THE INTRUSION PATHWAY: HOW NON-EVIDENTIAL CONTEXT DEGRADES LANGUAGE MODEL PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models lose accuracy when a prompt contains non-evidential context: irrelevant statements from unrelated domains, or random off-distribution tokens, with no competing answer and no semantic overlap with the query. Such noise pervades retrieval-augmented, tool-using, and long-context deployments, yet the failure has lacked a mechanistic account. We trace the corruption to the *Intrusion Pathway*, which starts with MLPs enriching both query-subject and irrelevant-token representations, parroting heads carry prefix-token content to the prediction position, and relational readers largely avoid unrelated sources. Nevertheless, preserved subject enrichment and selective reading do not prevent intrusion in the broader computation. We then localize part of this interference to a sparse set of attention heads, which we call *intrusion heads*. Their prefix-derived writes promote distractor-related content, while their reduced support for the correct answer arises primarily from weaker non-prefix contributions rather than direct suppression by the prefix. The same heads transfer across tasks and distractor types and are largely weight-disjoint from the heads that integrate relevant contextual evidence, enabling a class-conditional intervention. Replacing the corrupted last-token residual with its clean counterpart, an oracle probe that requires a matched clean forward pass, recovers 97.6% on average of the distractor-induced drop across four downstream tasks. A deployable mask level knockout of about ten heads, roughly 1% of the model's attention heads, recovers 57% at bounded cost to helpful context.

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), tool-call results from a browsing or search API (Schick et al., 2023), conversation snippets, and traces from agentic sub-loops (Yao et al., 2023). 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), 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; Niu et al., 2025; Fang et al., 2024; 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); its extraction machinery is the best-mapped circuit in the field (Wang et al., 2023; McDougall et al., 2023). *Counterfactual* context asserts a coherent competing answer via the query's own relation; it is the substrate of the knowledge-conflict literature (Longpre et al., 2021) and has been localized at every grain, from mid-layer FFNs to individual memory-vs-context heads. *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 of induction heads and task/function vectors (Olsson et al., 2022; Hendel et al., 2023). *Irrelevant* context is a coherent statement from an unrelated domain—no shared entity, relation, or topic with q—and contains one or more salient

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Regime	Context	Gold-logit shift
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."	−1.1
Irr. (multi-token)	"Computers perform ... efficiently."	−2.3
Random	"Calculator."	−0.2

Table 1: Context regimes for the running query "The capital of Germany is ___" ($a^\star$ = Berlin). The **Gold-logit shift** column shows the gold token (Berlin) logit change caused in nats by that regime. "Irr." abbreviates Irrelevant; the full multi-token prefix is "Computers perform mathematical simulations efficiently."

non-gold (distractor) tokens. Such context can substantially reduce answer accuracy (Shi et al., 2023; Wu et al., 2024). *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).

Among these five classes, irrelevant context is the least understood mechanistically. Shi et al. (2023) document the accuracy loss it causes, and Niu et al. (2025) identify heads responsible for contextual entrainment, the promotion of tokens that appear in the context. Consider again "The capital of Germany is," whose correct continuation is "Berlin." Prepending "Computers are useful" neither supports this answer nor proposes another capital, yet it lowers the gold token's logit. Promoting a distractor can reduce the gold token's probability through competition, but does not by itself establish why the gold token's own logit decreases. We therefore ask two questions: how does content that supplies no evidence about the answer reduce support for the gold token, and can the components that carry this interference be separated from those that let the model use golden context? Answering these questions would clarify why irrelevant context interferes with transformer computation and how much robustness can be recovered through a sparse, training-free intervention, rather than relying on noise-aware training or additional context-handling procedures (Fang et al., 2024; Yoran et al., 2024; Asai et al., 2024; Yu et al., 2024).

We answer the first question by tracing how context is represented and propagated, then localizing the writes associated with lost answer support. In §3, our analysis of the **Intrusion Pathway** follows MLP enrichment, parroting, and relational reading. The measured enrichment of the query subject is preserved under irrelevant context, while irrelevant-token representations are also enriched. Parroting heads carry prefix-token content to the last position, and relational readers largely avoid unrelated sources. Nevertheless, restoring the broader middle block's writes recovers answer support while also removing benefits from golden context, identifying an effective but non-selective intervention site.

In §4, we identify **intrusion heads**, whose attention to irrelevant context coincides with reduced direct support for the correct answer. Their prefix-derived writes promote distractor-related content, whereas their reduced gold support comes primarily from weaker contributions from the query. This localizes a component of the interference to changes in the heads' source contributions, rather than treating copying alone as the mechanism of answer degradation.

For the second question, intrusion heads are largely distinct from golden-context retrieval heads (Wu et al., 2025b; §4). Masking prefix attention in a fixed, retrieval-disjoint set of about ten heads per model, roughly 1%, recovers 57% of distractor-induced accuracy loss on average across four downstream tasks while largely preserving golden-context benefits. The set is selected once on factual recall, and the mask requires no matched clean forward pass at inference. Oracle replacement of the last-position residual with its matched clean counterpart recovers 97.6% on average (§5).

2 Experimental setting and background

For next-token circuit analysis, we compare a query q alone (nctx) with the prefixed input $c \parallel q$, holding the query, scored gold token g, and model fixed. **Task data.** The analyses for intrusion is characterized on two circuit benchmarks—factual recall and binding/retrieval. Factual recall uses LRE source–target pairs rendered through relation-specific templates (Hernandez et al., 2024). Bind-

ing/retrieval uses indirect object identification (IOI) templates (Wang et al., 2023). Irrelevant prefixes are approximately 12-token sentences (`short_irr`) or approximately 100-token WikiText-103 paragraphs (`long_irr`; Merity et al., 2016). Long paragraphs are filtered for subject, answer, and related lexical or semantic overlap. Appendix A provides construction and filtering details. **Models.** We use five open-weight base models—Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.1 (Jiang et al., 2023), Qwen2.5-7B (Qwen Team, 2024), Pythia-6.9B (Biderman et al., 2023), and GPT-2 XL (Radford et al., 2019)—through TransformerLens (Nanda & Bloom, 2023). Downstream evaluation is described separately. **Metrics.** Let $\ell_t(q)$ and $P(t \mid q)$ denote token t's logit and next-token probability at the last prompt position. We measure the gold-logit shift $\Delta\ell_g(c; q) = \ell_g(c \,\|\, q) - \ell_g(q)$ and token log-probability shift $\Delta_t(c; q) = \log P(t \mid c \,\|\, q) - \log P(t \mid q)$. Negative $\Delta\ell_g$ indicates that the prefix lowers the gold token's logit; negative Δ_g indicates that it lowers the probability assigned to that token. For a designated distractor token d appearing in the prefix, positive Δ_d indicates distractor entrainment. A decrease in gold probability can also result from increased competing logits, even when the gold logit is unchanged. We report paired changes averaged over the full dataset for each task, model, and evaluated prefix condition. Downstream, we report accuracy and recovery: the percentage of distractor-induced accuracy loss restored by an intervention.

