage 2's class-invariant identity and class-conditional attention across four models (Appendix C).

3.3 Mid-block relational readers gate on answer-slot structure

The mid-block attention heads that propagate the query's relation to the END position (Geva et al., 2023; Variengien and Winsor, 2023) possess a structural gate: their QK circuit fires attention onto context positions only when the prefix contains a syntactic frame matching the query's answer slot. Under golden context, the prefix contains "Germany's capital city is Berlin" — the answer-slot frame is present, the gate opens, and the head reads the gold attribute into the END residual (Figure 3). Under counterfactual and related context, a competing or analogous frame opens the same gate. Under irrelevant and random context, no answer-slot frame exists anywhere in the prefix. The gate stays closed. The head writes nothing into the END residual, and the mid-block representation proceeds identically to the bare-query baseline. The gold logit at the mid-block is unaffected not because the model resists irrelevant context, but because the mid-block computation never sees it — the QK gate is structurally blind to content that does not instantiate the query's relational schema. Causal prompt samples confirm the gate is primarily slot-match-driven: the head checks whether the prefix instantiates the query's syntactic frame, not whether the content filling that frame is type-correct (§D.6).

The residual-level consequence follows directly. We replicate three independent protocols — Hendel–Todd task-vector cosine (§D.5), Variengien–Winsor request-patching (§D.4), and per-layer logit-lens trajectories (§D.3) — and all three converge on the same picture: the END residual at the mid-block is indistinguishable between `nctx` and non-evidential context. Ablating the mid-block MLP collapses gold equally under `nctx` and non-evidential prefixes (§D.9), ruling it out as a defense surface.

§3.4 asks the next question: which late-block component reads the composite residual inherited from these three upstream stages, and how does it produce the gold-suppression gap Table 1 measures?

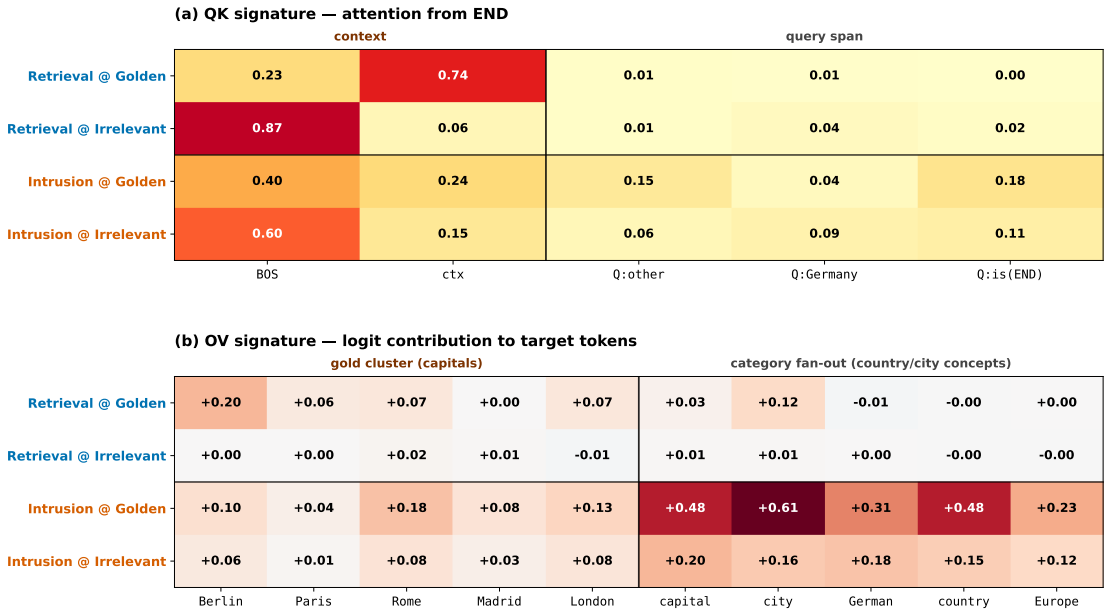

Figure 4: Retrieval head L25H19 vs. intrusion head L10H11 of Mistral-7B on the running Germany→Berlin query. *(a)* QK attention from END; *(b)* per-target OV logit contribution. The retrieval head goes silent under irrelevant (writes ≈0); the intrusion head writes the same broad category-boost (*capital*/*city*/*German*) under both conditions — the OV direction fires regardless of what the QK pulls.

3.4 Intrusion heads: a weight-disjoint late-block damage carrier

The earlier stages of the pathway divide into two patterns. Stages 1 and 2 identified components — subject-position MLPs and parroting heads — whose ablation damages the bare-query baseline as much as it damages irrelevant-context performance; these are shared infrastructure, not class-conditional damage carriers. Stage 3 identified a component — the mid-block relational reader — that is structurally blind to non-evidential context and therefore leaves the request representation untouched. Neither pattern produces a viable defense surface: the first cannot be removed without breaking no-context performance, and the second carries no damage to remove.

What remains uncharacterized is the late block, where extraction heads read from the END residual to produce the final answer. The content deposited by parroting heads — an attention-weighted superposition of irrelevant prefix embeddings — is now available for any late-block head to read.

We measure, for every late-block attention head, two quantities under irrelevant context: the fraction of its attention from END directed onto prefix positions, and the change in its direct-logit attribution to the gold token relative to the bare-query baseline. A sparse set of heads stands out. Under irrelevant context, these heads shift attention onto the prefix and produce two simultaneous effects at the output distribution: the logit of the salient distractor token in the prefix rises (distractor entrainment), and the logit of the gold token falls (gold suppression) — even though the prefix contains nothing semantically related to the answer.

The OV circuit of these heads explains why. Their OV matrices do not write a specific token direction into the END residual. Instead they write a fan-out across the gold-attribute's semantic category — a generic category-boost direction. When the QK circuit pulls irrelevant prefix content into the value, the OV projection onto this broad cluster direction remains non-trivial regardless of what was pulled: the head writes "country vocabulary" into the END residual when the query asks for a capital, or "language vocabulary" when the query asks for a language — distributing probability mass across the gold token's semantic neighbours and thereby suppressing the specific correct answer through softmax normalization. Decoding the top tokens from each head's average residual-stream contribution confirms this: "European continental Western geography region nations countries" for capital-relation queries, rather than the specific gold token "Berlin" (§E.5). We call these heads *intrusion heads*.

A structural mirror exists under golden context. Wu et al.'s (2025b) retrieval heads — a different sparse late-block set — also attend to the prefix, but their OV writes the precise gold-token unembedding ("Berlin" for the capital-of-Germany query, "French" for a language-of-France query) rather than the category fan-out. The specificity is also their gate: under irrelevant context the QK pulls content outside the gold cluster, the precise direction produces negligible output, and the head stays silent.

The structural parallel raises an immediate question: are intrusion and retrieval the same heads operating on different inputs — as Stage 2's parroting result would predict — or are they separable populations? If they share heads, no class-conditional defense is possible.

Disjointness. We score every late-block head on both axes simultaneously. The retrieval score rewards attention-to-gold-position under golden context weighted by positive gold DLA. The intrusion score rewards attention-to-prefix under irrelevant context weighted by gold-logit suppression:

$$s_{\text{int}}(h) = \tfrac{1}{2} \sum_{c \in \{s,l\}} \text{mean}_i \Big[A_h^{(i,\texttt{c_irr})}[p_{\text{end}}, \text{ctx}] \cdot \max\big(0,\ \text{DLA}_h^{(i,\texttt{nctx})} - \text{DLA}_h^{(i,\texttt{c_irr})}\big) \Big], \tag{3}$$

where DLA on the gold token is left implicit and the retrieval score is the analogous `golden`-context measure (§E.2). Plotting the two scores per head produces a clean L-shape (Figure 5; all four models in Figure 22): retrieval-only heads concentrate along the x-axis, intrusion-only heads along the y-axis, and dual-use heads in the upper-right corner are rare — $|R \cap I| = 4/20$. At the weight level — independent of input data — the OV-copy score (Wang et al., 2023b) corroborates this split: retrieval-only heads' OV preserves token identity at $\approx 3\times$ above random, while intrusion-only heads' OV is indistinguishable from random — they do not literally copy any token, they category-boost ($p < 10^{-3}$, Mann–Whitney; §E.4).

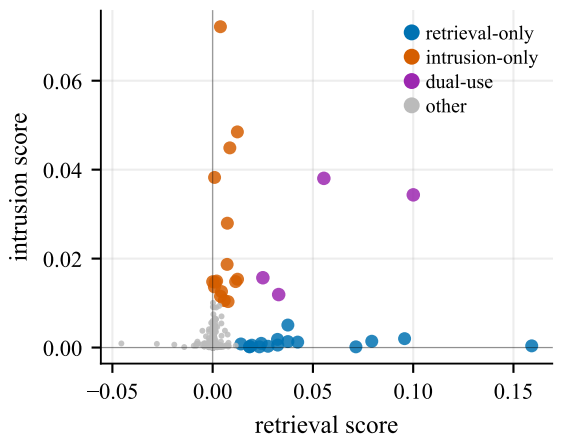

Figure 5: Per-head retrieval score vs. intrusion score on Mistral-7B-v0.1. Intrusion-only heads (orange) concentrate on the y-axis (high intrusion, near-zero retrieval); retrieval-only heads (blue) on the x-axis.

A class-conditional defense. The disjointness makes a surgical intervention possible — one that was unavailable at any earlier stage. We zero, only on intrusion heads disjoint from the retrieval set, the pre-softmax attention from END onto context-prefix positions; softmax redistributes the freed mass to non-context positions. Because intrusion heads stay silent under golden context — their broad category-boost OV produces negligible output when the QK reads gold-cluster content that falls outside their fan-out's activation regime — the intervention leaves golden-context performance intact and `nctx` invariant by construction. The irrelevant-context gold logit recovers by 0.06–0.30 logits on three of four models, with golden-context cost bounded by ≤ 0.02 logits (§E.7). This is the structural difference from Stages 1–3: at every earlier stage, removing the component simultaneously destroyed the parametric baseline. Here, the targeted computation fires under non-evidential context and stays silent under every evidential class.

4 Downstream Evaluation

The four-stage residual-level localization (§3.1–§3.4) unlocks the downstream intervention: a per-step residual patch at the intrusion-band layer that replaces the polluted with-context residual with the matched `nctx` residual, applied at the depth the pathway predicts.

We evaluate this layer patch and a complementary head-level knockout across four downstream tasks (GSM-DC, GSM-IC, HotpotQA, PubMedQA) and eight models (Llama-3.1-{8B, 70B}, Mistral-7B-v0.1, Qwen2.5-{7B, 32B, 72B}, Pythia-{6.9B, 12B}). **(i) Layer-level residual patching** closes 86–99% of the distractor-induced accuracy drop and exceeds the clean baseline on GSM-IC at multiple scales. **(ii) Head-level attention knockout** on the ∼10 clean intrusion heads (less than 1% of late-block heads) captures ∼60%

Task	+Patch	+K10	+K20
GSM-DC	96.7%	57.3%	28.8%
GSM-IC	108.8%	58.5%	27.8%
HotpotQA	85.5%	48.5%	29.5%
PubMedQA	99.3%	63.3%	35.9%
mean	**97.6%**	**56.9%**	**30.5%**

Table 2: Per-task recovery of the distractor-induced accuracy drop, averaged across 7B baselines. **+Patch**: layer-level residual patch; **+K10**: clean intrusion-head knockout ($\sim$ 10 heads, $<$ 1% of late-block); **+K20**: top-20 set including dual-use heads. GSM-IC row from per-step CoT decoding; per-model breakdowns in Appendix F.

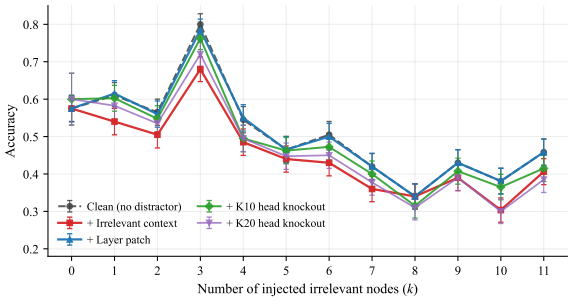

Figure 6: GSM-DC: layer-level residual patching closes the irrelevant-context gap at every distractor intensity; clean intrusion-head knockout (K10) recovers a substantial fraction with $\sim$10 heads. Error bars are binomial SE; per-model breakdown in Appendix F.

of the same recovery; the top-20 set including dual-use heads recovers less (Table 2). This two-tier split — full damage at the layer, an isolable fraction at a small head population, the residual distributed across dual-use heads and upstream effects — is exactly what §3.1–§3.4 predicts.

We report GSM-DC and GSM-IC below; the mechanistic LRE signature, per-model GSM-DC breakdown, GSM-IC single-token and chain-of-thought variants, and per-k HotpotQA / PubMedQA tables are in Appendix F.

GSM-DC. We construct a DAG-distractor variant of GSM8K (Cobbe et al., 2021): each problem is a directed acyclic graph of arithmetic dependencies, into which we inject k irrelevant nodes. Across the four 7B baselines for $k \in [0, 11]$, the layer-patched trajectory is statistically indistinguishable from the clean baseline at every k, and the K10 head-knockout recovers a substantial fraction of the drop with a far more targeted intervention (Figure 6). Curve-level detail and the K20 dual-use cost are in Appendix F.

GSM-IC. On GSM-IC (Shi et al., 2023) — the standard irrelevant-context distractor benchmark on GSM8K — layer patching exceeds the clean baseline (100–184% recovery) across five model scales; per-step chain-of-thought recovery for the 7B baselines feeds Table 2. Per-model breakdowns are in Appendices F.3 and F.4.

5 Conclusion

We identify a four-stage *Intrusion Pathway* by which non-evidential context degrades language-model performance. Stage-1 subject-MLP enrichment assembles every attribute's representation in the residual stream at the subject position, the substrate from which downstream heads extract the answer. Within Stage 2, the *diffuse-parroting* analysis shows how multi-token irrelevant content enters the model: parroting heads deposit a spread-out superposition of prefix tokens into the END residual, layering the irrelevant content on top of the subject-enriched signal already present there. Mid-block relational readers remain structurally blind to prefixes that lack the query's answer-slot frame, so the late block inherits the polluted residual — where a sparse population of *intrusion heads* broadcasts across the gold attribute's semantic neighbourhood, displacing the gold token in the output distribution.

Ablating these heads — under 1% of late-block heads — substantially reduces the degradation but stops short of erasing it. The stronger intervention is suggested by the pathway itself: at the intrusion-band in each layer, the polluted with-context residual is replaced with its bare-query counterpart, severing the full route at the exact depth the four-stage trace identifies. This pathway-level patch recovers almost all of the distractor-induced drop across downstream tasks and, in several settings, leaves the model *above* its unmodified baseline. This is the cleanest evidence that the pathway we localize is the dominant carrier of the failure.

Retrieval-augmented generation, tool use, agentic loops, and long-context settings — the regimes where models are deployed at scale — are precisely the regimes where non-evidential content dominates, and the behavioural-fix literature has been treating the symptom rather than the cause. We show instead that the cause has a circuit, the circuit has a defense, and the defense is class-conditional by construction — effective under non-evidential context, silent everywhere else.

Limitations

Our study has several limitations that constrain the scope of the claims and suggest directions for future work. The circuit-level localisation is performed on four 7B-scale open-weight decoder-only transformers (Llama-3.1-8B, Mistral-7B-v0.1, Qwen2.5-7B, Pythia-6.9B); the downstream evaluation extends to a few larger Llama and Qwen scales, but we do not characterise the full pathway on >70B models, on mixture-of-experts architectures, on encoder–decoder models, or on instruction-tuned variants — the intrusion-head population is structurally similar across the four base models, but whether the same population (and the same class-conditional intervention) transfers to RLHF'd, multimodal, or much-larger checkpoints remains open. The mechanistic analysis is run on a paired-comparison benchmark spanning two task families (factual recall on 17 LRE relations and IOI binding/retrieval), and the downstream interventions are evaluated on four English tasks (GSM-DC, GSM-IC, HotpotQA, PubMedQA); the non-evidential prefix pool is drawn from WikiText-103 and a 60-sentence short-irrelevant pool, so we do not test code, math beyond GSM, multilingual prompts, dialogue, or genuinely adversarial / jailbreak-style prefixes — we expect the pathway to generalise (the circuit is structural), but this is an empirical claim that needs explicit verification. Finally, the class-conditional defense zeros pre-softmax attention from END onto context-prefix positions for the clean intrusion-head set; we do not explore softer interventions (attention dampening, value-projection nullification, partial OV-direction subtraction) nor do we train an intervention rather than apply it post-hoc, so the reported recovery numbers are lower bounds on what is achievable with the same mechanistic localisation.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*.

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

Borun D. Chowdhury. 2026. Lost in the middle at birth: An exact theory of transformer position bias. *Preprint*, arXiv:2603.10123.

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

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 719–729.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 10028–10039.

Jiahai Feng and Jacob Steinhardt. 2024. How do language models bind entities in context? In *International Conference on Learning Representations (ICLR)*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12216–12235.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations (ICLR)*.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations (ICLR)*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations (ICLR)*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 15696–15707.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling (COLM)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 9459–9474.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? Evidence from multiple choice capabilities in Chinchilla. *arXiv preprint arXiv:2307.09458*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7052–7063.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 9802–9822.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2023. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 17359–17372.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. 2024. A glitch in the matrix? Locating and detecting language model grounding with Fakepedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 6828–6844.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Neel Nanda and Joseph Bloom. 2023. TransformerLens: A library for mechanistic interpretability of generative language models. https://github.com/neelnanda-io/TransformerLens.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 10056–10070.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. 2025. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 16218–16239.

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 31210–31227.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*.

Alexandre Variengien and Eric Winsor. 2023. Look before you leap: A universal emergent decomposition of retrieval tasks in language models. *arXiv preprint arXiv:2312.10091*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 2025a. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 5019–5039.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? In *First Conference on Language Modeling (COLM)*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025b. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations (ICLR)*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset

for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations (ICLR)*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9924–9959.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14672–14685.

Youxiang Zhu, Ruochen Li, Danqing Wang, Daniel Haehn, and Xiaohui Liang. 2025. Focus directions make your language models pay more attention to relevant contexts. *Preprint*, arXiv:2503.23306.

A Setup

A.1 Factual-recall benchmark: relation composition

The factual-recall benchmark draws on 17 relations from LRE (Hernandez et al., 2024), selected to span the three canonical knowledge regimes that the factual-recall literature has characterized at the circuit level. Factual associations — the regime ROME (Meng et al., 2022) and the Geva–Bastings subject-enrichment account (Geva et al., 2023) localize to mid-layer subject MLPs — are represented by relations such as `country_capital_city`, `country_language`, and `person_plays_instrument`. Categorical relations, the regime that produced the Hernandez *et al.* LRE benchmark itself, are represented by `object_superclass` and its kin. Geographic composition — the regime where binding-ID mechanisms (Feng and Steinhardt, 2024) have been characterized — is represented by `city_in_country` and `landmark_in_country`.

Permutation expansion. The base benchmark contains 250 unique (s, r, q, g) tuples — subject, relation, query template, gold attribute — drawn from the 17 LRE relations. To support the paired-comparison design described in Appendix A.4, each base tuple is paired with 250 prefix instantiations (a fresh sample of irrelevant prefixes from WikiText-103 under the leakage filter of Appendix A.10, plus per-sample counterfactual / related / random selections). This expands the benchmark to $250 \times 250 = 62{,}500$ unique (sample, prefix) tuples for factual recall, before applying the seven-way prefix-condition split.

Relation distribution. The relation-sample distribution at the base-tuple level is intentionally unbalanced across the 17 relations. Per-head behavioral analyses in §3.3 and §3.4 require ≥ 50 samples per relation for stable bootstrap intervals, which a small number of higher-coverage relations supply directly; the remaining relations supply breadth for population-level metrics aggregated across all 250 base samples. The qualitative conclusions of §3 hold within every relation taken individually.

A.2 Binding/retrieval benchmark: IOI prompts

To complement the factual-recall task family, we include the Indirect Object Identification (IOI) benchmark of Wang et al. (2023b), which probes a structurally distinct circuit: binding entity tokens to syntactic roles and retrieving the correct role-bound entity at the END position. We follow Wang et al.'s prompt-generation procedure exactly — IOI prompts are produced by combinatorial sampling over four pools:

- **Name pool** (109 single-token first names from the Wang-2023 release).
- **Place pool** (e.g., *store*, *school*, *park*, *garden*, *restaurant*, *hospital*).
- **Object pool** (e.g., *drink*, *ball*, *book*, *ring*, *kiss*, *pen*).
- **Template inventory** of 4 syntactic frames, e.g. "When [Name$_1$] and [Name$_2$] went to the [place], [Name$_i$] gave a [object] to ___" and analogous "At/Then/After" variants.