Notation. END is the last prompt position, and $x_p^{(L)}$ is the residual-stream vector at position p, layer L. Let $u_t = W_U[:, t]$ be token t's unembedding direction. Head h's QK circuit sets attention weights $\alpha_p = A_h[\mathrm{END}, p]$, while its OV circuit writes

$$o_h = \sum_p \alpha_p W_O^{(h)} W_V^{(h)} x_p^{(L)}, \tag{1}$$

with normalization and bias terms suppressed. We use the raw projection $\mathrm{DLA}_h(t) = \langle o_h, u_t \rangle$ for direct logit attribution, measuring directional alignment rather than an intervention-based causal effect.

We introduce these copying behaviours here to specify what it means for a head to write prefix-token content into the END residual, before examining the corresponding heads and their effects.

Induction versus parroting. Induction is prefix-matched copying: in the canonical pattern $[A][B] \ldots [A] \rightarrow [B]$, the repeated $[A]$ supplies a cue to attend to the earlier continuation $[B]$ and promote it as the next token (Olsson et al., 2022). Parroting instead promotes previously encountered tokens without requiring that matching-prefix cue. These terms describe behaviours, not mutually exclusive head populations: Goldowsky-Dill et al. (2023) describe a head that contributes to both. We use parroting for the antecedent-independent copying contribution studied here.

Single-source and diffuse parroting. Copying may concentrate on one token or span several, such as "Computers", "mathematical", and "simulations" in an irrelevant sentence. We call the latter *diffuse parroting*, descriptively rather than as a new mechanism (Figure 1). For prefix-content positions $\mathcal{P}_{\mathrm{content}}$, excluding BOS, stopwords, and query repetitions, define the attention-weighted identity reference and decompose the corresponding write as

$$\begin{aligned} u_*^{\mathrm{attn}} &= \sum_{p \in \mathcal{P}_{\mathrm{content}}} \alpha_p u_{t_p}, \\ o_h^{\mathrm{content}} &= c_h u_*^{\mathrm{attn}} + r_h. \end{aligned} \tag{2}$$

Here, $c_h > 0$ is a shared copy gain, and r_h is the deviation of the actual content write from the identity-only reference; it is not assumed small or orthogonal. When one source dominates the mixture, the reference reduces approximately to $c_h \alpha_{p_s} u_{t_{p_s}}$. Multiple sources need not receive equal attention. Appendix B.2.4 gives the derivation and assumptions; Appendix B.3.1 reports the held-out reconstruction test.

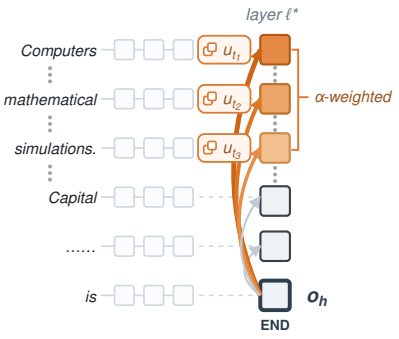


Figure 1: Diffuse parroting at layer ℓ^*: attention from END spreads over several prefix tokens, whose unembedding directions are written into the END residual o_h.

3 TRACING THE INTRUSION PATHWAY

We seek to explain how irrelevant context changes the computation of an answer already determined by the query. Recall "The capital of Germany is ___" and the two prefixes in our running example: "Computers are useful" and "Germany's capital city is Berlin." The former supplies no evidence about the answer, while the latter states it explicitly. Their different effects raise a question about how the model handles contextual information—not just which words it reads, but what it represents about them, carries to the answer position, and uses in the final prediction.

Subject and prefix enrichment. A natural place to look is the subject itself: does the sentence about computers disrupt information associated with Germany? Subject-position MLP writes make related attributes available in the subject representation (Geva et al., 2023). Their support for independently defined subject attributes, relative to another subject's attributes, provides a measure of this enrichment. These writes favor their own subject's attributes, and the paired preservation test shows that this measured signal remains effectively unchanged after an irrelevant prefix is added (Appendix B.1.1). Lower support for "Berlin" therefore need not reflect the disappearance of information associated with Germany.

Enrichment is not confined to the queried subject. MLP writes at irrelevant-token positions also favor words associated with those tokens over another distractor's associated words (Appendix B.1.2). In our running example, "Germany" and "Computers" thus differ in their relevance to the question, not in whether their representations can contain associated information.

The prefix effect also appears in the IOI-like setting. In "Springs compress under load. When Mary and John went to the store, John gave a drink to," answering the query requires identifying the participants' roles rather than recalling their factual attributes. The diagnostic detects no comparable factual-attribute enrichment at the tested name positions, but associated information still appears at the irrelevant-prefix positions. Prefix enrichment therefore also occurs when the query does not ask for a factual association.

Deleting the complete MLP writes does not isolate these associations. Cancelling subject-position writes lowers the final gold logit under both bare and irrelevant inputs, and the early-layer reduction also persists under activation replacement (Appendix B.1.3). These interventions establish the importance of the complete writes, rather than attributing their entire effect to the association signal measured within them.

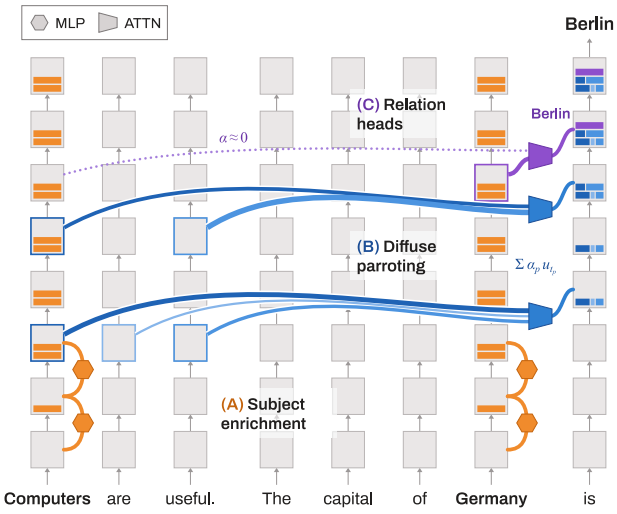


Figure 2: Irrelevant content can coexist with answer-supporting information at the prediction position: avoiding unrelated sources does not, by itself, remove their earlier contributions to the residual stream.

Parroting to END. How does information from a source such as "Computers" reach END, the position from which the answer is predicted? For copying to occur, a head must do more than attend to that source: its write must follow the source token's identity. In matched prompts, replacing the value read at one source position while holding the attention pattern fixed shifts the selected heads' writes toward the replacement token and away from the original, more strongly than in control heads. Since the attention weights remain unchanged, this identity-sensitive effect comes from the source-dependent OV write rather than altered QK routing (Appendix B.2.1).

The sentence about computers supplies no matching capital-answer pattern, but copying need not depend on such a cue. In controlled prompts, we remove the earlier matching cue while keeping the candidate copied token, its position, and the final cue fixed. Copying weakens but does not disappear (Appendix B.2.2). We use parroting for this surviving cue-independent contribution. The same heads

can also exploit an induction cue when one is present: induction and parroting describe behaviors, not mutually exclusive head populations.

Copying can also combine several sources. A prefix such as "Springs compress under load" contains several content-bearing words rather than a single candidate source. We compare the actual prefix-derived write with an attention-weighted mixture of source-token identity directions, then shuffle the pairing between tokens and attention weights while retaining the same tokens and weights. The correct pairing improves held-out reconstruction for a subset of the tested heads, supporting attention-weighted, multi-source copying (Appendix B.3.1). We refer to this behavior as diffuse parroting. The mixture captures a distributed identity component, rather than everything the head writes.