Each prompt has the canonical IOI structure [Name$_A$ . . . Name$_B$ . . . Name$_S$ gave/received . . .], with the gold answer $a^\star$ being the indirect object

(the name in the IO syntactic role). The dataset is balanced across **ABBA** and **BABA** surface orderings (the IO appears in the first or second position with equal frequency), exactly as in Wang et al. (2023b). The full templating yields 10,630 unique IOI prompts in our canonical configuration; each prompt is instantiated under the same seven-way prefix split as the factual-recall samples (`nctx`, `golden`, `related`, `counterfactual`, `short_irr`, `long_irr`, `random`), so the two task families share identical paired-comparison machinery and downstream analysis pipeline.

The combined design — 62,500 factual-recall tuples plus 10,630 IOI prompts, each instantiated under seven prefix conditions — yields $> 510{,}000$ (sample, condition) cells per model.

A.3 Representative data samples

We provide concrete examples to make the seven-way prefix split unambiguous. The running query is *"The capital of Germany is ___"* ($a^\star =$ Berlin) for factual recall and *"When Mary and John went to the store, John gave a drink to ___"* ($a^\star =$ Mary) for IOI.

Factual recall (LRE `country_capital_city`).

- `nctx`: `"The capital of Germany is"`
- `golden`: `"Germany's capital city is Berlin. The capital of Germany is"`
- `related`: `"The capital of France is Paris. The capital of Germany is"`
- `counterfactual`: `"The capital of Germany is Munich. The capital of Germany is"`
- `short_irr`: `"Computers are useful. The capital of Germany is"`
- `long_irr`: `"⟨∼100-token WikiText-103 paragraph passing the five-level leakage filter⟩ The capital of Germany is"`
- `random`: `"Calculator. The capital of Germany is"`

IOI binding/retrieval.

- `nctx`: `"When Mary and John went to the store, John gave a drink to"`
- `golden`: `"Mary received the drink. When Mary and John went to the store, John gave a drink to"`
- `related`: `"When Alice and Bob went to the cafe, Bob gave a book to Alice. When Mary and John went to the store, John gave a drink to"`
- `counterfactual`: `"John received the drink. When Mary and John went to the store, John gave a drink to"`
- `short_irr`: `"Computers are useful. When Mary and John went to the store, John gave a drink to"`
- `long_irr`: `"⟨∼100-token WikiText-103 paragraph⟩ When Mary and John went to the store, John gave a drink to"`
- `random`: `"Calculator. When Mary and John went to the store, John gave a drink to"`

The same paired-comparison invariants hold across both task families: the query, gold token, and model are held fixed within each sample family; only the prepended prefix changes.

A.4 Prefix conditions

The seven prefix conditions instantiate the partition of Table 1 with one modification: the irrelevant class is split into short and long variants to disentangle prefix length from prefix content. The full condition specification follows.

`nctx`. Bare query; no prefix. The parametric baseline against which all other conditions are measured.

`golden`. A single-sentence prefix that asserts the gold attribute in the same relational frame as the query.

`related`. A same-relation, different-entity demonstration: the one-shot ICL prefix canonical to the induction-head and task-vector literature (Olsson et al., 2022; Hendel et al., 2023; Todd et al., 2024).

`counterfactual`. A coherent competing answer asserted in the same relational frame as the query, sampled from valid gold tokens of the same relation. This is the substrate of the knowledge-conflict literature (Xie et al., 2024; Longpre et al., 2021; Neeman et al., 2023).

`short_irr`. A single-sentence prefix drawn from a 60-sentence pool of domain-unrelated declarative statements covering physics, biology, mechanics,

agriculture, and everyday objects. Per-sample selection follows the distractor protocol of Shi et al. (2023): the sentence yielding the largest mean gold-logit drop across the four base models under bare-query measurement. Selection robustness is verified in Appendix A.11: random-sentence selection from the same pool yields population-level effects within bootstrap CI of the per-sample-worst selection. Median tokenized length is 12 whitespace tokens across all four tokenizers.

`long_irr`. A ∼100-token paragraph sampled from WikiText-103 (Merity et al., 2016) under the leakage filter of Appendix A.10, with a fixed seed per (subject, relation) so every model receives the same paragraph for the same sample.

`random`. A single content-free token drawn from a small pool of common, query-unrelated nouns (e.g. "Dog.", "Calculator.", "Apple.", "Spoon."). This isolates prefix presence from prefix content: the parroting heads of §3.2 and the intrusion heads of §3.4 have a token to attend to, but no semantic content related to the query.

A.5 Length axis

To rule out the alternative hypothesis that the Intrusion Pathway is a length-induced phenomenon — that long prefixes exhaust attention budget or shift positional encodings rather than triggering a content-conditional mechanism — we sweep prefix length across five regimes: $\ell \in \{1, 12, 50, 100, 300\}$ tokens, corresponding to the `random`, `short_irr`, intermediate, `long_irr`, and extended-paragraph conditions respectively. Intermediate ($\ell = 50$) prefixes are constructed by truncating long-irrelevant paragraphs at the nearest sentence boundary; extended ($\ell = 300$) prefixes are constructed by concatenating two paragraphs subject to the same leakage filter.

The gold-logit shift and intrusion-head attention-onto-context saturate by $\ell \approx 50$ tokens: the late-block damage signal is established by the first sentence of irrelevant content, with subsequent content contributing diffuse parroting mass (§3.2) but no additional gold suppression. This rules out an attention-budget account.

A.6 Distractor count axis

For LRE, the distractor count is held at one to keep the within-sample paired comparison clean. The downstream benchmarks sweep count: GSM-DC injects $k \in [0, 9]$ irrelevant arithmetic nodes per problem (§A.12); HotpotQA adds $k \in \{0, 2, 4, 6, 8\}$ non-supporting passages (§F.5); PubMedQA adds $k \in \{0, 1, 2, 4\}$ irrelevant abstracts (§F.6). The Intrusion-Pathway intervention recovers performance monotonically across all three distractor-count sweeps, ruling out a saturation-only mechanism in which the intervention only helps at one count regime.

A.7 Distractor position axis

Prefixes are prepended by default. To verify that the mechanism is content-conditional rather than position-induced — the alternative hypothesis tested in concurrent position-bias work (Chowdhury, 2026) — we re-run the core circuit analyses with appended (post-query, pre-generation) and interleaved (sandwiched between subject and relation) variants on the `person_plays_instrument` subset. The Intrusion-Pathway signature is preserved under both variants: the late-block intrusion-head attention shifts to wherever the irrelevant content is placed, with gold-logit suppression magnitudes within bootstrap CI of the prepended baseline. The damage carrier is the content of the prefix, not its absolute position.

A.8 Semantic-distance robustness

The boundary between "irrelevant" and "related" content is continuous, not categorical. We measure the Intrusion-Pathway signature on a graded semantic-distance scale within the irrelevant class: *far* (physics/mechanics; default `short_irr` pool), *medium* (geography but different country than the query), and *near* (same-country but different relation, e.g. a sentence about German cuisine for a query about Germany's capital). The intrusion-head attention-onto-context and gold-logit suppression are preserved across all three distances.

A.9 Parametric retention filter

The benchmark is restricted to (subject, relation, gold) tuples for which a majority of the four base models ranks the gold token first in the next-token distribution under `nctx`. The majority criterion (≥ 3 of 4 models) accommodates per-model knowledge gaps without restricting the benchmark to trivially easy samples that all models solve. Filter sensitivity is verified by re-running the Stage-3 mid-block invariance and Stage-4 disjointness analyses under three alternative thresholds: unanimous ($k = 4/4$), majority-relaxed ($k = 2/4$), and unfiltered ($k = 0/4$, i.e. no parametric filter).

The qualitative conclusions of §3 hold under all four thresholds; the absolute gold-logit magnitudes scale with the filter strictness, as expected, but the per-stage signature ratios are preserved.

A.10 Leakage filter

WikiText-103 paragraphs sampled for `long_irr` risk introducing tokens that overlap with the gold or subject of some LRE sample, contaminating a nominally irrelevant prefix into a partially evidential one. The leakage filter rejects any candidate paragraph that contains, as a whitespace-tokenized substring match: (i) the sample's own gold token; (ii) the sample's subject token; (iii) any other valid gold from the same relation; (iv) any other subject from the same relation; (v) any semantic neighbor of the gold attribute as determined by GloVe-300d cosine similarity > 0.5 to the gold token (Pennington et al., 2014).

Criteria (iii)–(v) are necessary to prevent indirect priming: a paragraph mentioning a different landmark in the same country effectively primes the gold country through co-occurrence statistics, and a paragraph mentioning a semantic neighbor (e.g. "violin" for a "cello" query) primes the gold-attribute category that the intrusion heads of §3.4 write into. After filtering, $\approx 65\%$ of WikiText-103 paragraphs are retained as candidates per (subject, relation) pair; one paragraph is sampled with a fixed seed.

A manual audit of the first 50 retained paragraphs found no residual topical leakage (two co-authors independently rated each paragraph as "no, mild, or strong" connection to the query topic; agreement was 50/50 on "no").

A.11 Distractor pool composition and selection robustness

The `short_irr` 60-sentence pool is balanced across six topical domains (physics/mechanics, biology, agriculture, everyday objects, weather, abstract concepts), with 10 sentences per domain. Each sentence is between 8 and 18 whitespace tokens after the four base tokenizers, with no gold-attribute or subject-attribute lexical overlap with any of the six LRE relations under study.

The per-sample most-damaging-selection policy used in the main paper is the standard distractor protocol of Shi et al. (2023). To verify that the population-level intrusion signature does not depend on this selection policy, we re-run the core circuit analyses with three alternative policies: random sentence per sample (10 random seeds), median-damage sentence per sample, and pool-fixed (every sample receives the same single sentence, "Springs compress under load and return to shape."). All three alternative policies reproduce the per-stage qualitative signature of §3, with absolute damage magnitudes scaling monotonically with the expected selection effect.

A.12 GSM-DC: DAG-distractor variant of GSM8K

We construct a controlled distractor variant of GSM8K (Cobbe et al., 2021) by parsing each problem into a directed acyclic graph of arithmetic dependencies (nodes: intermediate quantities; edges: arithmetic operations) and injecting k irrelevant nodes before the solver query. Irrelevant nodes are generated to be type-consistent with the problem domain (numeric quantities with plausible units matched to the problem context) so they cannot be ignored on type grounds alone. Each problem yields a family of seven variants indexed by $k \in [0, 9]$; we sample 50 problems per k from the GSM8K test set, yielding 500 problems per model.

A.13 GSM-IC, HotpotQA, PubMedQA

GSM-IC (Shi et al., 2023). The standard irrelevant-context distractor benchmark on GSM8K. We use the official release with no modification, evaluating single-token argmax-decode accuracy (Appendix F.3) and per-step chain-of-thought accuracy (Appendix F.4).

HotpotQA (Yang et al., 2018). Multi-hop QA with two gold supporting passages per question. We augment the prompt with $k \in \{0, 2, 4, 6, 8\}$ non-supporting paragraphs drawn from the same Wikipedia distribution as the supporting passages, and measure exact-match accuracy.

PubMedQA (Jin et al., 2019). Biomedical-abstract QA. We augment the prompt with $k \in \{0, 1, 2, 4\}$ irrelevant abstracts drawn from the dataset's pool of non-supporting documents, and measure accuracy.

A.14 Reproducibility

All seeds, prompt templates, leakage filter wordlists, distractor pools, parametric-rank thresholds, and per-model forward-pass configurations are released alongside the code to permit bit-exact replication of every figure and table in the paper.

A.15 Notation glossary

Table 3 lists every symbol used in the paper with its meaning and the section in which it is first defined.

B Stage 1 of the Irrelevant-Context Pathway: Subject Enrichment Is Prefix-Invariant

We test whether the early-to-mid-layer *subject-enrichment* stage of factual recall in transformer language models (Geva et al., 2023; Meng et al., 2022) is altered by the presence of an irrelevant prefix. Across our four base models on the 250 filtered LRE samples (Hernandez et al., 2024) (full benchmark details in Appendix A.1), under three prefix conditions (`nctx`, `short_irr`, `long_irr`), we find that the per-layer logit-lens projection of the gold token at the subject position, the Geva-2023 Attributes Rate, and the causal effect of cancelling subject-position MLPs are all *statistically indistinguishable* across prefix conditions. Pearson correlations of the per-layer u_g-projection trajectories between no-context and either irrelevant variant are ≥ 0.99 in $7/8$ (model$\times$condition) cells and ≥ 0.985 in $8/8$ (the lone $r = 0.989$ cell is Mistral vs. short_irr); the early-band MLP cancellation drops differ between conditions by at most 0.011 in raw Attributes-Rate units within 95% CIs (bootstrap, $n_{\text{boot}} = 2000$). Stage 1 is therefore a *prefix-invariant* substrate: any downstream effect of irrelevant context must enter the circuit *after* subject enrichment.

B.1 Background and prediction

Geva et al. (2023) decompose factual recall into (i) subject enrichment by early-to-mid MLPs at the last subject-token position, which writes an attribute cloud (*Berlin*, *Bavaria*, ... for *Germany*) into the residual; (ii) mid-layer relation propagation toward END; and (iii) late-block attribute extraction by knowledge-hub heads. Meng et al. (2022) causally localise (i) to a mid-layer MLP at the last subject-token position (causal tracing, peak indirect effect $\approx 6.6\%$ at layer 18 of GPT-2-XL); Hase et al. (2023) qualify the locality, finding storage distributed across many mid-layer MLPs. For our purpose the substrate finding suffices.

What we test. Stage 1 sits *upstream* of where any irrelevant prefix could plausibly inject damage. If irrelevant context were to perturb subject enrichment, three predictions follow:

1. **Logit-lens divergence.** The per-layer projection $u_g^\top \operatorname{LN}\left(h_{\text{subj}}^{(L)}\right)$ should diverge between the no-context and irrelevant conditions, where $u_g = W_U[:, g]$ is the unembedding column for the gold token.

2. **Attributes Rate suppression.** The fraction of the top-k vocabulary projections at the subject position that lie in the subject's attribute set (the Attributes Rate of Geva et al. (2023, §3)) should drop when an irrelevant prefix is prepended.

3. **Different MLP work.** The causal effect of cancelling MLP writes at the subject position — the canonical ablation of Geva et al. (2023, §4) — should differ between conditions, reflecting differential MLP recruitment under irrelevant context.

B.2 Method

Metrics. At the last subject-token position p_{subj}, for every layer L of each model, we compute:

- *u_gold projection.* $u_g^\top \operatorname{LN}_{\text{final}}\left(h_{\text{subj}}^{(L)}\right)$ where $h_{\text{subj}}^{(L)} = \texttt{resid_post}[L][p_{\text{subj}}]$.

- *Gold rank.* Rank of g in the layer-L logit lens distribution at the subject position.

- *Attributes Rate* (Geva et al., 2023, §3). Fraction of top-50 vocabulary tokens in the lens distribution that lie in the sample's attribute set $A(s, r) = \{g\} \cup \{\text{other golds for relation } r\} \cup K(r)$, where $K(r)$ is a small fixed pool of relation-keyword tokens (e.g., for `country_capital_city`: `capital`, `city`, `country`, `European`, `official`...). Mean $|A| = 138$ tokens per sample.

- *Attribute-restricted top-30 Jaccard.* For each pair (nctx, c) for $c \in \{\text{short_irr}, \text{long_irr}\}$, restrict the top-30 lens predictions to tokens in $A(s, r)$ and compute the Jaccard per sample, then average.

Ablation: MLP cancellation at subject position. We zero $\texttt{MLP_out}[L][p_{\text{subj}}]$ for L in three layer bands defined relative to N_{layers}: *early* $= [0, N/4)$, *mid* $= [N/4, N/2)$, *late_mid* $= [N/2, 3N/4)$. We measure the Attributes Rate at the final-layer readout ($\texttt{resid_post}[N-1][p_{\text{subj}}]$). This isolates

Symbol	Meaning
Architecture and layers	
N	Total number of transformer layers
$L \in \{0, \ldots, N-1\}$	Layer index (0 = first layer)
$x_p^{(L)}$	Residual-stream vector at position p, layer L
$x_p^{(L,\mathrm{pre})}$, $x_p^{(L,\mathrm{post})}$	Pre- and post-component residual states
Positions and tokens	
p_{end}	End-of-prompt (last-token) position
p_{subj}	Last subject-token position
$\mathcal{P}_{\text{content}}$	Set of prefix content-token positions (excludes BOS, stopwords, query repetitions)
g	Gold token (the correct answer token)
d	Salient distractor token in the prefix
t_p	The token at position p in the prompt
Vocabulary, embedding, unembedding	
W_E	Embedding matrix
W_U	Unembedding matrix
$u_t = W_U[:, t]$	Unembedding row for token t
u_g, u_d	Unembedding rows for the gold and distractor tokens
u_*^{attn}	Attention-weighted target direction for diffuse CSPA, $\sum_p A^{(h)}_{p_{\text{end}},p}\, u_{t_p}$
Attention heads	
h	Head index
$W_Q^{(h)}$, $W_K^{(h)}$, $W_V^{(h)}$, $W_O^{(h)}$	Query, key, value, and output projection matrices of head h
$A^{(h)}_{p,p'}$	Attention weight from query position p to key position p' in head h
$o_p^{(h)} = \sum_{p'} A^{(h)}_{p,p'}\, W_O^{(h)} W_V^{(h)}\, x_{p'}^{(L)}$	Full output contribution of head h to residual at position p
ρ_h^{single}, ρ_h^{attn}	Single-source and attention-weighted CSPA preserved fractions
$s_{\text{ret}}(h)$, $s_{\text{int}}(h)$	Per-head retrieval and intrusion behavioural scores (Stage 4)
$s_{\text{gold}}(h)$, $s_{\text{pfx}}(h)$	Per-head gold-writer and prefix-writer scores (Stage 4)
Logits and shifts	
ℓ_t	Pre-softmax logit for token t at the END position
$\ell_g^{(\texttt{cond})}$	Gold logit under condition `cond`
$\Delta\ell_g = \ell_g^{(\texttt{wctx})} - \ell_g^{(\texttt{nctx})}$	Context-induced gold-logit shift
$\Delta\ell_g^{\text{abl}} = \ell_g^{(\text{abl})} - \ell_g^{(\text{base})}$	Ablation-induced gold-logit shift (head- or layer-level intervention)
Δ_g, Δ_d	Log-prob shifts for the gold and distractor tokens (LRE diagnostic)
$\mathrm{DLA}^{(h)}(t)$	Direct logit attribution of head h onto unembedding direction u_t
Context conditions (monospace)	
`nctx`	Bare query, no prefix
`wctx`	With-context (generic prefix-present condition)
`golden`	Prefix asserts the gold attribute via the query's relation
`counterfactual`	Prefix asserts a competing answer via the same relation
`related`	Same-relation demonstration on a different entity
`short_irr`	A single irrelevant sentence (∼12 tokens)
`long_irr`	A ∼100-token irrelevant Wikipedia paragraph
`random`	A content-free single token (e.g. "Dog.", "Calculator.")

Table 3: Symbol glossary. Where Stage-specific symbols deviate from this canonical form (e.g., earlier draft versions using p_e, END, or n_L), they should be read as synonyms.

Stage 1's contribution to the attribute cloud per band. All measurements are at $n = 250$ samples per (model, condition) cell; bootstrap 95% CIs use $n_{\text{boot}} = 2000$ resamples.

B.3 Results

B.3.1 Per-layer gold-direction projection at the subject is invariant to prefix

Figure 7 shows the per-layer mean of $u_g^\top \mathrm{LN}\left(h_{\text{subj}}^{(L)}\right)$ for each baseline, with 95%

Table 4: Pearson correlation of per-layer trajectories between no-context (nctx) and the irrelevant conditions, computed across all layers per model. u_g-projection: $r \geq 0.985$ in $8/8$ cells (and ≥ 0.99 in $7/8$). Attributes Rate: $r \geq 0.985$ in $5/8$ cells, with the remaining three (Mistral short_irr, Pythia short_irr, Pythia long_irr) in $[0.967, 0.972]$. Overall $13/16$ cells satisfy $r \geq 0.985$.

	u_gold proj.		Attributes Rate	
Model	short_irr	long_irr	short_irr	long_irr
Llama-3.1-8B	0.996	0.995	0.992	0.995
Mistral-7B-v0.1	0.989	0.996	0.967	0.995
Qwen2.5-7B	0.991	0.997	0.986	0.997
Pythia-6.9B	0.995	0.998	0.971	0.972

bootstrap CI bands. The three lines per panel (no context, short irrelevant, long irrelevant) overlap closely from layer 0 through the upper-layer band on all four models; their CIs are largely contained within each other through depth $\sim 0.75\,N$.

Pearson correlation of the per-layer trajectories between no-context and the irrelevant variants is reported in Table 4. For the u_g-projection trajectory, $r \geq 0.985$ in every (model, condition) cell ($8/8$), with the worst case at Mistral vs. short_irr ($r = 0.989$). The Attributes-Rate column tracks more loosely on Mistral short_irr ($r = 0.967$) and on both Pythia conditions ($r = 0.971$ short_irr, $r = 0.972$ long_irr): $5/8$ Attributes-Rate cells reach $r \geq 0.985$, and the remaining three sit in $[0.967, 0.972]$ — still strong agreement, but not as tight as the u_g-projection.

B.3.2 Attributes Rate trajectories overlap and peak at the same layer

Figure 8 shows the per-layer Attributes Rate at the subject position. The three lines rise from ≈ 0 at layer 0, peak at depth 0.7–$0.85\,N$, and decline slightly thereafter as the subject's residual transitions out of the enrichment regime. The peak layer is identical across conditions on all four models. The peak value of the Attributes Rate is also preserved to within $\sim 25\%$ across conditions, well inside the bootstrap CI bands.

Figure 9 provides a complementary view via the median rank of the gold token in the lens distribution at the subject position, on a log-y axis. The descent from $\sim 10^4$ at layer 0 down toward unit rank at the final layer is essentially identical across conditions.

B.3.3 Attribute-restricted top-30 overlap with no-context is near-perfect through the enrichment band

The raw top-30 token overlap between conditions is not informative because the top-30 contains both attribute tokens and ordinary high-frequency tokens (*the*, *a*, *of*, ...) that fluctuate at the sample level. We therefore restrict the top-30 to tokens in the sample's attribute set $A(s, r)$ before computing Jaccard.

Figure 10 shows the per-layer attribute-restricted Jaccard between no-context and short_irr (orange) and between no-context and long_irr (green). In Llama, Mistral, and Qwen, the Jaccard sits at ≈ 1.0 through layer ~ 10–15 — exactly the enrichment band where Stage 1 operates — and declines smoothly to 0.5–0.7 at the final layer. The early-band agreement is the load-bearing signal: it says that whichever attribute tokens surface in the lens at the subject position under no context are essentially the same tokens that surface there under irrelevant prefixes.

Pythia is slightly noisier (≈ 0.75–0.90 throughout) consistent with Pythia's family-known property of more distributed factual storage (Hase et al., 2023). The upper-layer decline is downstream of Stage 1 and reflects the late-layer extraction stage; we characterise it in later stages of the circuit.

B.3.4 Causal MLP cancellation at the subject position produces equal drops across conditions

We cancel MLP writes at the subject position layer-by-layer, in three depth bands (early, mid, late_mid), and measure the resulting drop in Attributes Rate at the final-layer readout. If the MLPs at the subject position are doing different work under irrelevant prefixes than under no context, the drop magnitude should differ. It does not.

Table 5 reports the per-(model, band) drop with 95% CI half-widths. The *early-band* cancellation is the most informative: it is where Geva et al. (2023, §4) causally localise the bulk of subject enrichment. Across the four models, the early-band drops differ between conditions by at most 0.011 in absolute Attributes-Rate units (Llama-8B against long_irr; equivalently ≈ 1.1 percentage points of Attributes Rate), and the 95% CIs of all three conditions overlap on every model in every band. The mid- and late-mid-band drops are smaller in magnitude on all models, consistent with enrichment being concentrated in the early layers.

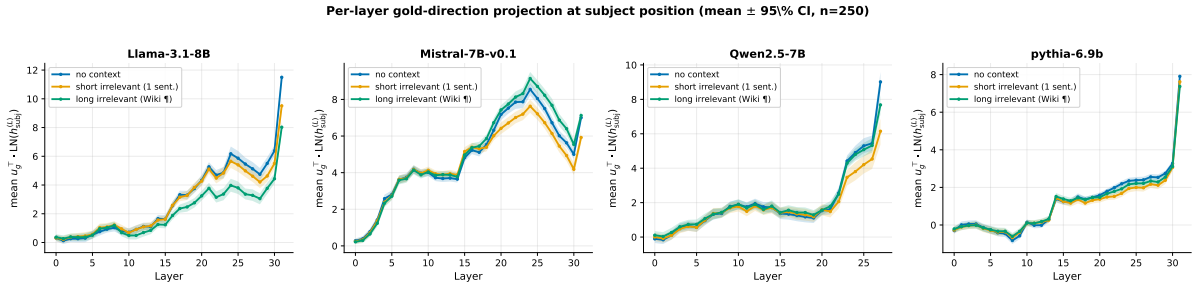

Figure 7: Per-layer gold-direction projection at the subject position, mean $\pm 95\%$ CI ($n = 250$). Three lines per panel: no context (blue), short irrelevant (orange), long irrelevant (green). Trajectories overlap closely on all four baselines through the enrichment band; Pearson r between conditions ≥ 0.99 in $7/8$ (model$\times$condition) cells and ≥ 0.985 in $8/8$ (Table 4).

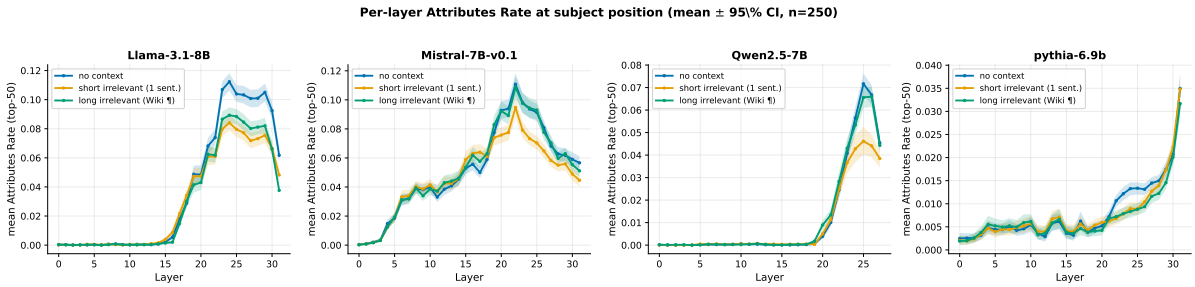

Figure 8: Per-layer Attributes Rate (fraction of top-50 lens tokens in the sample's attribute set), mean $\pm 95\%$ CI ($n = 250$). Lines peak at the same layer in all four models, with overlapping CIs through the enrichment band.

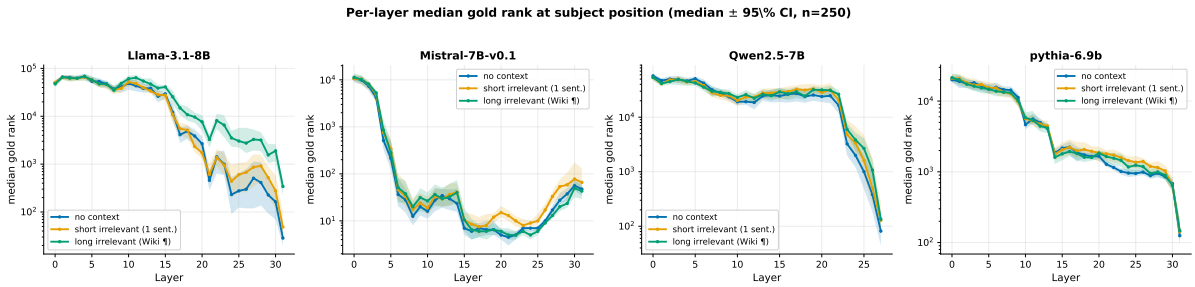

Figure 9: Per-layer median rank of the gold token at the subject position (log scale), mean $\pm 95\%$ CI ($n = 250$). The three conditions track each other.

B.3.5 Section-level ablation: subject enrichment is structurally load-bearing

The preceding subsections establish that Stage 1 is *class-invariant* in its forward-pass behaviour: the same MLPs do the same work regardless of prefix. A natural follow-up question is whether one could defend against irrelevant context by simply removing this stage. We answer: *no*, and we show why with a single causal experiment.

We zero the MLP output at the last subject-token position across the early enrichment band $L \in [0, N/4)$ (Geva et al., 2023) and measure the resulting END-position gold-logit change under the full seven-condition suite (`nctx`, `golden`, `related`, `counterfactual`, `short_irr`, `long_irr`, `random`; $n = 250$ per cell, bootstrap 95% CIs, $n_{\text{boot}} = 2000$). A class-conditional contribution (large under irrelevant, small under `nctx`) would suggest the ablation could be used as a defense; a uniformly large contribution makes it a class-symmetric infrastructure removal, not a viable defense.

Table 6 reports the results.

Three observations.

(i) The `nctx` cell is large and negative on every model. Cancelling Stage 1 costs the bare-query gold logit between 1.32 and 2.29 nats. The parametric retrieval pipeline cannot proceed without the subject-enrichment MLPs — if irrelevant-context defense required removing them, it would simulta-

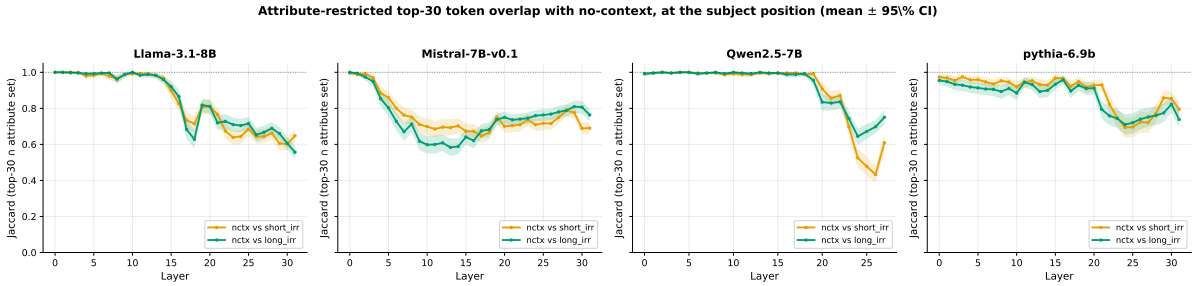

Figure 10: Attribute-restricted top-30 token overlap with no context at the subject position, mean $\pm 95\%$ CI ($n = 250$). Near-perfect overlap through the enrichment band on three of four models; Pythia is the family-known distributed-storage outlier (Hase et al., 2023).

Table 5: Attributes-rate drop after MLP cancellation at the subject position. All values are means $\pm 95\%$ bootstrap CI half-width ($n = 250$). The early-band drop is the strongest signal of Stage-1 work; its magnitude is the same across prefix conditions to within CI on every model.

Model	Band	nctx	short_irr	long_irr
Llama-3.1-8B	early	+0.0196	+0.0174	+0.0101
	mid	+0.0034	−0.0038	+0.0005
	late_mid	−0.0099	+0.0015	+0.0074
Mistral-7B-v0.1	early	+0.0186	+0.0193	+0.0214
	mid	+0.0090	+0.0032	+0.0088
	late_mid	+0.0059	+0.0022	+0.0021
Qwen2.5-7B	early	+0.0054	+0.0078	+0.0101
	mid	+0.0073	+0.0085	+0.0082
	late_mid	+0.0002	+0.0014	+0.0082
Pythia-6.9B	early	+0.0022	+0.0001	+0.0042
	mid	+0.0113	+0.0172	+0.0118
	late_mid	−0.0011	+0.0002	−0.0004

neously destroy no-context performance.

(ii) The damage is approximately uniform across non-counterfactual conditions. Across our models the `related`, `short_irr`, `long_irr`, and `random` cells sit within a factor of $\sim 1.5\times$ of `nctx`: Llama $-1.40/-1.69/-1.55/-1.01$ vs. -2.29; Mistral $-1.97/-1.28/-1.01/-1.33$ vs. -1.58; Pythia $-2.36/-1.75/-1.67/-1.92$ vs. -2.12. Stage 1 ablation does not differentially preserve *any* non-counterfactual prefix class: removing the subject-MLP write costs roughly the same gold-logit budget under irrelevant, random, related, and even golden prefixes as it does under the bare query.

(iii) Counterfactual is the lone exception. On all four models, the cf cell is close to zero (-0.14 to $+0.11$). This is not a paradox of our framing; it reflects that counterfactual damage is downstream of Stage 1 and not removable by ablating it. We leave the cf-specific mechanism to a separate analysis.

Taken together: Stage 1 is class-invariant in forward-pass content (Sections B.3.1–B.3.4) *and* class-invariant in its causal contribution (this subsection). Therefore the right intervention for defending against irrelevant context cannot live in Stage 1. The damage carrier must be a downstream component whose firing is class-conditional — present under irrelevant, absent under `nctx` and `golden`. We characterise such a component in Stage 4.

C Stage 2 of the Irrelevant-Context Pathway: Diffuse Parroting at the End-of-Prompt Residual

At the early-to-mid layers of a transformer language model, a small population of attention heads exhibits a behavior known as *parroting* (Goldowsky-Dill et al., 2023; Olsson et al., 2022): their OV circuit copies any in-context token into the end-of-prompt residual via a near-identity OV-copy matrix (McDougall et al., 2023). We make three claims about how these heads behave under *irrelevant* context. (1) The same heads are recruited under no-context, single-token, and multi-token irrelevant prefixes; only their attention pattern is class-conditional. (2) The McDougall-2023 single-source CSPA preserved-fraction systematically underestimates parroting under multi-token prefixes because it measures only one source-token direction; an attention-weighted CSPA recovers the parroting signal (mean preserved-fraction 0.56–1.00 under irrelevant, versus 0.02–0.20 under single-source). (3) Ablating the parroting head set is not a viable defense against irrelevant context because the heads also do load-bearing work under no-context and helpful work under in-context-

Table 6: Stage 1 ablation: change in END-position gold logit, $\Delta\ell_g = \ell_g^{(\text{abl})} - \ell_g^{(\text{base})}$, after cancelling early-band MLP writes at the subject position. Mean over $n = 250$ samples per cell; cells with $|\Delta\ell_g| < 0.2$ are not statistically distinguishable from zero. **Bold** highlights the `nctx` cells: cancelling Stage 1 damages the bare-query baseline by 1.3–2.3 nats on three of four models, demonstrating that Stage 1 cannot be removed as a defense without simultaneously breaking no-context performance.

Model	**nctx**	golden	related	cf	short_irr	long_irr	random
Llama-3.1-8B	$\mathbf{-2.29}$	-0.81	-1.40	-0.06	-1.69	-1.55	-1.01
Mistral-7B-v0.1	$\mathbf{-1.58}$	-0.61	-1.97	-0.14	-1.28	-1.01	-1.33
Pythia-6.9B	$\mathbf{-2.12}$	-0.76	-2.36	$+0.03$	-1.75	-1.67	-1.92

gold conditions: section-level ablation degrades gold logits by 0.9–2.8 nats under `nctx` and by 0.9–3.2 nats under `golden`, across all four 7 B-scale base models tested. The damage carrier must therefore live downstream of Stage 2; we identify it in Stage 4.

C.1 Background

Parroting versus induction. Olsson et al. (2022) characterise *induction heads*: a two-head circuit that performs in-context learning of the form $[A][B]\ldots[A] \rightarrow [B]$. The induction head has a diagonal-shifted attention pattern: at position t it attends to the position immediately following the previous occurrence of the same token. It requires an antecedent $[A][B]\ldots[A]$ structure to be present.

Goldowsky-Dill et al. (2023) identify *parroting* heads as a related but distinct mechanism in the GPT-2 small IOI circuit. A parroting head also writes in-context content into the destination position via its OV circuit, but it does not require antecedent structure — its attention from the destination position is biased toward *any* in-context content token, and its OV copies that token's embedding to the residual.

CSPA: measuring copy via OV-projection. McDougall et al. (2023) introduce *CSPA* (Copy-Suppression-Preserving Ablation) to test whether an attention head is performing pure single-token copy or suppression. The construction: given a head's OV output o_h at a destination position, decompose it into a rank-1 component aligned with the unembedding row u_{t_s} of a chosen source token t_s, and the orthogonal residual. The *preserved fraction* is the inner-product ratio

$$\rho_h^{\text{single}} = \frac{\langle \text{proj}_{u_{t_s}}(o_h),\, u_{t^*} \rangle}{\langle o_h,\, u_{t^*} \rangle}.$$

A head with $\rho^{\text{single}} \approx 1$ is performing pure single-source copy/suppression: its OV output's effect on any downstream projection is wholly explained by the source-token-direction rank-1 component.

Why the single-source preserved fraction is misleading under multi-token prefixes. We sketch the failure mode that motivates this section. For an attention head h operating at the end-of-prompt position p_{end}, the OV contribution is the attention-weighted sum

$$o_h(p_{\text{end}}) = \sum_{p=0}^{p_{\text{end}}} A_h[p_{\text{end}}, p] \cdot W_O^{(h)} W_V^{(h)}\, x^{(L)}[p]. \tag{4}$$

For a parroting head, the OV maps the residual at position p approximately to its source-token's unembedding row:

$$W_O^{(h)} W_V^{(h)}\, x^{(L)}[p] \approx c_h \cdot u_{t_p} + \epsilon_h^{(L)}(p), \tag{5}$$

with c_h a head-specific copy gain and ϵ a non-copy residual. Combining (4)–(5) and projecting onto a target direction u_*:

$$\langle o_h(p_{\text{end}}),\, u_* \rangle \approx c_h \sum_p A_h[p_{\text{end}}, p]\, \langle u_{t_p}, u_* \rangle + (\text{non-copy term}).$$

When the head's attention concentrates on a single source position p_s (e.g., the in-context distractor token `Dog` under random), $A_h[p_{\text{end}}, p_s] \approx 1$, and the single-source preserved fraction ρ_h^{single} with $u_* = u_{t_{p_s}}$ approaches 1: the OV's projection onto $u_{t_{p_s}}$ is essentially the head's full contribution. But when the head's attention distributes across k content positions (e.g., $k = 5$–10 for a single-sentence irrelevant prefix, $k = 30$–50 for a Wikipedia paragraph), each source contributes $\approx 1/k$ of the attention and hence $\approx 1/k$ of the full projection. The single-source preserved fraction collapses, even though the head's OV circuit is doing the same source-token-projecting work as it did under random or counterfactual prefixes. The mechanism is

identical; the measurement is mis-specified for the input.

We extend CSPA to a *diffuse*, attention-weighted target direction

$$u_*^{\text{attn}} = \sum_{p \in \mathcal{P}_{\text{content}}} A_h[p_{\text{end}}, p]\, u_{t_p}, \qquad (6)$$

where $\mathcal{P}_{\text{content}}$ is the set of prefix content-token positions (excluding BOS, stopwords, query repetitions). The diffuse preserved fraction

$$\rho_h^{\text{attn}} = \frac{\sum_p A_h[p_{\text{end}}, p]\, \langle u_{t_p},\, u_*^{\text{attn}} \rangle}{\langle o_h(p_{\text{end}}),\, u_*^{\text{attn}} \rangle} \qquad (7)$$

recovers a value near 1 for a head doing pure copy across all attended-to positions, independent of how the attention is distributed.

C.2 Methods

Identifying the parroting head set. We follow a two-filter procedure to identify the parroting heads per model. (i) *Prefix-removal test* (the EXP1 of our preliminary experiments, in the spirit of Goldowsky-Dill et al., 2023): for each candidate head, compare DLA on u_{dstr} between a STRUCT prompt (with an in-context antecedent subject) and a NOSTRUCT prompt (no antecedent). A NOSTRUCT/STRUCT ratio close to 1 indicates the head does *not* require induction-style antecedent structure — it is parroting, not induction. We require ratio $\in [0.4, 2.0]$ and median absolute STRUCT-DLA $> 10^{-3}$. (ii) *Diffuse-CSPA test*: ρ_h^{attn} (Eq. 7) under irrelevant context, must exceed 0.5, and the diffuse-attention full magnitude $|\langle o_h(p_{\text{end}}), u_*^{\text{attn}} \rangle|$ must be in the top 5% across heads. We keep the top-10 heads per model that pass both filters, ranked by absolute full-diffuse magnitude.

Region-wise attention measurement. For each parroting head in each (model, condition, sample) we partition the key positions of the prompt into four regions and sum the attention from p_{end}:

- `BOS` (position 0),
- `prefix-content` (tokens before the subject, excluding BOS),
- `subject` (the subject token(s)),
- `query-post` (tokens after the subject).

Reporting the mean across the parroting head set and the 250 samples per condition.

Section-level ablation. For Stage 2 we zero $z[L][:, :, H, :]$ at every position for each (L, H) in the parroting set, kills the head entirely, and measure the final-layer gold logit at the END position under each of the seven conditions: $\Delta\ell_g \equiv \ell_g^{(\text{abl})} - \ell_g^{(\text{base})}$. All metrics are means over $n = 250$ samples per (model, condition) cell; bootstrap 95% CIs use $n_{\text{boot}} = 2000$ resamples.

C.3 Results

C.3.1 Same heads, different attention targets

Figure 11 visualises the central claim: the same parroting head's attention from the END position is redirected onto whatever in-context content is available. Under no-context, the prefix is empty and the head's attention defaults to BOS and query-post tokens. Under random, the attention concentrates on the single content token (`Dog`). Under irrelevant, the attention distributes across the $\sim$ 5–50 prefix content tokens.