Copying "Computers" does not necessarily mean carrying all the information associated with it. A pure identity vector can already align with related words, so we compare the actual write with a gain-matched identity-only reference. Association-related output remains beyond what this reference predicts. However, the before-and-after comparisons do not establish that the OV transformation preserves the measured association profile (Appendices B.3.2–B.3.3). A copying write can therefore contain more than token identity without establishing intact transfer of the source representation.

Relational selectivity. Making information about "Computers" available at END does not, by itself, explain why "Berlin" loses support. The query also indicates what kind of information matters: "capital" calls for a city. We examine middle-layer relational readers selected for attending to answer-bearing spans under golden context, then vary the relational frame independently of the candidate answer's type. For the Germany query, this means comparing a city such as "Madrid" inside and outside the capital-answer frame, as well as placing a non-city word inside that frame. Across our analyses a candidate of the expected answer type attracts strong attention even without the matching frame and all reader populations assign little attention to the tested unrelated filler (Appendix B.4).

Middle-block interventions. Avoiding an irrelevant source is not the same as removing its influence from the answer computation (Figure 2). We therefore intervene on the broader middle block, rather than only the selected relational readers, replacing its attention heads' END-position writes with their context-free counterparts while retaining the prefixed input. Across the models and the short and long irrelevant-prefix conditions, this restores 41–90% of the gold-logit loss. Masking the same heads' END-to-prefix attention instead increases the gold logit by at most 0.17 (Appendix B.5). Blocking these direct reads does not reproduce the recovery obtained by restoring the writes (Figure 3). The restoration identifies an effective intervention site, but not a selective defence: it requires a context-free donor and also removes 45–87% of the gold-logit benefit from golden context (Table B.7). The remaining question is whether a more specific intervention can reduce the influence of "Computers are useful" without sacrificing the useful evidence in "Germany's capital city is Berlin."

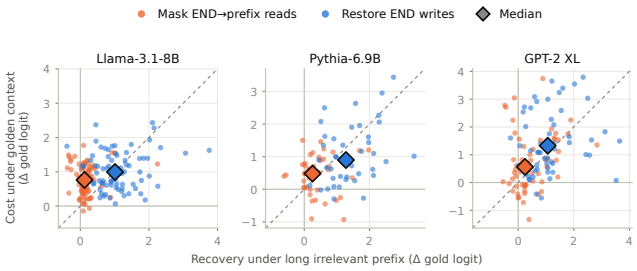


Figure 3: Per-fact recovery under the long irrelevant prefix versus cost under golden context, as gold-logit changes, for masking END→prefix reads and restoring END writes; diamonds mark medians.

The preceding analyses explain how contextual information becomes available, but do not yet isolate a selective way to remove its harmful effects. Cancelling MLP writes also removes useful answer computation. The tested parroting interventions do not reliably separate unwanted prefix contributions from helpful-context use (Appendix B.2.3). The relational readers largely avoid unrelated filler, yet consequential changes remain in the broader middle block. Restoring that block's writes reduces interference, but requires a context-free donor and also removes useful support from golden context.

These analyses trace how contextual information is represented and propagated, but do not isolate a selective defence. We next identify individual heads associated with reduced answer support and test interventions on their contextual reads.

4 INTRUSION HEADS: LOCALIZATION AND INTERVENTION

Identifying intrusion heads. A head's contextual read can change both what it adds from the prefix and what it contributes from the query. To identify the heads associated with lost answer support, we weight the reduction in each head's direct contribution to the correct-answer logit by its attention to the irrelevant prefix. This gives the intrusion score

$$s_{\text{int}}(h) = \frac{1}{N}\sum_{i=1}^{N} a_h^{(i,\text{irr})} \max\Big(0,\, \text{DLA}_h^{(i,\text{nctx})}(g_i) - \text{DLA}_h^{(i,\text{irr})}(g_i)\Big). \tag{3}$$

Here a_h sums attention from END to prefix positions, excluding BOS and the query, and the N discovery pairs give equal weight to short and long prefixes. We call the high-scoring population **intrusion heads** (I): their defining signature couples heads whose attention to the irrelevant prefix coincides with a reduction in their direct contribution to the correct-answer logit relative to the context-free run, rather than copying alone.

Having identified heads associated with reduced answer support, we inspect how their reads and writes change when irrelevant context is added. We select the intervention set $I^\star$ using the intrusion score on discovery queries and keep it fixed for evaluation on held-out queries, testing whether its behavior extends beyond the examples used for selection (Appendix C.1.1). Analyses of this set track three quantities: attention to the prefix and query, writes toward the correct answer, and distractor tokens, together with the associated words for each. In the context-free run, these heads attend to the query and BOS. Adding the irrelevant context shifts part of this attention toward the prefix. This change is accompanied by increased writing toward distractor tokens and their associated words, while their contributions to the correct answer weaken (Appendix C.1). Because distractor-directed writing does not enter the selection score, its increase constitutes an additional empirical finding. Figure 4 illustrates the source-to-destination net logit contributions for the running example.

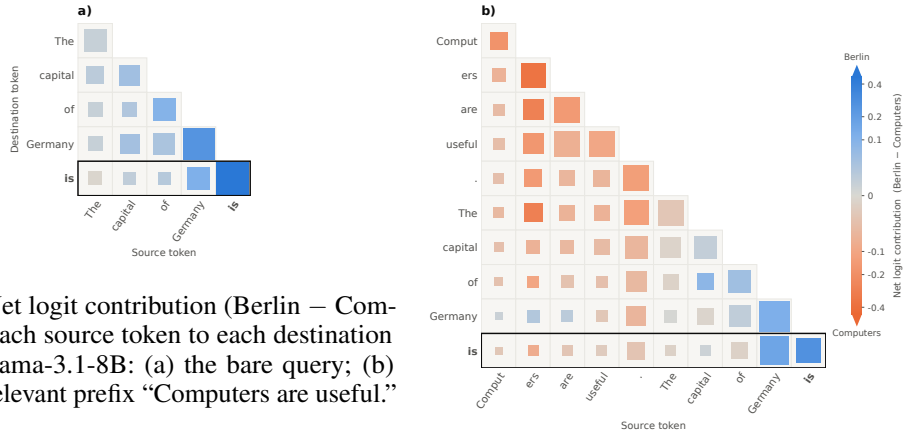


Figure 4: Net logit contribution (Berlin − Computers) of each source token to each destination token in Llama-3.1-8B: (a) the bare query; (b) with the irrelevant prefix "Computers are useful."

Prefix and non-prefix contributions. To locate these changes, we split each write as $o_h = P_h + U_h$, with P_h supplied by prefix positions and U_h by the remaining sources. Most of the increased distractor writing comes from P_h, but this contribution has nearly zero mean direct gold logit attribution. The gold decrease instead comes from smaller contributions within U_h, including subject, relation, and END positions (Appendix C.2.1). This distinction is also evident in the matched-filler comparison: distractor promotion persists, but much of the gold-attribution difference shrinks (Figure C.1). Thus, "Computers" is not simply read as a vector that opposes "Berlin": the prefix adds distractor-related content, but the loss of answer support comes from weaker query-derived contributions. This connects back to subject enrichment—information can remain available at the subject without contributing equally strongly to the answer.