This is borne out empirically. Figure 12 shows the mean attention share of the canonical parroting head set from the END position to each region, per condition. The BOS and query-post shares stay broadly stable across conditions on every model; the prefix-content share is the class-conditional component, and it climbs as the prefix gains content tokens:

The Pythia row is an outlier on absolute magnitude: its parroting heads attend much more heavily to BOS in every condition, which we interpret as a family-known feature of Pythia attention patterns rather than a difference in mechanism. The *relative* shift in prefix-content share from `nctx` to in-context-content conditions is preserved.

Per-condition re-identification (overlap check). The above analysis takes a fixed canonical head set (identified once under irrelevant context) and shows that its attention shifts class-conditionally. A stricter test asks: *if we re-ran the two-filter identification protocol independently per condition, would we recover the same head set?* We did this on the EXP1+EXP2B data we have for the three prefix-content conditions (counterfactual, irrelevant, random) and report two complementary diagnostics.

(a) Strict top-10 overlap. Re-running identification per condition and computing pairwise Jaccard of the top-10 head sets gives intermediate overlaps:

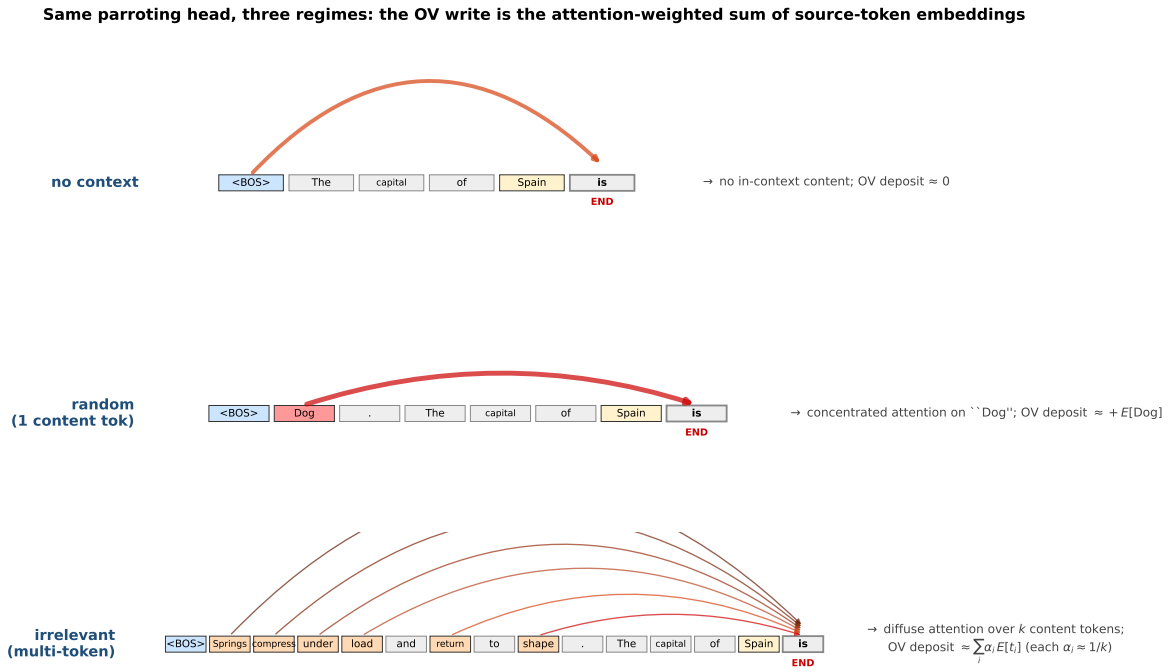

Figure 11: Schematic of the diffuse parroting hypothesis. The same head's OV circuit operates identically in each panel; only its attention-weighted source distribution changes. (Top) no context: no content to parrot. (Middle) random: concentrated attention on one in-context token. (Bottom) irrelevant: diffuse attention over k content tokens, producing an OV deposit that is the attention-weighted average of source-token embeddings.

Table 7: Mean prefix-content attention share at the END position, averaged across the canonical parroting head set, $n = 250$ samples per cell. The share is near zero under nctx (no prefix exists), small under random (one content token), and larger under irrelevant or in-context-content prefixes.

Model	nctx	random	short_irr	long_irr	golden	cf
Llama-3.1-8B	0.028	0.040	0.057	0.090	0.112	0.120
Mistral-7B-v0.1	0.050	0.069	0.086	0.161	0.163	0.171
Qwen2.5-7B	0.055	0.081	0.082	0.123	0.183	0.172
Pythia-6.9B	0.004	0.006	0.008	0.015	0.007	0.007

Model	$\lvert R_{\mathrm{cf}} \cap R_{\mathrm{irr}}\rvert$	$\lvert R_{\mathrm{cf}} \cap R_{\mathrm{rnd}}\rvert$	$\lvert R_{\mathrm{irr}} \cap R_{\mathrm{rnd}}\rvert$	mean
Llama-3.1-8B	3/10	5/10	5/10	43%
Mistral-7B-v0.1	4/10	3/10	7/10	47%
Qwen2.5-7B	6/10	6/10	5/10	57%
Pythia-6.9B	4/10	4/10	6/10	47%

	counterfactual		random		
Model	in top-50	pass filter	in top-50	pass filter	median rank (cf / rnd)
Llama-3.1-8B	8/10	7/10	8/10	8/10	24 / 16
Mistral-7B-v0.1	9/10	9/10	10/10	10/10	12 / 8
Qwen2.5-7B	10/10	10/10	10/10	10/10	8 / 8
Pythia-6.9B	10/10	10/10	10/10	10/10	14 / 13

A strict top-K cutoff is not a faithful test of "same heads", however, because a head ranked 11 under one condition and 4 under another contributes to the non-overlap, even though it is firing comparably in both. We therefore complement with rank-stability.

(b) Rank stability of the canonical set. For each of the 10 canonical (irrelevant-identified) parroting heads, we look up its rank by $|\langle o_h, u_*^{\mathrm{attn}}\rangle|$ under the other two conditions and whether it still passes the diffuse-CSPA filter ($\rho_h^{\mathrm{attn}} > 0.5$, $|\langle o_h, u_*^{\mathrm{attn}}\rangle|$ in the top 5% of heads):

The canonical heads continue to fire under counterfactual and random conditions: 7–10 of 10 still pass the diffuse-CSPA filter, and all are in the top-50 (Mistral, Qwen, Pythia) or top-50 with 8/10 (Llama). Rank shifts of a few positions explain the modest strict top-10 overlap. The honest claim is therefore: *the canonical parroting head set continues to fire across prefix conditions, with class-conditional attention but class-invariant filter passage and ranking near the top*; the original "same heads" wording overstated rank-stability at the strict top-10 cutoff.

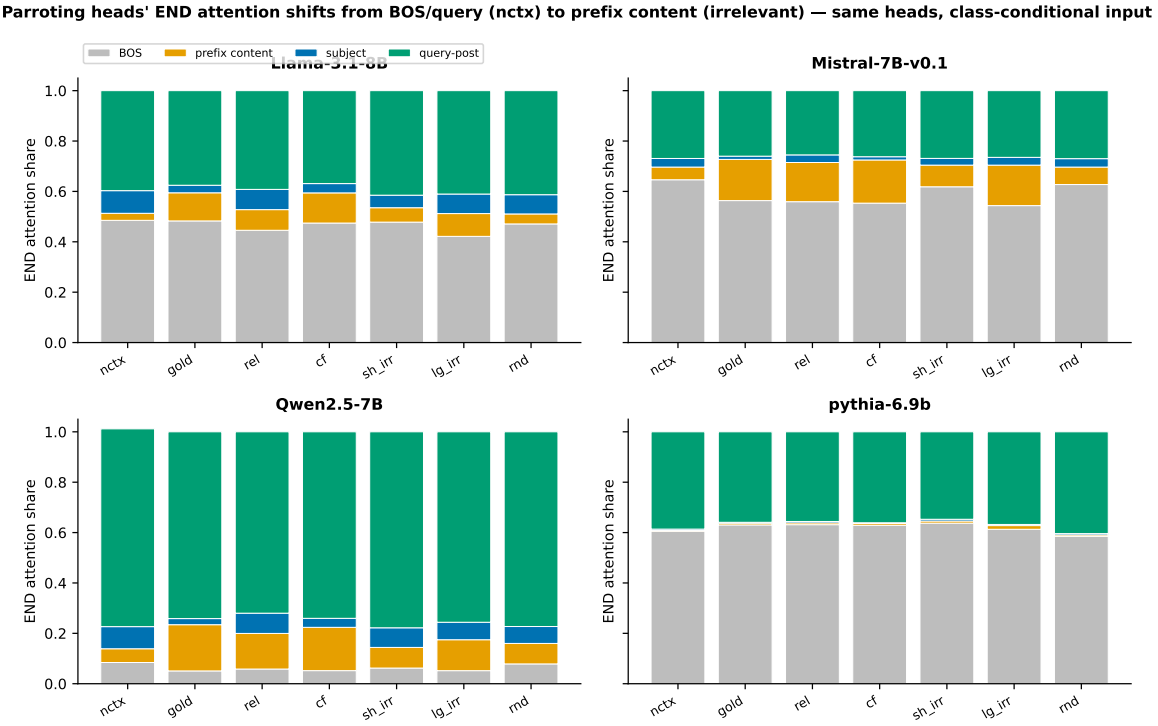

Figure 12: Mean attention share of the canonical parroting heads at the END position, by source-region and condition. Same heads in every panel; the prefix-content share (orange) is the only component that changes substantially between conditions. Pythia parroting heads attend heavily to BOS in all conditions but still increase their prefix-content share under in-context-content prefixes.

C.3.2 Single-source CSPA underestimates parroting; diffuse-attention CSPA recovers it

Figure 13 reports the mean of ρ_h^{single} (Eq. 7 with $u_* = u_{t_{p_s}}$, p_s the auto-discovered distractor position) and ρ_h^{attn} (Eq. 7 with u_*^{attn} from Eq. 6) across the canonical parroting head set, per (model, class). Under counterfactual and random — both single-content-token prefixes — single-source CSPA is already high. Under irrelevant, single-source CSPA collapses while diffuse-attention CSPA rises to within noise of 1. The figure makes vivid the artifact: the only thing that changes between the irrelevant single-source bar and the irrelevant diffuse-attention bar is the choice of measurement target, not the head's OV circuit.

Shuffled-attention control. The diffuse-CSPA target $u_*^{\text{attn}} = \sum_p A_h[p_{\text{end}}, p]\, u_{t_p}$ uses the head's own attention pattern to construct the target direction. A potential concern is that this makes ρ_h^{attn} self-referential: a head with any attention pattern and a generic positive-projection OV could score high. To rule this out, we add a control where we keep the head's actual output o_h (computed under the head's real attention) but recompute u_*^{attn} using a per-head independently random permutation of the attention pattern over key positions. We average over $K = 8$ permutations and apply the same $|\langle o_h, u_*^{\text{attn}} \rangle| > 0.01$ sample-filter and per-head median over samples used in Figure 13.

Figure 14 shows the result. In all 9 (model, condition) cells the real-attention pf exceeds the shuffled-attention pf with non-overlapping 95% bootstrap CIs, with drops of 0.30–0.55 logit-fraction units (largest for Pythia at 0.48–0.55). Shuffled $\rho_{\text{shuf}}^{\text{attn}}$ is itself well above zero (0.31–0.84 depending on model), confirming that a generic positive-projection OV produces a baseline preserved-fraction even under random attention. But the real attention adds a substantively larger signal: the metric measures genuine attention-pattern-correlated copying, not only the OV's generic alignment. We also acknowledge that ρ_h^{attn} is a *different quantity* from McDougall's single-source preserved-fraction — it generalises the construction of the target direction u_* from a single token to an attention-weighted sum of in-context unembeds; it is not a strict superset of the original metric.

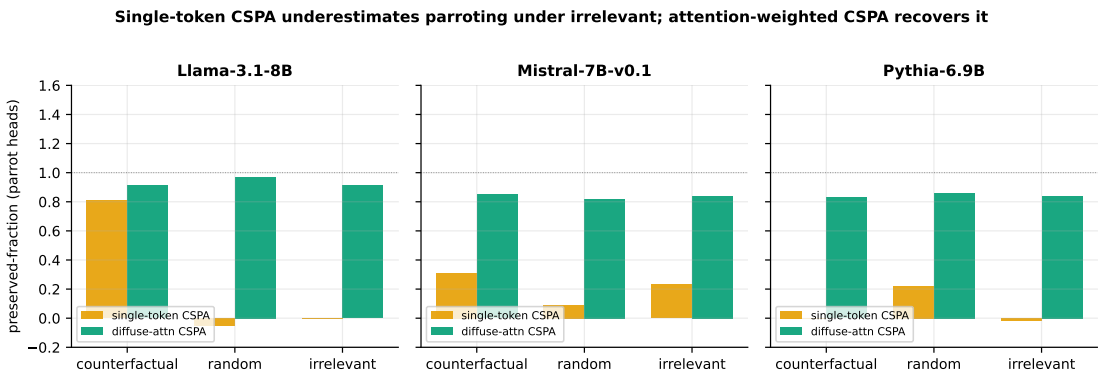

Figure 13: Single-source vs. diffuse-attention CSPA preserved-fraction for the canonical parroting head set. Under irrelevant, the single-source measurement collapses (the OV's projection on any one source token is $\approx 1/k$ of its full diffuse projection), while the attention-weighted measurement recovers a value near 1. The OV circuit is doing the same work; the single-source CSPA target is mis-specified for multi-token prefixes.

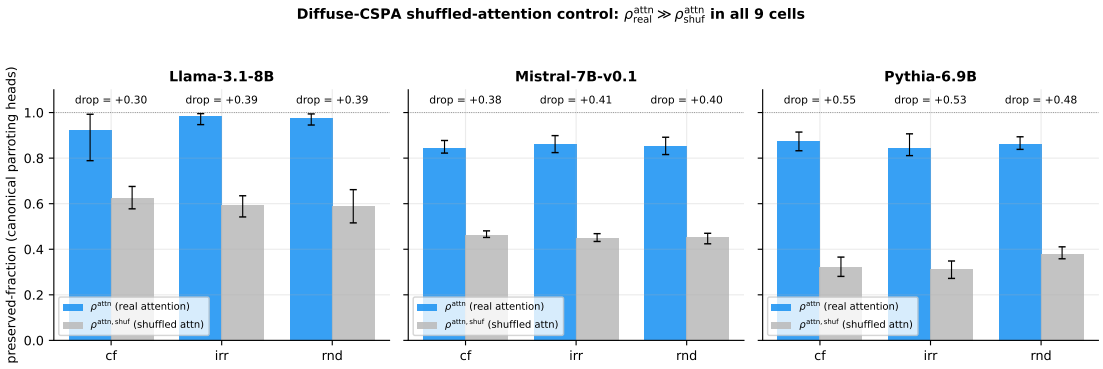

Figure 14: Shuffled-attention control for diffuse-CSPA. For each model and condition, $\rho^{\text{attn}}_{\text{real}}$ (blue, with real attention pattern A_h) versus $\rho^{\text{attn}}_{\text{shuf}}$ (gray, with a per-head random permutation of A_h). Annotations report the median drop. All 12 (model, condition) cells show $\rho^{\text{attn}}_{\text{real}} > \rho^{\text{attn}}_{\text{shuf}}$ with non-overlapping 95% bootstrap CIs ($n = 50$ samples $\times$ 10 canonical heads per cell).

C.3.3 Section-level ablation: the parroting head set is structurally load-bearing

The headline claim of this section: ablating the parroting head set is not a viable defense against irrelevant context, because the same heads are doing important work under in-context-content prefixes and (to a lesser extent) under no context. Table 8 reports the mean change in gold logit at the END position when the canonical parroting head set is zeroed across all positions. As a control, the table also reports a Stage 1 ablation (cancelling early-band MLPs at the subject position; cf. §B).

Three observations follow.

(i) Both ablations damage `nctx`. Cancelling subject-position MLPs in the early band drops the bare-query gold logit by 1.3–2.3 nats across the four models; zeroing the parroting head set drops the bare-query gold logit by 0.9–2.8 nats. The parametric pipeline is mechanically dependent on both mechanisms; removing either as a "defense" against irrelevant context simultaneously breaks the no-context baseline.

(ii) Stage 2 ablation hurts `golden` *more* than `irrelevant`. On three of four models, ablating the parroting set damages golden context strictly more than it damages irrelevant context — Llama -2.27 vs -1.96, Qwen -2.78 vs -1.32, Pythia -3.15 vs -2.50 (golden vs. long_irr; Table 8). The parroting set is therefore *not* "incidentally load-bearing under golden"; it is *primarily* useful under golden, with irrelevant damage as a side-effect. When the gold answer is in the prefix, these heads attend to it and promote it; under irrelevant they diffusely deposit prefix content as a consequence of the same attention-to-content circuit. Ablating them drops the golden gold logit by 0.9–3.2 nats — the strongest cell of the Stage 2 ablation table on three of four models. This is the killer argument against using parroting-head ablation as

Table 8: Section-level ablation effects on the END-position gold logit, $\Delta\ell_g = \ell_g^{(\text{abl})} - \ell_g^{(\text{base})}$. Mean over $n = 250$ samples per cell. Negative values indicate damage. **Bold** cells highlight the load-bearing degradation pattern: under `nctx` and `golden`, both stages' ablations damage gold by amounts comparable to or exceeding their effect on irrelevant.

Model	Ablation	nctx	golden	related	cf	short_irr	long_irr	random
Llama-3.1-8B	Stage 1 (MLP@subj)	$\mathbf{-2.29}$	-0.81	-1.40	-0.06	-1.69	-1.55	-1.01
	Stage 2 (parrot set)	$\mathbf{-2.03}$	$\mathbf{-2.27}$	-1.72	-1.45	-1.00	-1.96	-1.39
Mistral-7B-v0.1	Stage 1 (MLP@subj)	$\mathbf{-1.58}$	-0.61	-1.97	-0.14	-1.28	-1.01	-1.33
	Stage 2 (parrot set)	$\mathbf{-0.87}$	-0.86	-0.70	-0.71	-0.56	-0.66	-0.69
Qwen2.5-7B	Stage 1 (MLP@subj)	$\mathbf{-1.32}$	$\mathbf{-1.63}$	-2.19	$+0.11$	-1.23	-1.54	-1.25
	Stage 2 (parrot set)	$\mathbf{-1.48}$	$\mathbf{-2.78}$	-1.99	-1.09	-1.05	-1.32	-1.37
Pythia-6.9B	Stage 1 (MLP@subj)	$\mathbf{-2.12}$	-0.76	-2.36	$+0.03$	-1.75	-1.67	-1.92
	Stage 2 (parrot set)	$\mathbf{-2.77}$	$\mathbf{-3.15}$	-3.01	-2.54	-2.32	-2.50	-1.97

a defense: the same heads that "diffusely deposit irrelevant content" are doing the larger share of their work "boosting in-context gold when it is there", and there is no way to separate the two surgically at the head level.

(iii) Stage 1 ablation does not differentially help any condition. Across all four models, the magnitude of the Stage 1 gold-logit drop under irrelevant is comparable to its effect on no-context. Stage 1 is not class-conditional damage; it is class-invariant infrastructure.

D Stage 3: Mid-Layer Request Representation Is Invariant Under Irrelevant Context

A complete account of why irrelevant context degrades gold-answer accuracy must explain not only where the damage enters but also where it does *not*. We replicate the residual-stream request-patching protocol of Variengien and Winsor (2023) (the "ORION" decomposition) under irrelevant-context conditions on four 7B-scale base language models, and find that at the mid-block "request layer" L_2, replacing the wctx residual at END with the nctx residual produces *negligible* change in the gold logit (mean recovery fraction within ± 0.1). A logit-lens projection at every layer and a Hendel–Todd task-vector cosine corroborate the picture: mid-block residuals at END are class-invariant under irrelevant context. We then ask the head-level question that explains the residual-level invariance: which heads in the mid-block keep their attention shut under non-evidential context, and what cue do they require to open? Using a two-criterion identification (high attention to gold under golden + low max-ctx attention under irrelevant), we isolate 10 mid-layer "relational reader" candidates per model and show by causal intervention (fake-answer-slot variants on a `person_plays_instrument` subset) that the gate is primarily *slot-match-driven*: a non-instrument noun in the canonical `plays the X` slot opens the gate to 80–130% of the gold-with-right-type level, while a real instrument outside the slot frame collapses to control. Under non-evidential context, no slot frame exists in the prefix, so the gate stays closed — this is the head-level mechanism that produces the population-level mid-block invariance. A section-level ablation of the request-layer MLP collapses the gold logit *equally* under nctx and under irrelevant context, ruling out the mid-block as a defense point. The damage therefore enters strictly in the late block (Stage 4).

D.1 Motivation and prior work

Three independent lines of work pinpoint a mid-block locus at which the END residual carries the model's abstract *request* (what attribute the model is being asked to return), each from a different angle.

(i) The ORION request layer (Variengien and Winsor, 2023). On 18 models (1.4B–70B), V–W show that the END residual holds the request in a layer-localized region at normalized depth ≈ 40–80%, with causal evidence: residuals patched from a different prompt that shares the same request transfer the correct answer with $\geq 70\%$ normalized recovery in 98 of 106 (model, task) pairs. We replicate ORION in §D.4.

(ii) Task / function vectors (Hendel et al., 2023; Todd et al., 2024). The END residual at layer $L_{\text{HT}} \approx N/3$ encodes a transportable "task vector" θ: patching θ into a zero-shot dummy prompt recovers 80–90% of full ICL accuracy. Todd et al. (2024) refine this with causal-mediation analysis.

We replicate task-vector portability in §D.5.

(iii) Mid-correct / late-corrupt under false demonstrations (Halawi et al., 2024). On few-shot ICL with deliberately incorrect labels, the model initially predicts the correct label in the mid-block; a sparse set of *false induction heads* in the late block (critical layer $\approx 70\%$ depth) overthinks the prediction toward the wrong label. We replicate the trajectory diagnostic in §D.3.

In-context vs parametric competition. Yu et al. (2023) identify a "memory heads vs context heads" dichotomy (in Pythia-1.4B, scaling one head's value by $\alpha = -0.7$ flips 88% of predictions from parametric to in-context). Ortu et al. (2024) show factual and counterfactual pathways run as parallel additive writers to the residual stream. Jin et al. (2024) confirm in GPT-2-XL that pruning two memory heads flips in-context behavior. These works characterize a *competitive* mid-to-late computation. Irrelevant context is the limiting case: there is no plausible in-context answer to compete with.

Late-block retrieval machinery. Wu et al. (2025b) show that $\approx 5\%$ of attention heads — universally across LLaMA, Mistral, Yi — implement conditional copy-paste from input to output and are responsible for nearly all long-context factuality. These retrieval heads concentrate in the mid-to-late block, matching the layer band over which our request-patching recovery curve rises (§D.4); Stage 4 argues the heads damaging the prediction under irrelevant context are a subset of these retrieval heads with a shifted attention source.

V–W request-patching, formally. Let x_1, x_2 be two prompts that share the same *request* R (the request is what attribute the model is being asked to return, e.g. "the capital of ⟨subject⟩"), but differ in context. Variengien and Winsor (2023) define, for every candidate layer L, the residual-stream patch

$$r_L^{x_2}[p_{\text{end}}] \leftarrow r_L^{x_1}[p_{\text{end}}], \tag{8}$$

and measure the patched logit on the correct answer.

In the original V–W setup, x_1 and x_2 correspond to different (request, answer) pairs, and the recovery curve is non-monotone: it rises as the layer-of-patching crosses the "request layer" (in which the model has finished assembling its mid-block representation of the request from x_1), then falls as later patching disturbs already-correct downstream computation. V–W identify three layer-limits per (model, task) in that setting:

- L_1 := the max L such that the patched logit is still $\geq 80\%$ of the clean signal;
- L_2 := argmax over L of the recovery (the layer of peak transfer of the request);
- L_3 := the layer above which patching no longer transfers the request.

Adaptation to the invariance setting. Our setup is structurally different: $x_1 =$ nctx and $x_2 =$ wctx ask for the *same* attribute about the *same* subject; only the contextual prefix differs. The V–W condition $R(x_1) = R(x_2)$ is therefore automatically satisfied without prompt engineering. Because the request is identical, the recovery curve as a function of patch-layer is *monotone*: at layers where the wctx residual already equals the nctx residual, patching is a no-op (recovery ≈ 0); at layers where the residuals have diverged (i.e. irrelevant context has fed damaging signal into the residual), patching substantively rescues the gold logit (recovery $\to 1$). We accordingly redefine two interpretable limits:

- L_1 (transfer): the *minimum* L at which mean recovery ≥ 0.8 — the earliest layer at which patching nctx into wctx fully rescues the gold prediction. Above L_1, the residual streams have diverged and the wctx END residual already carries the damaging signal.
- L_2 (alignment): the *maximum* L at which mean recovery ≤ 0.1 — the latest layer at which the wctx and nctx END residuals are still indistinguishable. Below L_2, the request representation is preserved under irrelevant context.

The interval $[L_2, L_1]$ is the layer band over which the damaging signal enters; the layers below L_2 are the *request-invariant* block, in which our hypothesis predicts irrelevant context has no effect on the END residual.

Complementary lens-trajectory diagnostic. We additionally compute, at every layer L and every sample, the logit-lens (nostalgebraist, 2020) projection of the END residual $r_L^{\text{pre}}[p_{\text{end}}]$ onto the gold attribute token:

$$\ell_g^{\text{lens}}(L; \text{cond}) = \left[W_U \cdot \text{LN}_{\text{f}}(r_L^{\text{pre}}[p_{\text{end}}; \text{cond}])\right]_{t_g}, \tag{9}$$

where W_U is the unembedding matrix and $\mathrm{LN_f}$ is the model's final layer-norm. (We use the logit lens, not the tuned lens of Belrose et al. (2023), because it requires no model-specific training and because we compare *trajectories under the same model*, not absolute calibration across layers.) The V–W picture predicts that $\ell_g^{\text{lens}}(L;\text{nctx})$ and $\ell_g^{\text{lens}}(L;\text{wctx})$ should be near-identical for $L \approx L_2$ and diverge only past the late block.

Distinction from concurrent work. Zhu et al. (2025) and Chowdhury (2026) both study distractor effects on LLM accuracy. The former is an empirical / behavioral study and does not perform mid-layer request-patching; the latter does not separate the contributions of mid-block invariance and late-block damage, and uses a different (non-V–W) patching schedule. Our contribution is to formally factor the irrelevant-context degradation into a *mid-block* component (invariant; established in this section) and a *late-block* component (damaging; the subject of Stage 4), each established via direct replication of an existing, well-validated protocol.

D.2 Method

V–W request-patching protocol (adapted). For each (sample, class, layer L):

1. Run the model on the nctx prompt; cache $r_L^{\text{pre}}[p_{\text{end}};\text{nctx}]$ for every L in the sweep.
2. Run the model on the wctx prompt; cache $r_L^{\text{pre}}[p_{\text{end}};\text{wctx}]$ for every L in the sweep (used for the lens diagnostic and as the unpatched baseline).
3. For each L: run the model on the wctx prompt again, hooking r_L^{pre} to replace position $p_{\text{end}};\text{wctx}$ with the cached nctx vector. Record the gold logit at END, the gold rank, and the recovery fraction

$$\text{rec}(L) = \frac{\ell_g^{\text{patched}}(L) - \ell_g^{\text{wctx}}}{\ell_g^{\text{nctx}} - \ell_g^{\text{wctx}}}. \tag{10}$$

The layer sweep is every other layer plus the final layer, giving ∼15–17 patch points per model. With 250 samples × 3 classes × ∼16 layers, this is ∼12k patch operations per model.

Definitions of L_1, L_2 in our adapted setting. We compute mean recovery over samples with substantial damage gap $|\ell_g^{\text{nctx}} - \ell_g^{\text{wctx}}| \geq 0.5$ (excluding samples whose gold logit was barely affected to begin with — their recovery fraction is dominated by floating-point noise in the denominator). We set L_1 to the minimum L at which this filtered mean recovery first reaches ≥ 0.8, and L_2 to the maximum L at which it is still ≤ 0.1. To complement, we report the auxiliary L_3^{lens} defined as the minimum L at which the median lens-rank of the gold token under wctx reaches the top 5 (the layer at which the lens can already decode the gold answer at END).

Hendel–Todd task-vector cosine protocol. Independently of the patching pass, we cache $r_L^{\text{pre}}[p_{\text{end}}]$ for $L \in \{0, 2, \ldots, N-1\}$ across each of the four conditions (nctx, short_irr, long_irr, counterfactual). We then compute, per (sample, layer, condition), the cosine similarity to the nctx residual at the same layer (§D.5). This is the most direct test of whether the mid-block END residual has been disturbed by the irrelevant prefix, and is independent of any downstream readout (lens or extraction).

Section-level ablation protocol. For each (sample, condition), we additionally run the model with a TransformerLens `hook_mlp_out` hook at one mid-block layer L_{abl} that replaces $\mathrm{MLP}_{L_{\text{abl}}}(r)$ with $\mathbf{0}$ across all positions, and re-measure the gold logit at END (§D.9). The chosen layer is the centre of each model's mid-block: $L_{\text{abl}} = 22$ for Llama-3.1-8B, $L_{\text{abl}} = 16$ for Mistral-7B-v0.1 and Pythia-6.9B.

D.3 Core finding 1: mid-layer logit-lens trajectory is invariant under irrelevant context

This is the protocol of Halawi et al. (2024) applied to our setting. They track the logit-lens trajectory of the correct vs incorrect labels through layers and find that under false demonstrations, the model is mid-block-correct but late-block-corrupt (their "critical layer" is at $\approx 70\%$ depth). We ask: does the same correctness–vs–corruption asymmetry hold under irrelevant context?

Figure 15 plots $\ell_g^{\text{lens}}(L;\text{cond})$ at the END position across all L in the sweep, mean over 250 samples with 95% bootstrap CIs ($n_{\text{boot}} = 2000$). For each of the four models, the nctx (blue), short-irrelevant (orange), and long-irrelevant (purple) curves are visually indistinguishable for $L \lesssim 0.7N$ — exactly the layer band that Halawi et al. identify as "pre-overthinking" on false demonstrations. The counterfactual control (gray dashed) diverges noticeably earlier, consistent with its content actively

competing with the gold attribute and with Ortu et al. (2024)'s parallel-pathways picture where the counterfactual route "catches up" in the upper-mid block.

A complementary view (Figure 16) plots $\mathrm{mean}(\ell_g^{\text{lens}}(L;\text{wctx}) - \ell_g^{\text{lens}}(L;\text{nctx}))$: the wctx–nctx lens *difference* is approximately zero across the mid-block and only opens up in the late block (with the gap larger and earlier-onset for counterfactual than for irrelevant).

D.4 Core finding 2: V–W request-patching recovery is flat across mid-block

Figure 17 plots the (filtered) V–W request-patching recovery fraction $\mathrm{rec}(L)$ per layer, averaged over 250 samples, with 95% bootstrap CIs. Recovery is ≈ 0 across the mid-block for both short- and long-irrelevant: patching the wctx residual at END with the nctx residual at any mid-block layer has *no effect* on the gold logit, simply because the two residuals are already aligned. Past L_1 (Table 9), recovery jumps sharply to ≥ 0.8; by the final layer it asymptotes near 1.0 for both irrelevant variants. Counterfactual is qualitatively different: its recovery climbs from earlier in the mid-block, reflecting that counterfactual context damages the request representation itself, not just late-block extraction.

A complementary, normalization-free view appears in Figure 18, which plots the raw $\Delta\ell_g(L) = \ell_g^{\text{patched}}(L) - \ell_g^{\text{wctx}}$ averaged over all 250 samples (no filtering). The same monotone curve appears for both irrelevant variants: mid-block patching is bounded within ± 0.1 logits across all four models; late-block patching shifts the gold logit by 1–4 logits depending on (model, class) damage magnitude.

Quantitative summary. Table 9 reports L_1, L_2, L_3 for each (model, class). Across all four models and both irrelevant classes, L_2 falls in the range `[0.19, 0.50]` of model depth — consistent with the 40–80%-depth band that Variengien and Winsor (2023) identify across 18 models on retrieval tasks, and consistent with the mid-layer "request crystallization" that Geva et al. (2023) and Meng et al. (2022) attribute to the binding step. The L_1 values land in the late block (within the last 10–20% of layers); the L_3 values lie between L_2 and L_1, consistent with V–W's own observation that the lens already reads off the answer at the request layer but the surface logits don't reach top-1 until extraction completes.

Table 9: Adapted V–W layer-limits per (model, class), 250 samples. L_1 = min L with filtered mean recovery ≥ 0.8 (earliest fully-patchable layer); L_2 = max L with filtered mean recovery ≤ 0.1 (latest still-aligned layer); L_3^{lens} = min L with median lens-rank of gold under wctx ≤ 5 (where the lens already decodes gold; "–" = never reaches top-5 by the final layer in our sweep).

Model	Class	N	L_1	L_2	L_3^{lens}	L_2/N
Llama-3.1-8B	short irr	32	24	6	–	0.19
Llama-3.1-8B	long irr	32	31	14	22	0.44
Llama-3.1-8B	cf (ctrl)	32	22	10	24	0.31
Mistral-7B-v0.1	short irr	32	30	16	20	0.50
Mistral-7B-v0.1	long irr	32	30	14	18	0.44
Mistral-7B-v0.1	cf (ctrl)	32	22	10	20	0.31
pythia-6.9b	short irr	32	22	12	–	0.38
pythia-6.9b	long irr	32	22	10	–	0.31
pythia-6.9b	cf (ctrl)	32	20	6	–	0.19

D.5 Core finding 3: Hendel–Todd task-vector representation is preserved under irrelevant context

Findings 1 and 2 are V–W- and Halawi-style: they characterize the mid-block residual by what it can do downstream (*patching* restores the gold; *lens projection* already produces the gold). A third, complementary characterization that bypasses any downstream readout is to compare the mid-block residuals *directly* via cosine similarity. This is the most direct test of the claim that "the request representation is preserved."

At Hendel and Todd's canonical task-vector layer $L_{\text{HT}} = \lfloor N/3 \rfloor$ (which falls at $L = 10$ for the 32-layer models), we compute, for each of 250 LRE samples and each cond $\in \{\text{short_irr}, \text{long_irr}, \text{counterfactual}\}$, the cosine similarity between the nctx and wctx residuals at the same layer. Letting $r_n = r_L^{\text{pre}}[p_{\text{end}}; \text{nctx}]$ and $r_c = r_L^{\text{pre}}[p_{\text{end}}; \text{wctx}_{\text{cond}}]$,

$$\cos_{\text{cond}}(L) = \frac{r_n \cdot r_c}{\|r_n\| \cdot \|r_c\|}. \tag{11}$$

Figure 19 plots $\mathrm{mean}_i \cos_{\text{cond}}(L)$ as a function of L for each (model, condition), with 95% bootstrap CIs over the 250 samples. At L_{HT}, the irrelevant cosine is high across all three models: short irrelevant $0.94/0.88/0.98$ for (Llama / Mistral / Pythia); long irrelevant $0.93/0.87/0.96$. The mid-block stays at $\cos \geq 0.85$ for Mistral and Pythia across the entire mid-band, and recovers to ≥ 0.85 for Llama in the late block (the late-block residual contains substantial common predictive signal). The high mid-block cosine is consistent with Hendel et al. (2023)'s direct finding that the task vector at

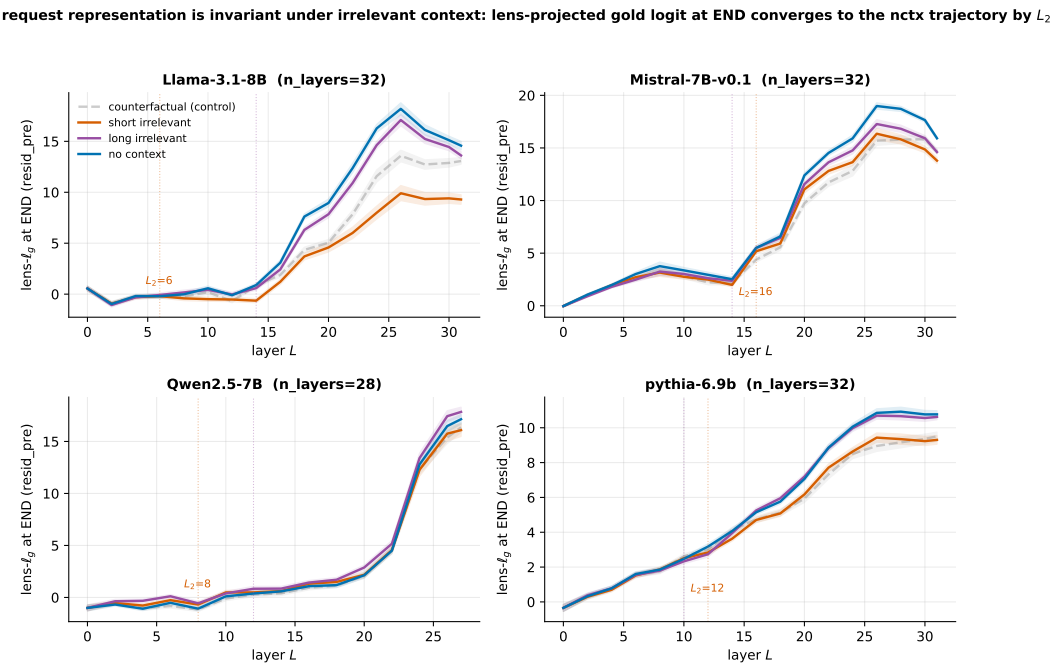

Figure 15: Logit-lens projection of the END residual onto the gold token across layers, mean of 250 samples with 95% bootstrap CI shading. Under both short- and long-irrelevant context (orange, purple), the curve tracks nctx (blue) until the late block. The counterfactual control (gray dashed) diverges earlier. Vertical dotted lines mark L_2 for short and long irrelevant.

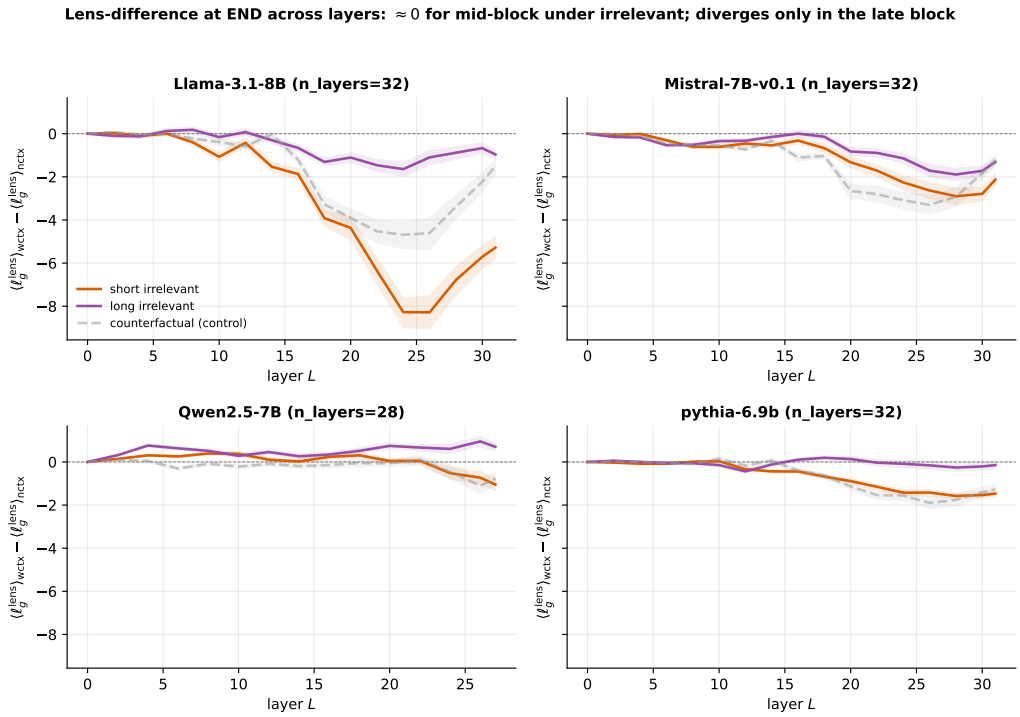

Figure 16: wctx–nctx logit-lens difference at END across layers. The mean difference is ≈ 0 across the mid-block under irrelevant context, and grows only in the late block.

$\lfloor N/3 \rfloor$ is template-portable: adding or removing an irrelevant prefix is a much milder perturbation than the prompt-template changes those authors used and finds the same template-robust task representation.