Restoring answer-supporting writes. The non-prefix contribution still supplies useful support for "Berlin," so deleting the whole head would remove part of the answer computation along with the distractor-related write. We instead test whether correcting the context-dependent change restores that support. On held-out queries, replacing the END writes of the unfiltered population I with

their context-free counterparts improves the gold logit more than matched-head controls, while the irrelevant prefix remains in the input. Interpolating as $o_h^{(\lambda)} = o_h^{\text{irr}} + \lambda(o_h^{\text{nctx}} - o_h^{\text{irr}})$ produces progressively stronger recovery as the correction moves from no replacement toward the full clean write (Appendix C.2.2). These interventions establish that the selected outputs are consequential, but still require a matched context-free pass. The source split suggests a finer alternative that does not: restrict the heads' prefix access while retaining the remaining sources.

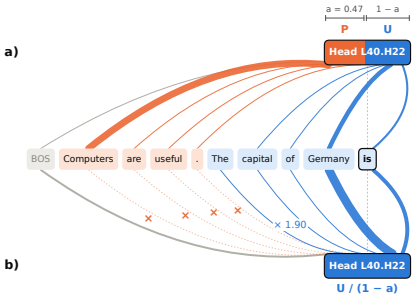


Figure 5: At fixed head inputs, masking combines prefix removal with amplification of the remaining contribution. It leaves the head in place and requires no clean donor.

Targeted attention masking. We implement this alternative by masking END-to-prefix attention in the fixed intervention set. Softmax redistributes the blocked attention across non-prefix positions, so the mask both removes the prefix write and strengthens the remaining reads (Figure 5). At fixed head inputs,

$$o_h^{\text{mask}} - o_h = -P_h + \frac{a_h}{1 - a_h} U_h, \qquad a_h < 1. \tag{4}$$

We separate these operations by removing P_h without redistributing its attention, amplifying U_h while retaining P_h, and applying both changes together. Prefix removal supplies most of the distractor reduction, whereas amplification increases the edited write's direct gold support. Where mask improves the final prediction, its benefit can arise from both removal and amplification; applying the same gain to the complete heads' write does not reproduce that benefit (Appendix C.2.3). The full mask can therefore improve gold probability and lower distractor support on the same held-out queries without deleting the heads or consulting a clean run (Appendix C.2.4).

Direct versus copier-mediated access. Does this contextual read depend on information first copied into END? We replace the measured copiers' contribution to the intrusion heads' END query inputs with its context-free counterpart. This barely changes their attention and produces no substantial beneficial recovery of prediction loss. A complementary intervention targets the prefix itself: its keys and values are replaced with those from corresponding token positions in a filler run, leaving END unpatched. In settings where masking improves prediction, this intervention recovers part of that improvement. The consequential access can therefore be direct: the tested copier-to-query route does not account for the effect of changing what these heads read from the prefix (Appendix C.3).

Comparison with entrainment and retrieval heads. The distractor-directed signal also invites comparison with contextual entrainment, in which heads just promote the tokens already present in the prefix (Niu et al., 2025). This behavioral overlap does not establish that the same head population is involved. We therefore compare the independently ranked intrusion heads with the entrainment set. It contains none of the twenty highest intrusion-scored heads (Appendix C.4). The overlap in behavior thus does not amount to rediscovering that entrainment population.

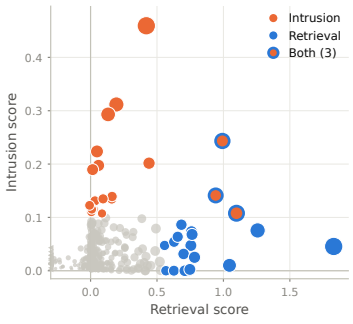


Figure 6: Retrieval score versus intrusion score for each head in GPT-2 XL.

Because the intrusion score identifies heads through reduced gold support under irrelevant context, we must also check whether masking them preserves answer support when the prefix is helpful. In our running example, reducing interference from "Computers are useful" should not prevent the model from using "Germany's capital city is Berlin." We therefore compare intrusion heads with contextual retrieval heads (Wu et al., 2025b), scored by their attention to the answer's position in the golden prefix and their gold-directed write. Comparing the independently ranked populations before any exclusions shows that their top-scoring heads are different, although some overlap remains (Figure 6; Appendix C.5). With golden context, this targeted mask retains the benefit supplied by the evidence (Appendices C.2.4 and C.5). Thus, restricting contextual access in the selected intrusion heads can reduce interference while largely preserving the model's ability to use helpful context, rather than achieving improvement by blocking context everywhere.

5 DOWNSTREAM EVALUATION

The pathway identified in §3 suggests a practical intervention with two requirements. The targeted intrusion heads must be separable from the heads that integrate evidential context, as §4 establishes. They must also remain useful intervention targets beyond the task on which they were identified, so that a model-specific set can be identified once on a simple benchmark and reused. We test this second requirement through independent head re-identification and fixed-set intervention on four downstream tasks.

Evaluation setting. We evaluate four base models: Llama-3.1-8B, Mistral-7B-v0.1, Qwen2.5-7B, and Pythia-6.9B and use the same four benchmarks to assess cross-task head stability and downstream intervention effectiveness. Each pairs an original task input with controlled irrelevant context; the clean condition retains the original question and any supporting evidence. GSM-DC (Yang et al., 2025) injects k type-consistent but irrelevant arithmetic nodes into GSM8K dependency graphs. GSM-IC (Shi et al., 2023) inserts one irrelevant sentence into the original problem. HotpotQA (Yang et al., 2018) retains the two supporting passages and adds k non-supporting Wikipedia paragraphs. PubMedQA (Jin et al., 2019) retains the supporting abstract and appends k irrelevant abstracts. These settings include in-domain distractors, extending the evaluation beyond the unrelated prefixes used for head identification. We report recovery as the percentage of distractor-induced accuracy loss restored by an intervention; task-specific metrics, sample sizes, and per-k results are provided in Appendix D.

Cross-task stability. If intrusion reflects a shared context-processing operation rather than a task-specific subroutine, different tasks should recruit overlapping intrusion-head populations, consistent with findings that head-level roles are reused across tasks (Wu et al., 2025b; Merullo et al., 2024). We test this prediction within each of the four base models by independently recomputing the intrusion score in Eq. 3 on factual recall, binding/retrieval (IOI), and the four downstream tasks, and comparing the resulting rankings and selected head sets. As a within-task reference, independent rankings from $1{,}000$ random halves of the factual-recall data yield Spearman correlations of 0.97–0.99 across models. Table D.1 summarizes cross-task overlap among the ten-head sets.

Fixed-head attention masking. We first test the practical intervention suggested by §4. For the frozen, retrieval-disjoint intrusion-head set, we mask the pre-softmax attention scores from END to context positions, redistributing attention to non-context positions while leaving all other heads unmodified. This suppresses the selected attention edges rather than deleting the heads or their weights. The intervention recovers 48.5–63.3% of the distractor-induced accuracy loss across the four tasks (Table 2), without a matched clean forward pass. The GSM-DC sweep further shows that its benefit is not restricted to a single distractor count (Figure 7). A broader knockout that includes dual-use heads recovers less (Appendix D), consistent with the importance of preserving useful context readers rather than indiscriminately suppressing context attention.

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 drop (7B baselines): +Patch layer residual patch, +K10 clean intrusion-head knockout ($\sim$10 heads), +K20 top-20 incl. dual-use. (>=100% = above baseline)

Oracle residual patching. To compare this sparse intervention with a broader oracle correction, we replace the entire END-position residual at the model-specific intrusion-band layer with its counterpart from the matched clean input. The substitution is applied per generation step, while other token positions are not directly patched. This oracle recovers 97.6% of the distractor-induced loss on average, ranging from 85.5% on HotpotQA to 108.8% on GSM-IC (Table 2). Recovery above 100% means that the patched model exceeds the unmodified clean baseline. One plausible explanation is that replacing the accumulated contamination at END, while retaining potentially useful contextual representations at other positions, allows subsequent layers to combine a restored answer state with those representations under reduced interference. Because the remaining contextual states still come from the distracted run, this intervention need not merely reproduce the clean computation. The patch therefore provides an oracle reference for how much loss can be reversed at the localized residual state, whereas knockout measures what can be recovered through a fixed, sparse intervention.

Together, these results turn the mechanistic localization into a reusable intervention: a fixed mask affecting approximately 1% of attention heads recovers 56.9% of distractor-induced accuracy loss on average across four tasks. It requires neither downstream head reselection nor a matched clean forward pass at inference. Oracle patching, which does require that clean pass, marks the ceiling: nearly all of the loss is reversible at the localized state, and knockout is the share of that ceiling reachable at deployment.

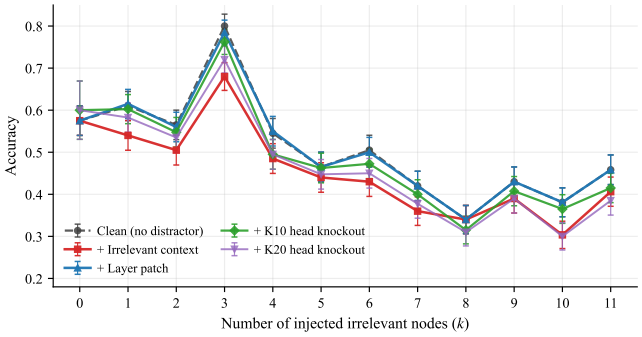


Figure 7: GSM-DC across k: layer patching closes the gap at every k; K10 recovers a large fraction with $\sim$10 heads. Per-model breakdowns in Appendix D.

6 Conclusion

We provide a mechanistic account of how non-evidential context weakens answer support, distinguish this loss from contextual token promotion alone, and use that understanding to guide a targeted intervention. We find answer support can weaken despite preserved subject enrichment and selective reading. Broad write restoration recovers support but requires a context-free donor and also removes golden-context benefits. We then localize a recoverable component of the interference to a sparse set of intrusion heads. These heads are largely distinct from the tested entrainment and golden-context retrieval populations. This localization guides a selective intervention that restricts contextual reads in the identified intrusion heads.

The resulting training-free mask targets roughly 1% of all attention heads. Applied unchanged across four downstream tasks, it recovers 57% of distractor-induced accuracy loss on average while largely preserving golden-context benefits. Broader oracle residual patching recovers 97.6% on average but required a matched clean pass. Retrieval-augmented generation, tool use, agentic loops, and long-context settings motivates this problem: helpful evidence and non-evidential material can arrive through the same context channel. A substantial part of contextual interference can be localized and mitigated within the computation itself, without retraining the model or broadly disabling its access to useful evidence.

Limitations

Our findings concern the evaluated open-weight decoder-only base models, with four 7B-scale models forming the main evaluation and GPT-2 XL included in circuit diagnostics. Broader applicability to substantially larger models, mixture-of-experts architectures, encoder–decoder models, and multimodal systems requires separate validation. The circuit analyses use factual recall and IOI binding/retrieval, while downstream evaluation covers four English benchmarks with controlled distractors, including in-domain material. These settings do not establish robustness for code, multilingual prompts, multi-turn dialogue, mathematics beyond the evaluated GSM tasks, or adversarial and jailbreak-style inputs.

Our measurements concern selected components, not complete representations or their training origins; causal evidence comes from interventions rather than logit attribution alone. The fixed mask requires internal attention access, model-specific head selection from paired inputs, and specified context positions, but does not determine relevance. Recovery remains partial, and helpful-context preservation is empirical rather than guaranteed. Oracle patching requires matched clean activations and provides a diagnostic reference, not a deployable solution or formal upper bound. Comparisons with softer and learned interventions remain future work. Our experiments do not establish a minimal head set or an optimal trade-off between robustness and helpful-context preservation.

AI USE STATEMENT

In this work, we used generative AI tools, including Claude (Anthropic), to aid and polish the writing and to assist with formatting the manuscript in LaTeX, including the layout of tables and figures and the formatting of references. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code is provided in the supplementary material. Section 2 specifies the models, task data, and metrics, and Appendix A describes prompt construction, prefix conditions, and filtering. The interventions and head populations are defined in Sections 3 and 4 and Appendices B and C. The downstream protocol is described in Section 5 and Appendix D, and the head-count and selection-penalty settings are given in Appendix F.

REFERENCES

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Borun D. Chowdhury. Lost in the middle at birth: An exact theory of transformer position bias, 2026. URL https://arxiv.org/abs/2603.10123.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 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)*, pp. 719–729, 2024.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 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*, pp. 10028–10039, 2024.

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

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

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

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

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

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 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)*, pp. 2567–2577, 2019.

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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 9459–9474, 2020.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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*, pp. 9802–9822, 2023.

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

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. 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*, pp. 16218–16239, 2025. doi: 10.18653/v1/2025.acl-long.791. URL `https://aclanthology.org/2025.acl-long.791/`.

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

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

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024. URL `https://arxiv.org/abs/2412.15115`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

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

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

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

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 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*, pp. 5019–5039, 2025a. doi: 10.18653/v1/2025.acl-long.250. URL https://aclanthology.org/2025.acl-long.250/.

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

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

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. How is LLM reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13329–13347. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.674. URL https://aclanthology.org/2025.emnlp-main.674/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2369–2380, 2018.

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

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

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

A Setup

A.1 Factual-recall benchmark: relation composition

The factual-recall benchmark draws on 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 & Steinhardt, 2024) have been characterized — is represented by `city_in_country` and `landmark_in_country`.

Relation distribution. 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. (2023), 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** (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 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. (2023). 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.

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 Length axis

We sweep prefix lengths $\ell \in \{1, 12, 50, 100, 300\}$ tokens. The 50-token prefixes truncate long-irrelevant paragraphs at the nearest sentence boundary; 300-token prefixes concatenate two paragraphs under the same leakage filter. Gold-logit shifts and intrusion-head attention to context saturate by $\ell \approx 50$; additional content increases diffuse parroting (§3) without further gold suppression.