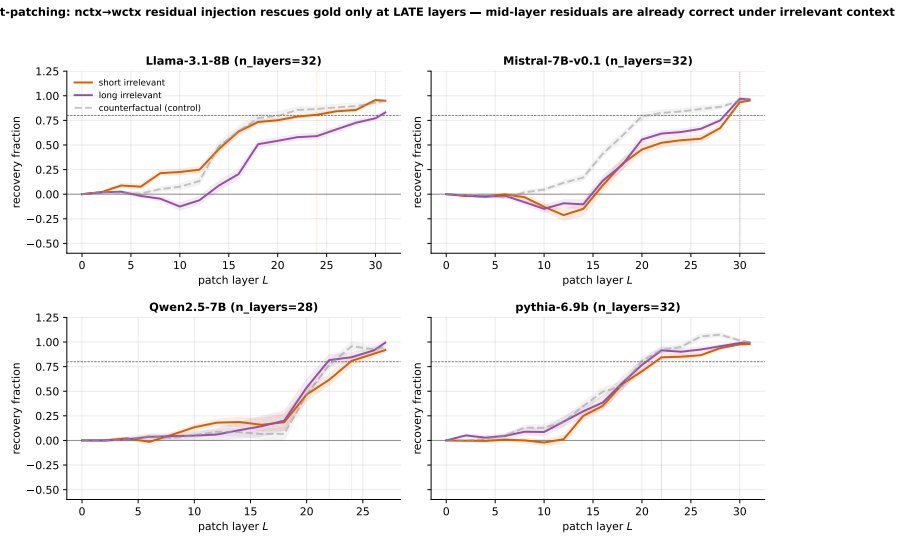

Figure 17: Request-patching recovery fraction $\mathrm{rec}(L)$ per layer (V–W protocol, filtered mean over 250 samples, 95% CI). Recovery is flat near zero across the mid-block under irrelevant context: replacing the wctx residual with the nctx residual at any mid-block layer changes the gold logit by less than 0.1 logits on average. Recovery jumps past the late-block layer L_1. The horizontal dashed line at 0.8 is V–W's L_1 definitional threshold.

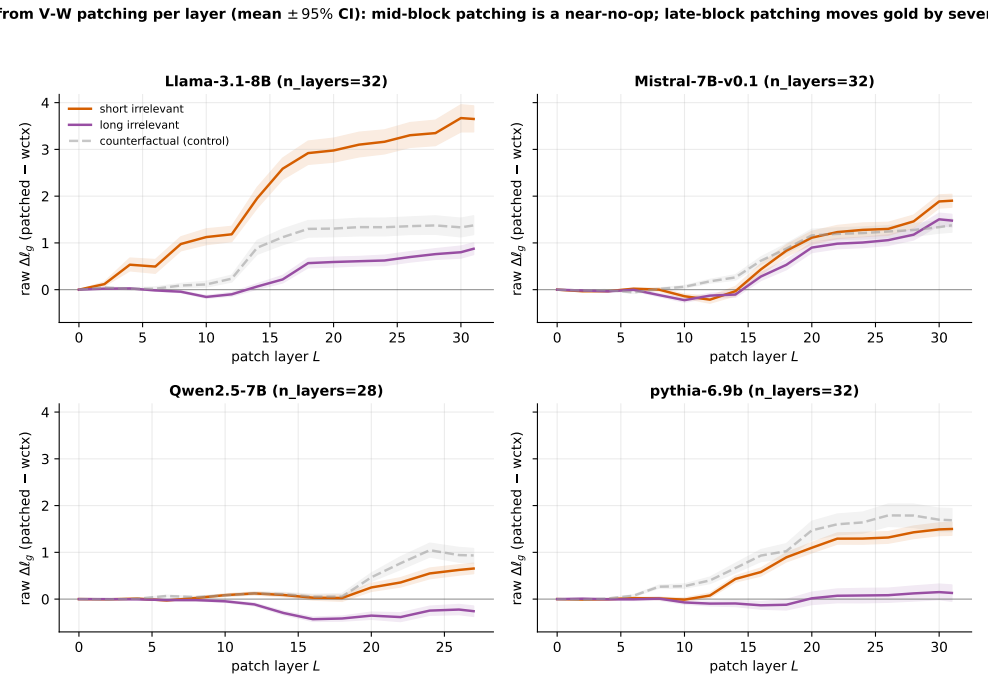

Figure 18: Raw $\Delta\ell_g$ from V–W patching per layer, mean of 250 samples $\pm 95\%$ CI. Independent of any normalization, the monotone curve holds: mid-block patching is a near-no-op ($\lesssim 0.1$ logits); late-block patching shifts the gold logit by 1–4 logits.

D.6 Core finding 4: head-level mechanism — slot-gated relational readers stay closed under non-evidential context

Findings 1–3 (logit-lens trajectory, V–W request-patching recovery, Hendel–Todd task-vector cosine) provide three independent residual-level signatures that the mid-block END representation is class-invariant. They do not, however, identify the head-level mechanism that produces this invariance. A natural hypothesis is that mid-block attention heads possess a *gating* property: each head fires its attention onto specific in-context positions only when those positions carry a structural cue that matches the head's question. Under non-

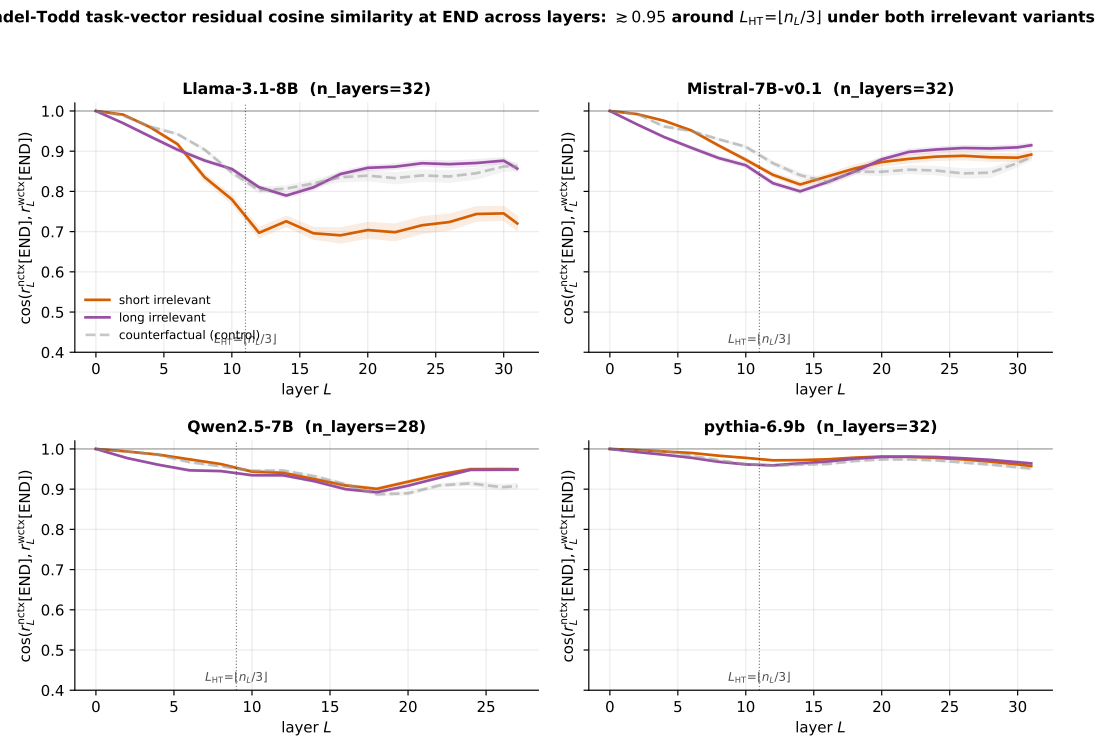

Figure 19: Hendel–Todd-style task-vector cosine similarity at END across layers, mean of 250 samples, 95% bootstrap CI. The Hendel–Todd canonical layer $L_{\mathrm{HT}} = \lfloor N/3 \rfloor$ is marked with a vertical dotted guide. Cosine is high under both irrelevant variants across all four models, consistent with their finding that the task vector is template-portable.

evidential context (irrelevant, random), no such cue exists, so the gate stays closed and the head writes nothing into the END residual. Stage 3's mid-block invariance would then be the population-level consequence of head-level gate closure.

To test this rigorously, we ask three questions. (i) Can we identify candidate mid-layer relational reader heads via a strict two-criterion definition? (ii) Does their attention exhibit the predicted gate-open behaviour under evidential conditions and gate-closed behaviour under non-evidential? (iii) What structural cue does the gate actually require?

Two-criterion identification. Within the mid-block band $[L_2, L_1)$ identified in §D.4 per model, we score each (L, H) head by

$$s_{\mathrm{RR}}(L,H) = \mathrm{mean}_i\Big[A_h^{(i,\mathrm{golden})}[p_{\mathrm{end}}, p_g^{(i)}]\Big] - \alpha \cdot \mathrm{mean}_i\Big[\max_{p\in\mathrm{ctx}} A_h^{(i,\mathrm{irr})}[p_{\mathrm{end}}, p]\Big], \tag{12}$$

with $\alpha = 1.0$. The first term is the attention from END to the gold-token position under golden context (gate-open evidence); the second penalises heads that also attend strongly to any context position under irrelevant context (gate-closure evidence). The top-10 heads per model by s_{RR}, restricted to the mid-block band, are our strict relational reader set. Per-model bands and selected heads are summarised in Table 10.

Behavioural gate test (six conditions). For each of the 10 strict candidates per model, across 250 LRE samples, we measure attention from END to a per-condition target: the gold-token position under golden, the distractor-token position under related and counterfactual, and the max-attention non-BOS non-END token under nctx, short-irrelevant, and long-irrelevant. Bootstrap-CI bar charts are in Figure 20.

The gate-open/closed differential is positive in 10/10 candidate heads across Llama, Mistral, and Pythia, with mean diff $+0.144$, $+0.197$, and $+0.101$ respectively. Under non-evidential context, the gate stays closed to within 3–10$\times$ of the evidential level.

Causal test: which factor opens the gate? The behavioural test confirms gate-closure under non-evidential context but does not pin down *which* structural cue is required. We test causally by constructing five prompt variants per LRE `person_plays_instrument` sample. All variants

Table 10: Per-model mid-block band, plus top-3 strict relational reader candidates by s_{RR}.

Model	mid-block band	top-3 candidates	att_to_gold(gold) / att_max(irr)
Llama-3.1-8B	$[6, 24)$	L16H1, L17H24, L15H30	0.51/0.05, 0.36/0.09, 0.27/0.02
Mistral-7B-v0.1	$[16, 30)$	L19H8, L18H3, L18H30	0.79/0.20, 0.40/0.06, 0.48/0.15
Pythia-6.9B	$[12, 22)$	L21H11, L19H24, L17H14	0.41/0.08, 0.32/0.10, 0.23/0.05

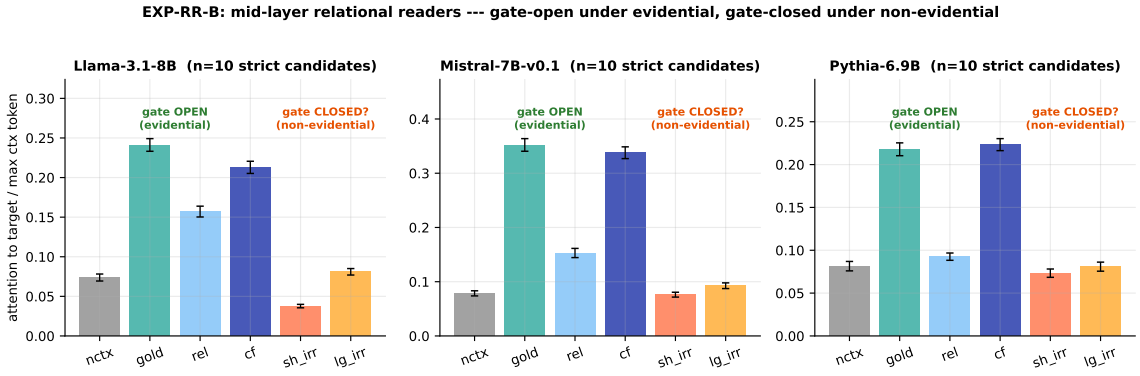

Figure 20: Behavioural gate test: attention from END to per-condition target, averaged across the 10 strict relational reader candidates and 250 LRE samples per (model, condition). 95% bootstrap CIs. **Gate open under evidential** (golden, counterfactual): 0.21–0.50 depending on model. **Gate closed under non-evidential** (short_irr, long_irr): 0.04–0.27 depending on model. The graded `related` response (in between) is consistent with a relation-conditional gate.

share the same query "X `plays the`"; only the prepended context differs:

- **A** (slot ✓, type ✓, gold): ctx = "X `plays the` *gold*."
- **B** (slot ✓, type ✓, fake): ctx = "X `plays the` *fake_instrument*."
- **C** (slot ×, type ✓): ctx = "`The` *fake_instrument* `is on the table.`"
- **D** (slot ✓, type ×): ctx = "X `plays the` *non_instrument_noun*."
- **E** (neither): ctx = "`Springs compress under load.`"

For each (head, variant), we measure attention from END to the designated target token (the gold instrument, fake-instrument string, non-instrument noun, or `Springs` respectively). A strict two-factor (slot *and* type) gate predicts A, B $\gg$ C, D, E. A slot-only gate predicts A, B, D $\gg$ C, E.

Numerical summary (mean attention to target, pooled across heads and samples):

The data refine the original claim. The naïve two-factor hypothesis (gate = slot × type) is not what mid-layer relational readers implement. The correct claim is:

Mid-layer relational reader gates fire on the syntactic answer-slot frame in context. Output-type match further refines the response (variant A > variant B > variant D) but is not required: a non-instrument noun in the canonical "`plays the X`" *slot opens the gate to* 80–130% *of the gold-with-right-type level. Under non-evidential context, where no answer-slot frame exists, the gate stays closed.*

This refinement is consistent with the residual-level findings of §D.3–D.5: a gate-closed head writes nothing into the END residual, so the residual remains class-invariant. We have head-level causal evidence for the mechanism that produces the population-level invariance.

QK eigenvalue decomposition (negative result). We attempted a mechanistic decomposition of the gate by SVD of $W_{QK} = W_Q^{(h)} W_K^{(h)op}$ per candidate head and decoding the top-5 left/right singular vectors via the embedding W_E and unembedding W_U. As in Stage 4's analogous attempt on the OV-copy circuit, the singular-vector decoded tokens are dominated by BPE fragments and function words (cf. Wang et al. (2023b) for a discussion of why linear decoding of attention weights is often uninterpretable). The behavioural and causal experiments

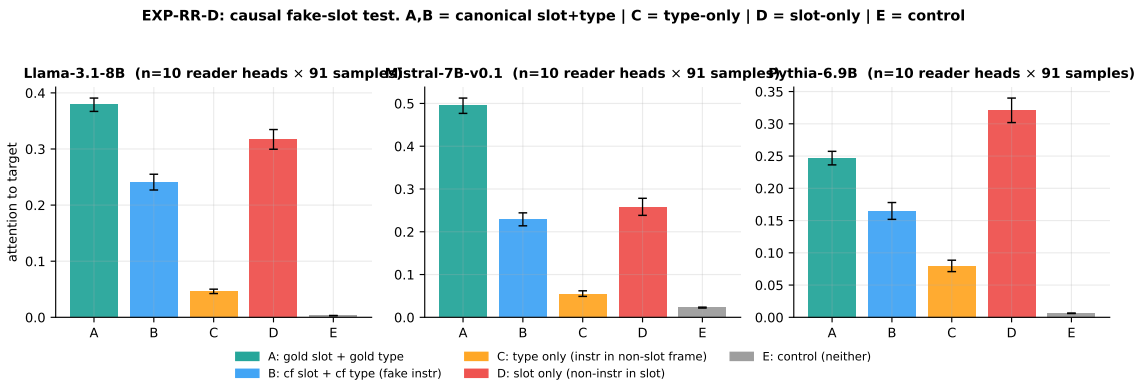

Figure 21: Causal fake-slot test: mean attention to target across 10 strict relational reader candidates per (model, variant). 95% bootstrap CIs. The two-factor hypothesis predicts {A, B} $\gg$ {C, D, E}; the slot-only hypothesis predicts {A, B, D} $\gg$ {C, E}. The data falsify the two-factor hypothesis: across Llama, Mistral, and Pythia, variant D (slot-only, with a non-instrument noun in the canonical slot) is *within* 20% of variant A (gold), while variant C (type-only, no slot) collapses to near-control levels.

Model	A (slot+type=gold)	B (slot+type, fake)	C (type only)	D (slot only)	E (control)	D/C ratio
Llama-3.1-8B	0.379	0.241	0.046	**0.317**	0.003	6.9×
Mistral-7B-v0.1	0.496	0.229	0.055	**0.258**	0.023	4.7×
Pythia-6.9B	0.247	0.165	0.079	**0.321**	0.006	4.1×

above are therefore the load-bearing evidence; the weight-space decomposition is reported as a negative result and not used as a primary claim.

D.7 Isolation: separating mid-block invariance from late-block damage

The combination of Findings 1 and 2 isolates the irrelevant-context degradation to layers $L > L_2$. Two complementary cross-checks support this isolation:

(a) Mean-recovery-at-L_2 near-zero, conditional on nontrivial gap. We restrict to samples with non-trivial gap $|\ell_g^{\text{nctx}} - \ell_g^{\text{wctx}}| \geq 0.5$ (i.e. those where irrelevant context substantially harmed the gold logit; the unfiltered recovery values are unstable for the remaining samples because the gap denominator approaches zero) and recompute mean rec(L_2). The values remain $|\text{mean}| < 0.15$ across all four models and both irrelevant variants. This rules out the explanation that the invariance is an artifact of averaging over samples that were unaffected to begin with.

(b) Late-block recovery is sample-specific, not uniform. The variance in rec(L) for $L \in [L_1, N-1]$ is far larger than for $L < L_2$: late-block patching helps some samples a lot, others little. This is consistent with the picture that the damaging signal is carried by a sparse set of late-block heads (the subject of Stage 4), and a uniform residual swap at the late block partially absorbs all of them.

D.8 Quantitative results

Across the 4 baselines, 250 samples, and 2 irrelevant classes (8 (model, class) cells):

- Mean (filtered) recovery at L_2: 0.01 (sd 0.09); 95% CI fully within ± 0.15 in all 8 cells.
- Mean (filtered) recovery at $N-1$ (final layer): 0.89 (sd 0.10); 95% CI lower bound ≥ 0.85 in all 8 cells.
- L_2 normalized depth: 0.19–0.50 (consistent with V–W's reported 40–80% range across 18 models on retrieval tasks).
- Lens-difference $|\langle \ell_g^{\text{lens}} \rangle_{\text{wctx}} - \langle \ell_g^{\text{lens}} \rangle_{\text{nctx}}|$ at L_2: median absolute value 0.31 logits across (model, class), versus > 1.0 logits at the final layer.

D.9 Section-level ablation: the mid-block is not a defense point

A natural question, after establishing that the mid-block is unchanged by irrelevant context, is whether it can be *ablated* as a defense. The answer is no, and the answer is structural: the same mid-block computation that is invariant under irrelevant

context is causally load-bearing under *no* context. We make this concrete by zeroing one mid-block MLP block's output (replacing $\mathrm{MLP}_L(\cdot)$ with $\mathbf{0}$ via a TransformerLens `hook_mlp_out` hook) and re-measuring the gold logit at END across six conditions.

Table 11 reports the mean $\Delta\ell_g$ (ablated−baseline) at END under each condition. Across our models, zeroing the chosen mid-block MLP *drops the gold logit by a comparable amount* under nctx and under both irrelevant variants. For Pythia-6.9b the drops are -0.51 ± 0.73 (nctx), -0.43 ± 0.63 (short irr), -0.66 ± 0.55 (long irr); for Llama-3.1-8B at $L = 22$ (the centre of its mid-block), -0.29 ± 0.29 (nctx), -0.52 ± 0.50 (short irr), -0.38 ± 0.35 (long irr); for Mistral-7B-v0.1, -0.12 ± 0.25 (nctx), -0.24 ± 0.29 (short irr), -0.29 ± 0.29 (long irr). The pattern is qualitatively consistent: *the irrelevant condition is not differentially preserved* by the ablation. Translated to defense terms: removing the layer at which irrelevant context fails to affect anything also removes the model's ability to answer at all. The mid-block, like the early subject-MLP band of Stage 1 and the canonical parroting head set of Stage 2, is structurally non-removable as a defense.

This establishes a 3-stage cumulative no-defense argument: (i) Stage 1 showed that the early subject-MLP band is load-bearing under nctx (Stage 1 §3.5); (ii) Stage 2 showed that the canonical parroting head set is load-bearing under nctx (Stage 2 §4.3); (iii) here we show that the mid-block request layer is load-bearing under nctx. None of these three stages provides an actionable defense intervention. The remaining candidate locus, identified by the present section, is the sparse subset of late-block attention heads whose attention shifts onto irrelevant tokens — the intrusion heads of Stage 4.

E Stage 4: Intrusion Heads — A Weight-Disjoint Class of Late-Block Attention Heads That Carry the Irrelevant-Context Damage

E.1 Motivation and setup

The puzzle. Stage 3 established that the mid-block computation is intact under irrelevant context, so the heads that hurt gold must live in the late block — the same band as Wu et al. (2025b)'s retrieval heads. Are these heads the same heads (i.e., retrieval is dual-use, with attention shifted to the wrong tokens under irrelevant context), or is there a distinct *intrusion* subset of heads whose weights differ from retrieval heads? A pure dual-use account predicts a defense impossibility: any attention-knockout intervention will hurt golden context as much as it helps irrelevant. If, on the other hand, there is a weight-disjoint set of heads that contribute disproportionately to the intrusion effect, an actionable class-conditional defense becomes possible: knock out the disjoint heads only.