A.5 Distractor count axis

To test content dependence against the position-bias alternative in concurrent work (Chowdhury, 2026), we repeat the core circuit analyses with distractors appended after the query but before generation, or interleaved between subject and relation. Both preserve the Intrusion-Pathway signature: late-block intrusion-head attention follows the irrelevant content, and gold-logit suppression remains within the prepended default's bootstrap CI. Damage therefore depends on content rather than absolute position.

A.6 Distractor position axis

We repeat the core circuit analyses with distractors appended after the query, before generation, or interleaved between subject and relation. Late-block intrusion-head attention follows the distractor, and gold-logit suppression remains within the prepended baseline's bootstrap CI.

A.7 Semantic-distance robustness

We test far (physics/mechanics; default `short_irr`), medium (geography about another country), and near (same country, different relation) irrelevant prefixes. Intrusion-head attention to context and gold-logit suppression persist across all three distances.

A.8 Leakage filter

For `long_irr`, we reject WikiText-103 paragraphs containing whitespace-tokenized substring matches to the sample's gold or subject token, other valid golds or subjects from the same relation, or gold-token semantic neighbors with GloVe-300d cosine similarity > 0.5 (Pennington et al., 2014).

A.9 Distractor pool composition and selection robustness

The `short_irr` pool contains sentences from physics/mechanics, biology, agriculture, everyday objects, weather, and abstract concepts. Sentences have 8–18 whitespace tokens, with no gold-attribute or subject-attribute lexical overlap across the tested LRE relations.

Robustness checks replace per-sample most-damaging selection (Shi et al., 2023) with random selection, median-damage selection, or one fixed sentence for all samples. All three preserve the qualitative per-component signature; absolute damage scales monotonically with the expected selection effect.

B Intrusion Pathway

Conditions retain the main-text notation: `nctx` (bare query), `far_irr` (type-disjoint LRE prefix; neither the query's relation type nor its gold attribute is supplied), and `long_irr` (irrelevant WikiText prefix).

B.1 MLP ENRICHMENT

B.1.1 SUBJECT ATTRIBUTES AND PRESERVATION

Subject writes favor their own attribute banks: peak LRE excess is 0.073–0.395 and Wikipedia excess is 0.040–0.329 (Table B.1a). Figure B.1 compares top-20 Attributes Rates.

Table B.1: **Entity-specific MLP enrichment.** Own-minus-other attribute-bank projection at the last subject or distractor token, read through the unembedding at a frozen final-normalization scale (Geva et al., 2023). Peak selection excludes the final layer; brackets are 95% paired-bootstrap intervals conditional on the selected layer. Final-layer values are point estimates.

Model	Bank	Layer	Peak excess [95% CI]	Final layer
(a) Subject position				
Llama-3.1-8B	LRE	30	+0.151 [+0.039, +0.267]	+1.217
	Wikipedia	30	+0.231 [+0.178, +0.285]	+1.370
Mistral-7B-v0.1	LRE	23	+0.229 [+0.135, +0.318]	+0.228
	Wikipedia	29	+0.186 [+0.113, +0.283]	+0.488
Qwen2.5-7B	LRE	24	+0.395 [+0.289, +0.499]	+0.844
	Wikipedia	26	+0.329 [+0.269, +0.394]	+1.014
Pythia-6.9B	LRE	21	+0.160 [+0.098, +0.226]	+0.279
	Wikipedia	30	+0.198 [+0.145, +0.255]	+0.467
GPT-2 XL	LRE	45	+0.073 [+0.021, +0.129]	−0.013
	Wikipedia	0	+0.040 [+0.028, +0.052]	−0.044
(b) Distractor position				
Llama-3.1-8B	Category	30	+0.110 [+0.075, +0.144]	+1.049
	Wikipedia	30	+0.272 [+0.223, +0.321]	+1.482
Mistral-7B-v0.1	Category	28	+0.126 [+0.107, +0.145]	+0.152
	Wikipedia	30	+0.165 [+0.137, +0.193]	+0.555
Qwen2.5-7B	Category	26	+0.386 [+0.330, +0.442]	+0.922
	Wikipedia	26	+0.292 [+0.239, +0.345]	+0.309
Pythia-6.9B	Category	30	+0.173 [+0.106, +0.236]	+0.353
GPT-2 XL	Category	38	+0.112 [+0.088, +0.139]	−0.027
	Wikipedia	0	+0.081 [+0.071, +0.093]	−0.022

Banks are defined independently of the tested MLP: LRE attributes or Wikipedia content words for subjects; category co-members and category words, or Wikipedia content words, for distractors. Source/context tokens, the gold answer, generic query-category words, and other valid answers are excluded.

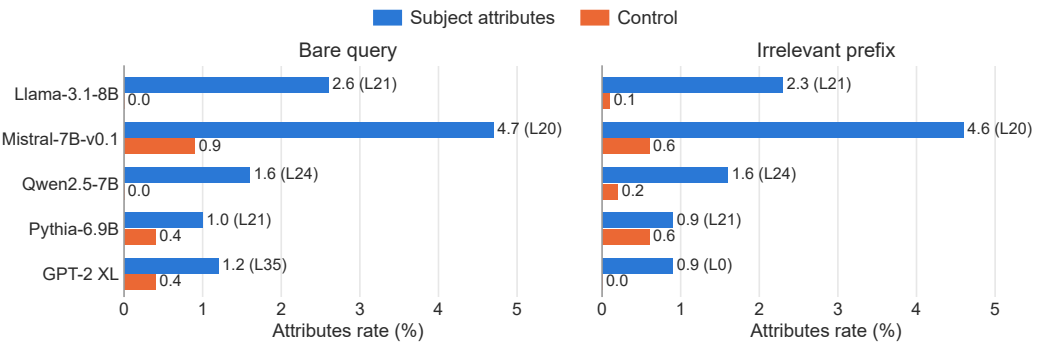


Figure B.1: **Subject Attributes Rate.** Percentage of top-20 projected tokens in the Wikipedia bank at the peak layer, with matched controls. Parentheses identify the layer. Values are point estimates.

B.1.2 ASSOCIATIONS AT IRRELEVANT-TOKEN POSITIONS

Distractor-position writes also favor their own associations: category excess is 0.110–0.386, with positive Wikipedia comparisons (Table B.1b).

B.1.3 CONTRIBUTIONS OF COMPLETE MLP WRITES

Cancelling early subject writes lowers the gold logit by 3.10–5.46 under bare queries and 3.05–4.60 under `far_irr`. Figure B.2 includes both subject and distractor edits with function-word controls. These interventions act on complete writes, including content outside the attribute banks.

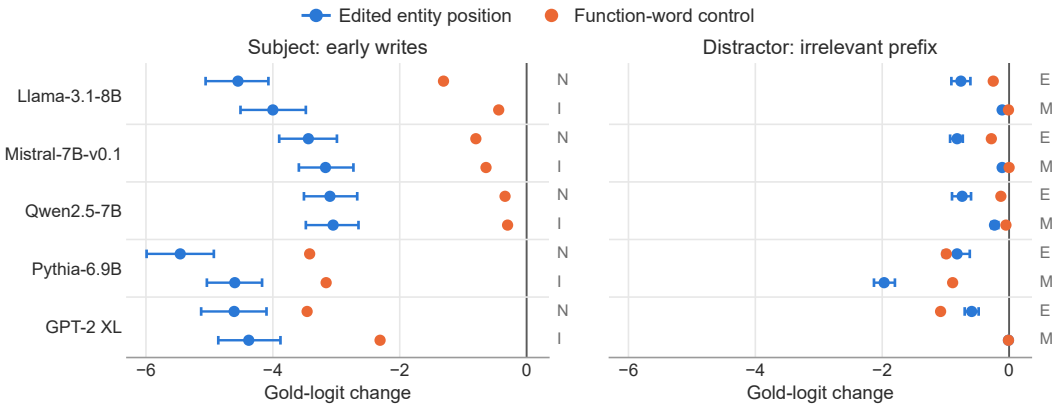


Figure B.2: **MLP-write cancellation.** Gold-logit change at END relative to the unedited run of the same prompt. Left: early-band subject edits under bare queries (N) and `far_irr` (I). Right: distractor edits in the early (E) or middle (M) band under `far_irr`. Bars show 95% paired-bootstrap intervals for entity-position edits; controls are point estimates at query function words (left) or prefix function words (right).

B.2 SOURCE-TOKEN COPYING

B.2.1 SOURCE SUBSTITUTION WITH FIXED ATTENTION

Source substitution shifts the source identity contrast by $+0.370$–$+4.690$ across five models and produces a positive END shift when attention is fixed (Figure B.3). Matched prompts have identical token counts and differ in one token ID, replacing B with D without a matching induction cue.

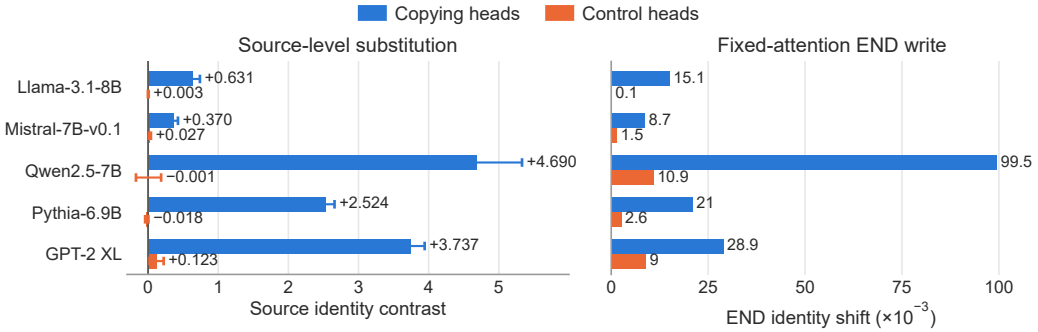


Figure B.3: **Identity substitution.** Left: change in the source OV write's replacement-versus-original identity projection, with 95% confidence intervals. Right: source-value substitution at fixed attention, measured in the END write (10^{-3} units; point estimates). Copying analyses use ten heads per model ranked by own-source OV identity projection minus mean projection onto other sources under `far_irr`. The intervention comparison in §B.2.3 uses eight heads.

B.2.2 COPYING AFTER CUE REMOVAL

Replacing $A\,B \cdots A$ with $C\,B \cdots A$ retains 7–26% of copying heads' attribution to B; induction controls range from -1 to 4% (Figure B.4). These ratios describe cue sensitivity within the tested heads; they do not compare absolute no-cue copying magnitudes across populations.

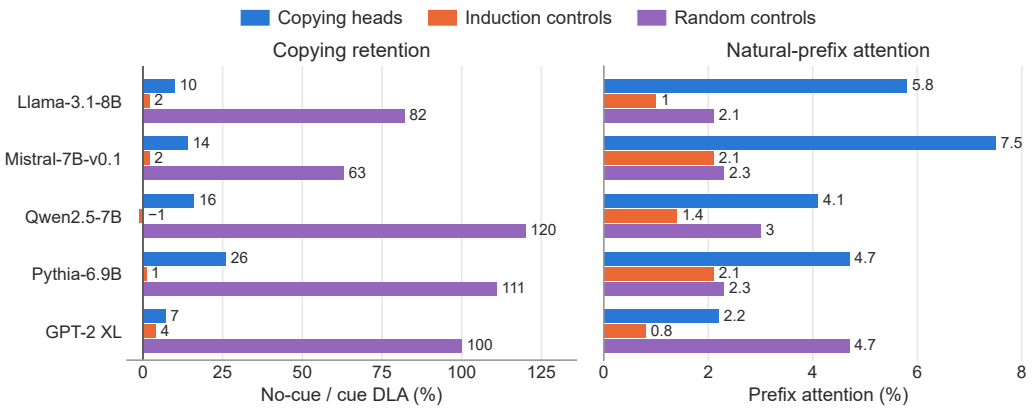


Figure B.4: **Cue removal and natural-prefix attention.** Left: median no-cue/cue direct-logit-attribution ratio, keeping B, its position, and the final cue fixed. Right: prefix attention in natural prompts without an induction cue. Induction controls are independently selected by prefix matching on repeated random sequences. Random controls are included in both panels. All values are point estimates.

B.2.3 Copying-head interventions across context types

The comparison tests eight copying heads under useful and irrelevant prefixes separately and together (Table B.2). Under mixed prefixes, edge masking changes the gold logit by -0.026 to -0.145 across the five models.

Table B.2: **Copying-head interventions.** Gold-logit change relative to the unedited run. Mixed context concatenates the golden and irrelevant prefixes. Whole-head edits zero or donor-replace the head output; END-only zeroing is restricted to END. Prefix-write removal deletes only prefix-source contributions, while prefix-edge masking blocks attention to those sources. Values are point estimates.

Model	Edit	Bare	Golden	Irrelevant	Mixed
Llama-3.1-8B	Whole-head zeroing	-0.025	-0.118	-0.034	-0.085
	Whole-head donor replacement	-0.003	-0.095	-0.019	-0.062
	END-only zeroing	-0.025	-0.095	-0.035	-0.070
	Prefix-write removal	0.000	-0.080	-0.009	-0.030
	Prefix-edge masking	0.000	-0.068	0.000	-0.035
Mistral-7B-v0.1	Whole-head zeroing	-0.016	-0.102	-0.005	-0.084
	Whole-head donor replacement	-0.013	-0.100	$+0.006$	-0.079
	END-only zeroing	-0.019	-0.107	-0.010	-0.081
	Prefix-write removal	0.000	-0.088	$+0.008$	-0.060
	Prefix-edge masking	0.000	-0.075	$+0.010$	-0.056
Qwen2.5-7B	Whole-head zeroing	-0.076	-0.201	-0.085	-0.251
	Whole-head donor replacement	-0.036	-0.161	-0.045	-0.168
	END-only zeroing	-0.049	-0.180	-0.052	-0.200
	Prefix-write removal	0.000	-0.139	-0.021	-0.155
	Prefix-edge masking	0.000	-0.133	-0.011	-0.145
Pythia-6.9B	Whole-head zeroing	$+0.027$	-0.279	-0.114	-0.232
	Prefix-write removal	0.000	-0.015	$+0.081$	-0.066
	Prefix-edge masking	0.000	$+0.003$	$+0.105$	-0.072
GPT-2 XL	Whole-head zeroing	$+0.019$	$+0.034$	$+0.016$	$+0.040$
	Prefix-edge masking	0.000	-0.026	-0.002	-0.026