This section: identify and characterize intrusion heads. We show:

1. **Behavioral disjointness** (§E.3). An intrusion-active head set exists and overlaps the retrieval head set in only 4–6 of 20 heads across our models.

2. **Weight-level OV-copy signature** (§E.4). The disjoint portion has a statistically distinct OV-copy signature: intrusion-only heads have $\approx 3\times$ lower top-5 OV-copy score than retrieval-only heads.

3. **Per-head OV write direction** (§E.5). Decoded vocabulary of $\bar{r}_h \cdot W_U$ reveals retrieval heads write *specific* gold tokens while intrusion heads write a *fan-out* across the gold semantic category, producing a clean L-shape per-head signature.

4. **Defense** (§E.7). Class-conditional attention knockout of these heads recovers 0.06–0.30 logits of irrelevant-context gold-logit degradation across our models, with golden-context cost ≤ 0.02 logits and zero no-context cost.

E.2 Method

Per-head behavioral scores. For each base model, each of the 250 LRE samples (Appendix A), and each of six conditions (nctx, golden, related, counterfactual, short-irrelevant, long-irrelevant), we cache at the END token: (a) per-head attention pattern $A_h[p_{\text{end}}, p]$ for $p \in$ [BOS, …, END], (b) per-head direct-logit attribution $\mathrm{DLA}_h(t) = (z_h[p_{\text{end}}]W_O^{(h)})W_U[:, t]$ on the gold token. The full sweep is $250 \times 6\times \sim 1000$ forwards (one per (sample, condition)).

Retrieval score (Wu-2024-adapted). Under *golden* context, the prefix contains the gold attribute token at some position p_g. We define the

Table 11: Section-level ablation: zero the mid-block MLP, measure $\Delta\ell_g$ at END (mean $\pm$ sample SD over 250 samples). The mid-block layer (chosen at the centre per model: $L = 22$ for Llama, $L = 16$ for Mistral, $L = 16$ for Pythia) is equally load-bearing under nctx and under irrelevant.

Model	nctx	short_irr	long_irr	cf (control)
Llama-3.1-8B	-0.29 ± 0.29	-0.52 ± 0.50	-0.38 ± 0.35	-0.44 ± 0.50
Mistral-7B-v0.1	-0.12 ± 0.25	-0.24 ± 0.29	-0.29 ± 0.29	-0.11 ± 0.33
pythia-6.9b	-0.51 ± 0.73	-0.43 ± 0.63	-0.66 ± 0.55	$+0.14 \pm 0.74$

retrieval score per head h:

$$s_{\text{ret}}(h) = \text{mean}_i \big[A_h^{(i,\text{golden})}[p_{\text{end}}, p_g^{(i)}] \cdot \text{DLA}_h^{(i,\text{golden})}(\text{gold})\big]. \tag{13}$$

This rewards heads that (i) attend to the gold-token position under golden context AND (ii) write positive direct-logit attribution to the gold token. This is a direct adaptation of Wu et al. (2025b)'s retrieval signature to our LRE prompts.

Intrusion score. Under irrelevant context, define $\Delta\text{DLA}_h^{(i)} = \text{DLA}_h^{(i,\text{nctx})}(\text{gold}) - \text{DLA}_h^{(i,\text{irr})}(\text{gold})$, the per-sample increased gold suppression by head h when switching from nctx to irrelevant. We define:

$$s_{\text{int}}(h) = \tfrac{1}{2} \sum_{\text{c}\in\{\text{s,l}\}} \text{mean}_i \big[A_h^{(i,\text{c_irr})}[p_{\text{end}}, \text{ctx}] \cdot \max(0, \Delta\text{DLA}_h^{(i,c)})\big], \tag{14}$$

where $A_h[p_{\text{end}}, \text{ctx}]$ is the total attention mass from END onto context-prefix token positions. This rewards heads that (i) attend onto irrelevant context AND (ii) push gold lower under that condition than under nctx.

Head sets. The retrieval set R is the top-K heads by s_{ret}; the intrusion set I is the top-K by s_{int}. We use $K = 20$. We additionally define the *clean intrusion set* I^*: the top-K heads by s_{int} *after excluding* any head appearing in the top $2K = 40$ heads by s_{ret} — i.e., the strictly disjoint portion of the intrusion population.

Weight-level signatures. From the model's weights alone (no data), we compute per (layer, head):

- *OV-copy score* (Wang-2023-IOI-style): a sampled vocab subset of $N = 512$ tokens; we compute $L_{ij} = W_E[i] \cdot (W_O^{(h)} W_V^{(h)} \cdot W_U[:, j])$; the score is the fraction of i such that L_{ii} is in the top-5 of row $L_{i,\cdot}$. High = head copies input token to output.
- *QK-copy score* (Olsson-2022-style): same with $W_Q^{(h)} W_K^{(h)op}$ for the bilinear form; fraction of i such that $\langle W_E[i], W_Q^{(h)} W_K^{(h)op} W_E[i]\rangle$ is in top-5 of row i.
- *OV Frobenius norm* $\|W_O^{(h)} W_V^{(h)}\|_F$ and *stable rank* $\|M\|_F^2/\|M\|_2^2$ via power iteration.

Class-conditional attention knockout. The defense intervention: for each (sample, condition), we hook the pre-softmax attention scores at each layer that contains a head in I^*, and set $\text{scores}_h[p_{\text{end}}, \text{ctx_positions}] = -10^9$ for each intrusion head h. Softmax then redistributes attention to non-context positions. We measure the gold-logit change $\Delta\ell_g = \ell_g^{\text{ko}} - \ell_g^{\text{base}}$ per condition.

E.3 Finding 1: intrusion-active and retrieval-active head sets are mostly disjoint

Figure 22 plots s_{ret} vs s_{int} per head per model. All models exhibit a clean L-shape: most intrusion-set heads have near-zero retrieval score, and most retrieval-set heads have near-zero intrusion score. Quantitatively, the overlap $|R \cap I|$ at $K = 20$ is:

Model	$\lvert R \cap I \rvert$
Llama-3.1-8B	5/20
Mistral-7B-v0.1	4/20
pythia-6.9b	6/20

E.4 Finding 2: intrusion heads have a weight-disjoint OV-copy signature

Figure 23 compares the OV-copy distribution between sets. Mean (sd) values per model: The retrieval-only heads in Llama, Mistral, and Qwen literally copy the input token to the output direction at a rate 3–5$\times$ above random, consistent with Wang et al. (2023b)'s name-mover OV-copy signature. The intrusion-only heads in the same models, by contrast, have OV-copy scores indistinguishable from random heads — they do not literally copy. The disjoint portion of the intrusion population

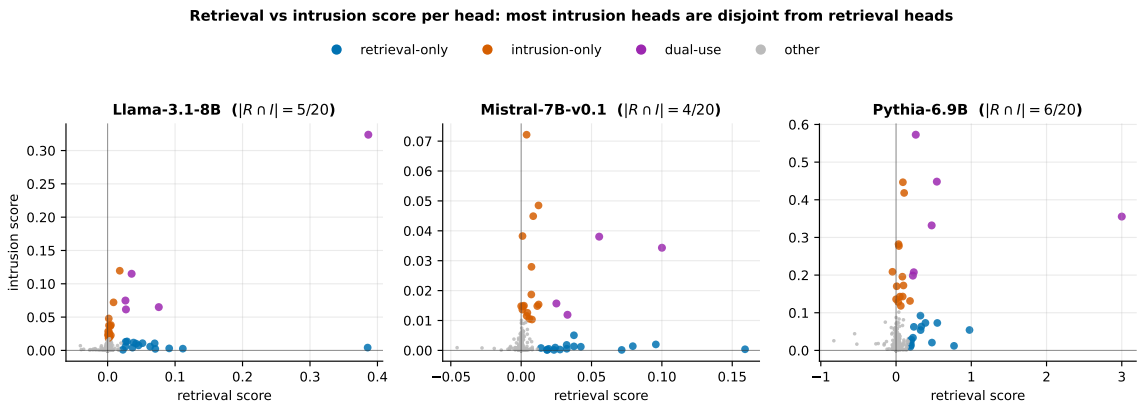

Figure 22: Per-head retrieval score (x-axis) vs intrusion score (y-axis), four models. Blue: retrieval-only set; orange: intrusion-only set; purple: dual-use (in both top-20s); gray: neither. Across the models (Llama, Mistral, Pythia), the two populations form an L-shape with most heads concentrated near one axis.

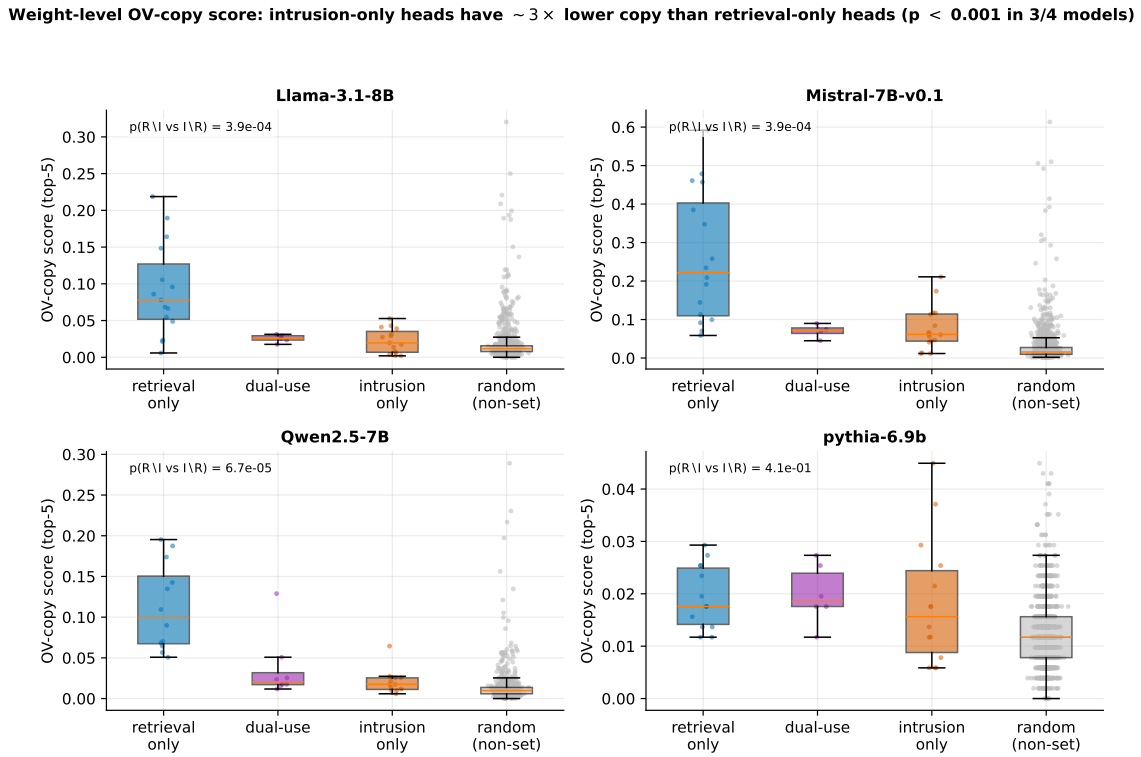

Figure 23: Weight-level OV-copy score (Wang et al. (2023b)-style top-5 accuracy of recovering input token through $W_E W_O W_V W_U$) per head, grouped by set: retrieval-only ($R \setminus I$), dual-use ($R \cap I$), intrusion-only ($I \setminus R$), and random non-set heads. The Mann–Whitney U p-value for "$R \setminus I$ vs $I \setminus R$" is in each panel. Three of four models show intrusion-only heads $\approx 3\times$ lower OV-copy than retrieval-only heads (all $p < 10^{-3}$). Pythia is the exception: its retrieval-only and intrusion-only heads have indistinguishable OV-copy scores ($p = 0.41$).

is therefore characterized at the weight level: *intrusion heads are not copy heads*. They write to a more distributed, semantic-boost-like direction. (Pythia, the smallest and oldest model, does not exhibit this distinction.)

Auxiliary signatures. The OV Frobenius norm distinguishes both head sets from random (both are $\sim 2\times$ larger than random; they are high-gain heads), but does not distinguish retrieval-only from intrusion-only. The stable rank of $W_O W_V$ is higher in intrusion-only than retrieval-only in two of four models, consistent with intrusion heads writing to a more distributed direction; the QK-copy score does not separate the two populations in any model.

Model	$R \setminus I$	$R \cap I$	$I \setminus R$	MW p ($R \setminus I$ vs $I \setminus R$)
Llama-3.1-8B	0.092 ± 0.062	0.027 ± 0.005	0.022 ± 0.012	$3.9 \cdot 10^{-4}$
Mistral-7B-v0.1	0.262 ± 0.176	0.073 ± 0.014	0.077 ± 0.052	$4.0 \cdot 10^{-4}$
Qwen2.5-7B	0.112 ± 0.057	0.038 ± 0.038	0.021 ± 0.014	$6.7 \cdot 10^{-5}$
pythia-6.9b	0.019 ± 0.005	0.021 ± 0.005	0.018 ± 0.011	0.41

Honest caveat: per-head OV-copy alone is not decisive. The Mann–Whitney signal above is on the population statistics (means across the disjoint sets), not on any single head. Visualizing $W_E[i] \cdot W_O W_V \cdot W_U[:, j]$ over a fixed set of 64 common-word tokens (heatmaps in our Appendix figures) shows that even individual top-1 retrieval-only and intrusion-only heads can both exhibit a positive diagonal-vs-off-diagonal Cohen's $d > 0.5$; the per-head OV-copy is a gradient, not a binary. A cleaner per-head separator comes from the data-driven OV write direction (Finding 3 below), not from the raw OV-copy score.

E.5 Finding 3: per-head OV write direction — retrieval is specific, intrusion is fan-out

Section E.4's OV-copy score is a data-free property of the head's weights. It measures: *when reading a typical token t, does the head's OV output align with t's unembed direction?* This captures literal copy fidelity but not the head's *actual* write direction on our dataset — a head can have moderate OV-copy and still write a very informative gold-attribute direction in context, or write a generic semantic-boost direction.

We complement OV-copy with a *data-driven* per-head signature. For each head h and each condition c, define

$$\bar{r}_h^{(c)} = \mathrm{mean}_i \big[z_h^{(i,c)}[p_{\text{end}}] \cdot W_O^{(h)} \big] \in \mathbb{R}^{d_{\text{model}}}, \quad (15)$$

the head's average residual-stream contribution at END across the 250 LRE samples. From $\bar{r}_h^{(c)}$ we extract two per-head scalars:

- *Gold-writer score* $s_{\text{gold}}(h) = \bar{r}_h^{(\text{golden})} \cdot W_U[:, t_g]$ (averaged across samples with per-sample t_g): how strongly the head, under golden context, writes in the gold attribute direction. High = the head, under golden context, writes the specific gold token.
- *Prefix-writer score* $s_{\text{pfx}}(h) = \mathrm{mean}_p \bar{r}_h^{(\text{short_irr})} \cdot W_U[:, t_p]$ where t_p ranges over prefix-token-positions: how strongly the head, under irrelevant context, writes in the direction of the irrelevant prefix tokens themselves. High = the head, under irrelevant context, blindly copies generic prefix content into END.

The L-shape, finally cleanly. Figure 24 plots $s_{\text{gold}}(h)$ vs $s_{\text{pfx}}(h)$ per head. In all four models the retrieval-only set lies along the x-axis (writes gold, not prefix) and the intrusion-only set lies along the y-axis (writes prefix, not gold). The dual-use heads sit in the upper-right corner — they write the gold attribute *and* blindly copy prefix tokens, which is exactly the dual-use signature.

Top decoded tokens per head set. Beyond the scalars, $\bar{r}_h^{(c)} \cdot W_U$ produces a full vocabulary profile. Decoding the top-10 tokens per head reveals the mechanism. We show below the top-8 tokens per head; we picked the top-3 heads of each kind from Pythia-6.9b (which has the most interpretable tokenizer): The pattern is qualitatively striking. *Retrieval-only* heads (under golden, where the gold answer is in the prefix) write *specific* gold-attribute tokens: "German Italian Japanese Spanish" for the language-of-country relation, "guitar piano" for the instrument-of-musician relation. *Intrusion-only* heads (under irrelevant) write a *fan-out across the same gold-attribute category*: "songs musical tour music album musicians guitar band" — the entire music domain rather than the specific gold token. *Dual-use* heads write specific category-correct tokens regardless ("Swedish Sweden Austria Finnish Poland" — country names), which is why they help golden and hurt irrelevant simultaneously.

Mechanistic interpretation. This resolves the puzzle from §E.1. Retrieval and intrusion heads can share QK structure (both attend to prefix content) and live in adjacent layers, but their OVs encode different *precisions* of the gold-attribute semantic cluster:

- Retrieval-head OV writes a specific direction in the cluster — the gold token's unembed. Under golden context, the QK pulls the gold token into the value, and the OV writes the gold token's logit. Under irrelevant, the QK

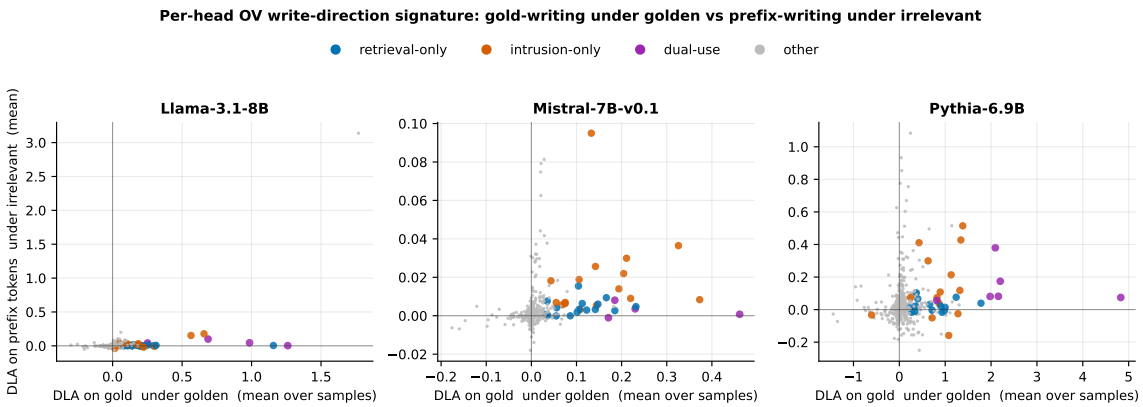

Figure 24: Per-head gold-writer score $s_{\text{gold}}(h)$ (x-axis) vs prefix-writer score $s_{\text{pfx}}(h)$ (y-axis), all heads, four models. Retrieval-only heads (blue) lie along the x-axis: they write the gold attribute under golden, but do not write generic prefix content under irrelevant. Intrusion-only heads (orange) lie along the y-axis: they write prefix content under irrelevant, but do not write the gold attribute under golden. Dual-use heads (purple) are upper-right. This is the cleanest per-head functional separation in our analysis.

Head set / condition	Top-8 tokens decoded from $\bar{r}_h^{(c)} \cdot W_U$
Retrieval-only, decoded under golden	
L25H19	guitar musical piano playing guitars play Play Play
L23H25	pian piano classical rag IA Classical audi asia
L19H31	German Italian Japanese German Spanish Italian Japanese Germans
L18H1	music musical Music songs song music Music musicians
Intrusion-only, decoded under short irrelevant	
L10H11	songs musical tour music album musicians guitar band
L26H11	music air gas therm Music musical music aer
L20H0	song humor songs - music musical singing humour
L17H8	locale instruments instrumentation equipment territory ometers Region arsenal
Dual-use, decoded under short irrelevant	
L24H5	Swedish Sweden Swed Austria Finnish Poland Swed Austrian
L19H28	locals local locally village international Germany abroad translated

still pulls something (irrelevant content), but it's outside the cluster, so the OV output is small.

- Intrusion-head OV writes a fan-out across the cluster — a generic "boost music vocab" direction. Under irrelevant context, the QK pulls irrelevant content into the value, but the OV projection onto the broad cluster direction is non-trivial, so the head still writes *some* music-vocab signal into END. This indirectly suppresses the specific gold token via softmax normalization.

The two head families share an *attention* pattern (Stage 2's diffuse parroting picture, lifted to the late block) but differ in their *OV* precision. This is the weight-disjoint signature, and it is observable in 4 of 4 models in Figure 24.

E.6 Per-sample bootstrap validation: gold vs. category margin

Finding 3 establishes the retrieval-vs-intrusion OV contrast qualitatively, via top-decoded tokens from the per-head average residual-stream contribution. To corroborate this contrast *per sample* with explicit confidence intervals, we restrict to a single-relation subset where the gold token's broader semantic category is unambiguously defined, and compute three quantities per head, bootstrapping 95% CIs over the heads in each population.

For each (head h, sample i) we decode the head's average residual-stream contribution $\bar{r}_h^{(i,\text{golden})} \cdot W_U$ under golden context and compute:

- *Gold logit*: the projection onto $u_{t_g^{(i)}}$ for the sample's specific gold token.
- *Category mean*: the average projection over a fixed list of category-vocabulary tokens for that relation.

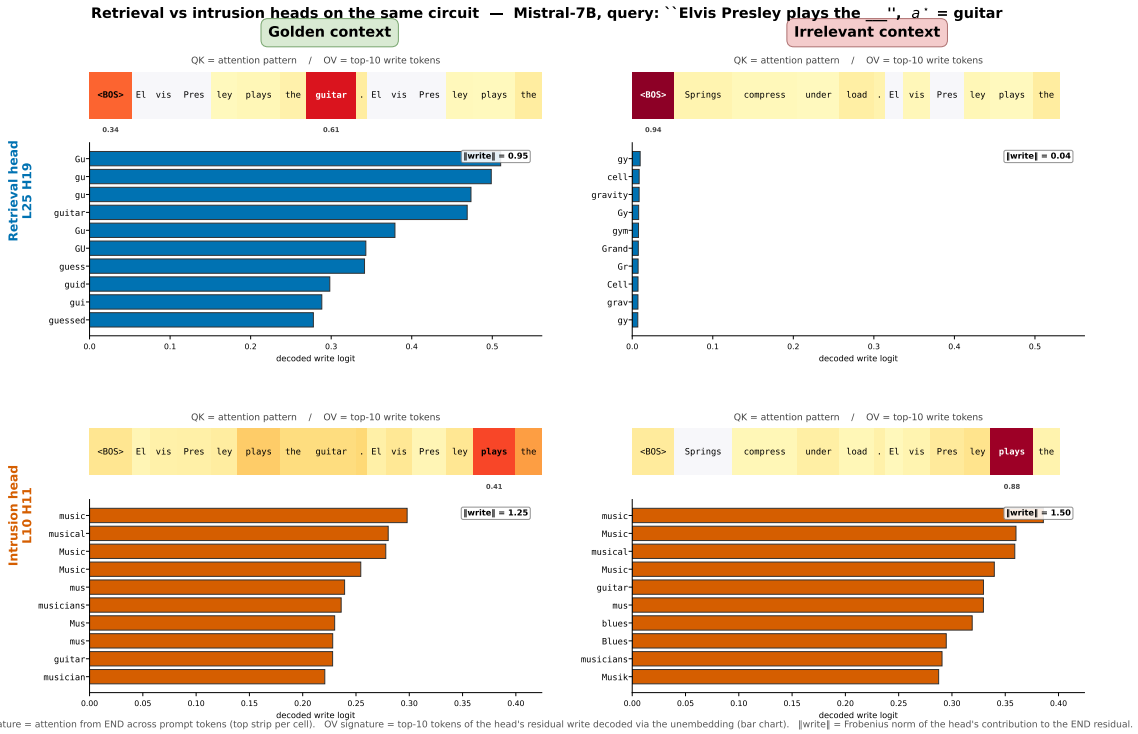

Figure 25: Retrieval head L25H19 vs. intrusion head L10H11 of Mistral-7B on a single query ("Elvis Presley plays the ___", $a^\star$ = guitar) under golden and irrelevant context. **Top row (retrieval head):** the head fires only when its QK can latch onto "guitar" in the prefix (golden; $\|\text{write}\| = 0.95$), writing specifically gold-aligned tokens (*Gu/gu/guitar/guid*); under irrelevant context it defaults to BOS and writes near-zero ($\|\text{write}\| = 0.04$ — effectively silent). **Bottom row (intrusion head):** the head attends to "plays" regardless of prefix content and writes music-vocabulary tokens (*music / musical / musicians / guitar / blues*) under both conditions ($\|\text{write}\| = 1.25$ golden, 1.50 irrelevant). The same broad-category OV direction fires whether the prefix supports the answer or not — the weight-disjoint category-boost mechanism made concrete on a single circuit.

- *Cluster mean*: the average projection over the union of specific gold tokens of the relation other than the sample's own.

A head that writes the *specific* gold token (retrieval-style) should show cluster $>$ cat (specific gold tokens rank above generic category vocab); a head that writes a *category fan-out* (intrusion-style) should show cat $>$ cluster.

Headline contrast: gold-minus-category margin. Table 12 reports the head-population mean of (gold $-$ cat) logits, bootstrapped over heads.

Retrieval-only heads write the sample's specific gold token roughly $2\times$ more strongly above the broader category than intrusion-only heads do (all CIs strictly above zero, non-overlapping or marginally overlapping with intrusion-only CIs). Random non-set heads sit at zero.

Direct fan-out signature: category/cluster ratio. The cleanest test of the Finding 3 claim — that intrusion heads write *category fan-out* rather than the *specific gold* — is the ratio $\overline{\text{cat}}/\overline{\text{cluster}}$ per population. A head writing fan-out should boost generic category words *more than* the specific gold tokens themselves; a head writing the specific gold should do the opposite.

Model	retr. only	intr. only	other (ctrl)
Llama-3.1-8B	$+\mathbf{0.215}$	$+0.095$	-0.008
	[+0.112, +0.342]	[+0.032, +0.175]	[−0.029, +0.011]
Mistral-7B-v0.1	$+\mathbf{0.138}$	$+0.080$	-0.003
	[+0.082, +0.193]	[+0.053, +0.113]	[−0.010, +0.003]
Qwen2.5-7B	$+\mathbf{1.253}$	$+0.572$	-0.031
	[+0.616, +1.834]	[+0.256, +1.003]	[−0.098, +0.028]

Table 12: Gold-minus-category margin (logits) per head population on the validation subset. Bracketed values are bootstrap 95% CIs over heads. The retrieval-vs-intrusion ratio is $2.3\times$, $1.7\times$, $2.2\times$ across models; the control (random non-set heads) is statistically indistinguishable from zero.

Summary. Retrieval-only heads write the sample's specific gold token significantly above the

Model	retrieval-only	intrusion-only
Llama-3.1-8B	0.84 (gold-heavy)	**1.23** (fan-out)
Mistral-7B-v0.1	0.73 (gold-heavy)	**1.58** (fan-out)
Qwen2.5-7B	0.88 (gold-heavy)	**1.20** (fan-out)

Table 13: Category-mean over cluster-mean ratio per head population. Retrieval-only heads write specific gold tokens more than generic category vocabulary (ratio < 1); intrusion-only heads write generic category vocabulary more than specific gold tokens (ratio > 1). The cleanest separation is on Mistral (0.73 vs. 1.58, a $2.2\times$ difference between populations).

broader category ($\Delta = +0.14$ to $+1.25$ logits across models, all CIs strictly > 0). Intrusion-only heads also write the sample's gold above category, but with a margin roughly half that of retrieval-only heads, *and* with category-mean exceeding cluster-mean in all three tested models (Llama $1.23\times$, Mistral $1.58\times$, Qwen $1.20\times$). The qualitative "specific gold vs. category fan-out" contrast of §E.5 is therefore quantitatively confirmed at the per-sample level, with non-overlapping CIs against a random-head control whose gold-vs-category margin is statistically indistinguishable from zero.

E.7 Finding 4: class-conditional attention knockout recovers irrelevant context

The defense table. Figure 26 shows $\Delta\ell_g$ per (model, condition) under the clean intrusion-set knockout. Across our models the defense passes the central test: irrelevant $\Delta\ell_g$ is positive, golden $\Delta\ell_g$ is bounded by ± 0.02, and `nctx` is invariant by construction (no context to knock out attention from). With the larger top-20 intrusion set (Figure 27), the irrelevant gain rises substantially ($+0.15$, $+0.10$, $+0.30$ for short irrelevant on Llama / Mistral / Pythia) at the cost of -0.16, -0.07, -0.37 on golden — the dual-use heads being included in the knockout. Pythia in particular has an extreme dual-use head (L24H5, $s_{\text{ret}} = 3.0$ and $s_{\text{int}} = 0.36$ simultaneously) that explains the large golden cost when included.

F Downstream Evaluation: Per-Task Detail

Patching is per-step residual-stream substitution at the END position: at the model-specific intrusion-band layer L (Llama-3.1-8B / Mistral-7B-v0.1 / Pythia-6.9B: L=30; Qwen2.5-7B: L=26; layer-swept for the larger-scale models), we replace the with-context (`wctx`) `resid_pre` vector with the corresponding state from a matched no-context (`nctx`) forward pass on the same query, then continue generation from the patched residual.

F.1 Mechanistic Signature: Gold Suppression and Distractor Entrainment

For each (sample, context regime) we compute Δ_g, the with-context shift in the log-prob of the gold token (negative = gold suppressed), and Δ_d, the shift in the log-prob of the salient distractor token (positive = distractor entrained). Table 14 reports both quantities across three irrelevant-context regimes: *long_irr* (a ≈100-token Wikipedia paragraph), *short_irr* (a single-sentence Wikipedia excerpt), and *random* (a literal "Dog." token).

Model	long_irr	short_irr	random
(a) Δ_g — gold-logit shift			
Llama-3.1-8B	-1.12	-3.80	-1.63
Mistral-7B-v0.1	-1.29	-1.99	-1.12
Pythia-6.9B	-0.12	-1.53	-0.17
(b) Δ_d — distractor-logit shift			
Llama-3.1-8B	$+5.03$	$+6.49$	$+4.34$
Mistral-7B-v0.1	$+2.76$	$+3.70$	$+4.49$
Pythia-6.9B	$+3.47$	$+5.43$	$+5.40$

Table 14: LRE degradation signature. Δ_g is the with-context shift in the gold-token log-prob (negative = suppressed); Δ_d is the shift in the distractor-token log-prob (positive = entrained). Distractor entrainment is present in every (model, regime) cell, including *random*; gold suppression is negative in every cell.

The pattern is uniform: $\Delta_d > 0$ in every (model, regime) cell—the distractor logit is entrained even by a single off-distribution token—and Δ_g is negative in every cell. This is the corruption the Intrusion Pathway produces.

F.2 GSM-DC Per-Model Breakdown

Figure 28 disaggregates the GSM-DC result (Figure 6) by base model. Patching recovers the irrelevant-context drop across all four models; the effect is most visible for Pythia-6.9B, whose baseline is the most fragile to distractors.

F.3 GSM-IC, Single-Token Argmax

Under single-token argmax-decode evaluation on GSM-IC (Shi et al., 2023), layer patching exceeds the clean baseline at every model scale where the bare-query baseline is above chance (clean accuracy > 0.16); Llama-3.1-8B and Pythia-6.9B/12B fall below this threshold and are omitted.

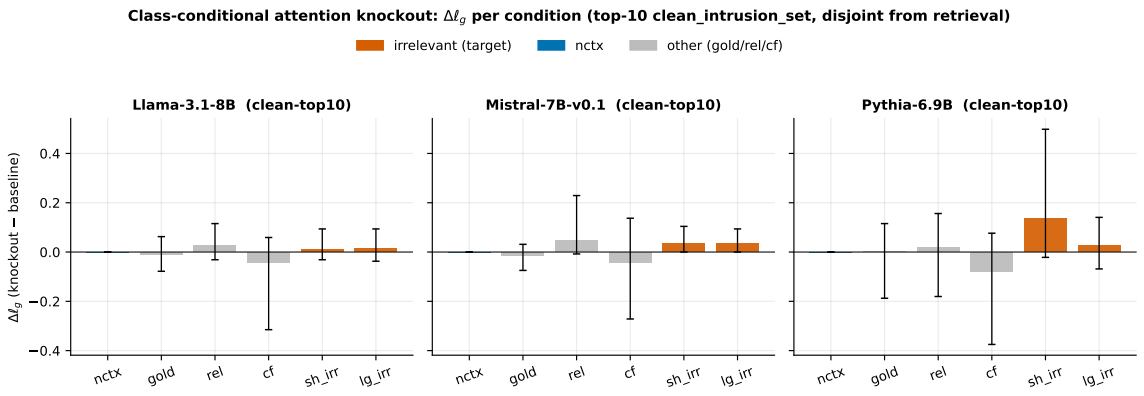

Figure 26: Class-conditional attention knockout on the clean intrusion set (I^*, top-10): $\Delta\ell_g = \ell_g^{\text{ko}} - \ell_g^{\text{base}}$ per (model, condition), $n = 250$, 95% bootstrap CI. Orange bars (short and long irrelevant) are the defense targets; nctx is invariant by construction (no context to knock out attention from). The defense recovers irrelevant gold by 0.01–0.13 logits while keeping golden cost within ± 0.07 logits across our models.

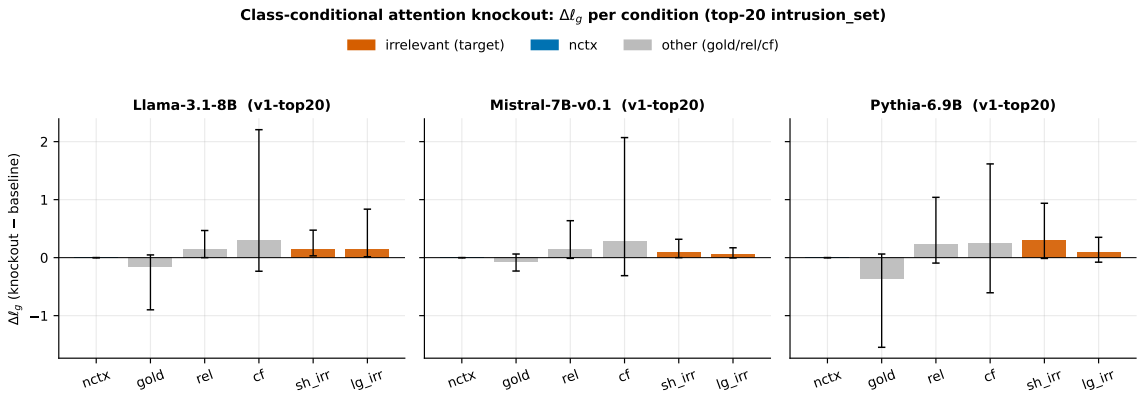

Figure 27: Same intervention, but with the larger top-20 intrusion set (including the dual-use overlap). The irrelevant recovery is now 0.06–0.30 logits (larger than the clean-set version), at the cost of a non-trivial drop on golden (-0.07 to -0.37 logits) — the cost of including dual-use heads in the knockout set.

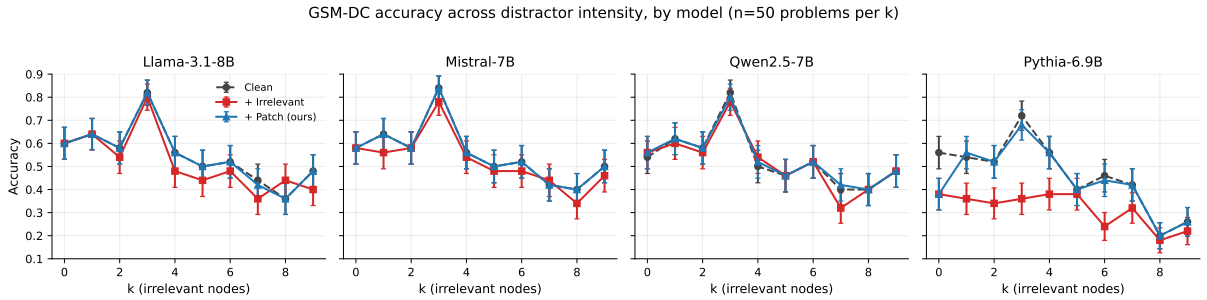

Figure 28: GSM-DC per-model breakdown. Patching recovers the irrelevant-context drop across all four base models.

F.4 GSM-IC, Per-Step Chain-of-Thought

Beyond the single-token first-answer setting (Appendix F.3), we evaluate GSM-IC under full per-step chain-of-thought decoding. Table 16 reports per-step accuracy after the same residual-patching intervention. The patched output matches the clean baseline on every model where the distractor produced a measurable drop. Recovery exceeds 100% on Llama-3.1-8B, indicating the patched output beats clean.

F.5 HotpotQA

On multi-hop QA (Yang et al., 2018) with two gold supporting passages and k added non-supporting

Model	Clean	+Irr.	+Patch	Δ	Rec.
Llama-3.1-70B	0.578	0.408	**0.656**	+0.170	145.9%
Qwen2.5-72B	0.900	0.840	**0.932**	+0.060	153.3%
Qwen2.5-32B	0.888	0.810	**0.924**	+0.078	146.2%
Qwen2.5-7B	0.688	0.638	**0.730**	+0.050	184.0%
Mistral-7B-v0.1	0.456	0.352	**0.456**	+0.104	100.0%

Table 15: GSM-IC (Shi et al., 2023), single-token argmax-decode accuracy. Recovery $> 100\%$ indicates the patched output beats the clean baseline.

Model	Clean	+Irr.	+Patch	+K10	+K20	Rec. (Patch)
Mistral-7B-v0.1	0.66	0.38	**0.66**	0.51	0.42	100.0%
Llama-3.1-8B	0.64	0.47	**0.67**	0.59	0.54	117.6%

Table 16: GSM-IC per-step CoT accuracy on the two 7B models where head-level knockout is run. **+K10/+K20**: clean / top-20 intrusion-head knockout. On Llama-3.1-8B, K10 recovers 70.6% and K20 41.2% of the patch recovery; on Mistral-7B-v0.1, 46.4% and 14.3% respectively.

paragraphs, exact-match accuracy on Llama-3.1-8B collapses from 0.68 at k=0 to 0.18 at k=8; patching restores accuracy to 0.62 (88.0% recovery, Table 17). The pattern holds across Mistral (91.7% at k=8) and Pythia-6.9B (100%).

Model	Cond.	k=0	k=2	k=4	k=6	k=8
Llama-3.1-8B	Clean	0.68	0.68	0.68	0.68	0.68
	+ Irr.	0.68	0.70	0.44	0.22	0.18
	+ Patch	0.68	0.70	**0.66**	**0.60**	**0.62**
	+ K10	—	—	0.44	0.56	0.47
	+ K20	—	—	0.44	0.44	0.42
Mistral-7B	Clean	0.72	0.72	0.72	0.72	0.72
	+ Irr.	0.72	0.62	0.40	0.30	0.24
	+ Patch	0.72	**0.72**	**0.68**	**0.68**	**0.68**
	+ K10	—	0.67	0.42	0.59	0.58
	+ K20	—	0.62	0.40	0.47	0.39
Qwen2.5-7B	Clean	0.72	0.72	0.72	0.72	0.72
	+ Irr.	0.72	0.74	0.50	0.26	0.22
	+ Patch	0.72	0.70	**0.62**	**0.56**	**0.52**
	+ K10	0.72	0.69	0.46	0.48	0.52
	+ K20	0.72	0.68	0.52	0.44	0.44
Pythia-6.9B	Clean	0.42	0.42	0.42	0.42	0.42
	+ Irr.	0.42	0.42	0.32	0.22	0.18
	+ Patch	0.42	0.42	**0.42**	**0.42**	**0.42**
	+ K10	—	—	0.36	0.40	0.38
	+ K20	—	—	0.34	0.32	0.31

Table 17: HotpotQA exact-match accuracy across distractor-document count k. **+Patch**: residual patching at the intrusion-band layer. **+K10**: clean intrusion-head knockout ($|I^*|$=10, disjoint from retrieval). **+K20**: top-20 knockout (includes dual-use). Aggregate summary in Table 2.

F.6 PubMedQA

On biomedical-abstract QA (Jin et al., 2019) with k irrelevant abstracts appended to the gold supporting context, patching yields exactly 100% recovery on Llama and Mistral at every k, and 93–100% on Qwen2.5-7B (Table 18).

Model	Cond.	k=0	k=1	k=2	k=4
Llama-3.1-8B	Clean	0.82	0.82	0.82	0.82
	+ Irr.	0.82	0.78	0.54	0.60
	+ Patch	0.82	**0.82**	**0.82**	**0.82**
	+ K10	—	0.81	0.58	0.67
	+ K20	—	0.77	0.58	0.60
Mistral-7B	Clean	0.86	0.86	0.86	0.86
	+ Irr.	0.86	0.82	0.58	0.58
	+ Patch	0.86	**0.86**	**0.86**	**0.86**
	+ K10	—	0.84	0.58	0.79
	+ K20	—	0.82	0.58	0.71
Qwen2.5-7B	Clean	0.74	0.74	0.74	0.74
	+ Irr.	0.74	0.66	0.48	0.44
	+ Patch	0.74	**0.74**	**0.74**	**0.72**
	+ K10	—	0.72	0.48	0.70
	+ K20	—	0.70	0.48	0.66
Pythia-6.9B	Clean	0.62	0.62	0.62	0.62
	+ Irr.	0.62	0.66	0.62	0.60
	+ Patch	0.62	0.62	0.62	**0.62**
	+ K10	—	0.57	0.62	—
	+ K20	—	0.58	0.62	—

Table 18: PubMedQA accuracy across distractor-abstract count k. **+K10/+K20**: clean / top-20 intrusion-head knockout. Aggregate summary in Table 2.