B.2.4 Attention-weighted reference for copying

For a head h, let $\alpha_p = A_h[\mathrm{END}, p]$ and let $w_{h,p} = W_O^{(h)} W_V^{(h)} x_p^{(L)}$ denote its unweighted source OV write, omitting normalization and bias terms. Its complete END write and prefix-content contribution

are

$$o_h = \sum_p \alpha_p w_{h,p}, \qquad o_h^{\text{content}} = \sum_{p \in \mathcal{P}_{\text{content}}} \alpha_p w_{h,p}. \tag{B.1}$$

The content set excludes BOS, stopwords, and query repetitions; attention is not renormalized over it. With source-token unembedding u_{t_p}, a shared positive copy-gain reference is

$$\begin{aligned} w_{h,p} &= c_h u_{t_p} + \epsilon_{h,p}, \qquad c_h > 0, \\ o_h^{\text{content}} &= c_h u_*^{\text{attn}} + r_h, \qquad u_*^{\text{attn}} = \sum_{p \in \mathcal{P}_{\text{content}}} \alpha_p u_{t_p}, \quad r_h = \sum_{p \in \mathcal{P}_{\text{content}}} \alpha_p \epsilon_{h,p}. \end{aligned} \tag{B.2}$$

The residual includes source-dependent gains and other represented content; it need not be small or orthogonal. A dominant source gives a single-source reference; several contributing sources give a diffuse mixture. For any target direction u_t,

$$\langle o_h^{\text{content}}, u_t \rangle = c_h \sum_{p \in \mathcal{P}_{\text{content}}} \alpha_p \langle u_{t_p}, u_t \rangle + \langle r_h, u_t \rangle. \tag{B.3}$$

Thus contributions depend on token geometry as well as attention. The held-out comparison permutes identity–attention pairings while preserving both multisets:

$$u_*^{\text{shuffle}} = \sum_{p \in \mathcal{P}_{\text{content}}} \alpha_p u_{t_{\pi(p)}}. \tag{B.4}$$

Equal weights can make the shuffle uninformative. This tests an identity component of the write. CSPA (McDougall et al., 2023) instead preserves negative source-wise projections with destination-prediction filtering, evaluated against mean ablation.

B.3 Multi-source copying and association output

B.3.1 Attention-weighted source mixtures

Correct identity–attention pairing improves held-out reconstruction for 2–7 of ten selected copying heads per model, compared with 0–1 control heads (Table B.3). The reported cosine ranges describe alignment with this component.

Table B.3: **Multi-source reconstruction.** Number of heads whose paired-bootstrap interval for shuffled-minus-correct reconstruction error is above zero. Cosines give alignment between copying writes and the identity mixture.

Model	Copying heads	Control heads	Cosine range
Llama-3.1-8B	2/10	1/10	0.03–0.23
Mistral-7B-v0.1	3/10	1/10	0.06–0.18
Qwen2.5-7B	3/10	0/10	0.06–0.25
Pythia-6.9B	6/10	0/10	0.03–0.15
GPT-2 XL	7/10	1/10	0.07–0.20

B.3.2 Association-related output

Copying writes show positive association excess in all five models; held-out profile errors and cosines measure source-profile preservation separately (Table B.4).

B.3.3 Selective edits to association content

Table B.5 removes a rank-12 association component with the identity direction projected out and attention fixed. Identity-removal and matched random edits are controls; identity and final-answer readouts accompany the association changes.

B.4 Relational readers: frame and answer type

Unrelated-filler attention is 0.008–0.012, compared with 0.194–0.273 for framed gold answers (Figure B.5). The controlled alternatives separate answer type from relational frame.

Table B.4: **Association output through OV.** Copying-minus-identity-reference association projection (excluding the source token), with 95% intervals and layer/head-matched random controls. Profile error and cosine compare source/output association coordinates; one scalar gain is fitted on the discovery half and fixed on held-out data. Profiles are held-out medians; control excesses are point estimates.

	Association excess		Profile alignment	
Model	Copying heads [95% CI]	Control	Rel. error	Cosine
Llama-3.1-8B	$+0.017$ $[+0.009, +0.025]$	$+0.001$	1.000	$+0.005$
Mistral-7B-v0.1	$+0.008$ $[+0.005, +0.011]$	$+0.001$	1.025	$+0.024$
Qwen2.5-7B	$+0.086$ $[+0.048, +0.127]$	-0.005	0.989	$+0.067$
Pythia-6.9B	$+0.095$ $[+0.079, +0.112]$	$+0.003$	0.996	$+0.087$
GPT-2 XL	$+0.161$ $[+0.133, +0.190]$	-0.028	1.000	-0.011

Table B.5: **Association-subspace removal.** (a) Association-output change; brackets are 95% intervals, rounded to four decimals. Random subspaces match rank and norm. (b) Identity projection and final gold logit before/after association removal. Controls and before/after values are point estimates.

(a) Association-output change

Model	Association removal [95% CI]	Identity removal	Random subspace
Llama-3.1-8B	-0.0005 $[-0.0009, -0.0000]$	$+0.0010$	-0.0001
Mistral-7B-v0.1	$+0.0001$ $[+0.0001, +0.0001]$	$+0.0000$	-0.0000
Qwen2.5-7B	-0.0035 $[-0.0052, -0.0019]$	$+0.0006$	$+0.0002$
Pythia-6.9B	-0.0055 $[-0.0078, -0.0032]$	$+0.0000$	-0.0004
GPT-2 XL	-0.0163 $[-0.0214, -0.0116]$	-0.0041	$+0.0001$

(b) Identity and answer readouts

	Identity projection		Gold logit	
Model	Before	After	Before	After
Llama-3.1-8B	0.387	0.384	18.921	18.909
Mistral-7B-v0.1	0.206	0.206	15.188	15.184
Qwen2.5-7B	3.041	3.033	18.529	18.519
Pythia-6.9B	1.573	1.552	16.209	16.233
GPT-2 XL	2.113	2.085	12.816	12.807

B.5 Restoring attention writes and residual states

B.5.1 Middle-block attention and MLP outputs

We replace END outputs in the middle depth third (approximately 33–67%) with matched bare-query outputs, retaining the prefixed input. For gold logit or gold-versus-best-competitor margin s, the signed restoration ratio is

$$R_s = 100\,\frac{s_{\text{restored}} - s_{\text{ctx}}}{s_{\text{nctx}} - s_{\text{ctx}}}. \tag{B.5}$$

When context lowers s, 100% denotes full recovery. When context raises s, a positive ratio denotes removal of that benefit. Table B.6 retains the baseline changes so both cases are visible; Qwen's `far_irr` gold and margin baselines are positive. Golden-context results use the same definition (Table B.7).

Figure B.6 compares attention, MLP, and joint restoration in the same middle window. The modelwise panels retain each signed effect and its reported uncertainty.

B.5.2 Restoration across depth

Table B.8 compares complete early, middle, and late depth thirds. Figure B.7 replaces the END residual at individual layers, with subsequent layers continuing on the prefixed run. The two experiments vary the restored site and intervention scope separately.

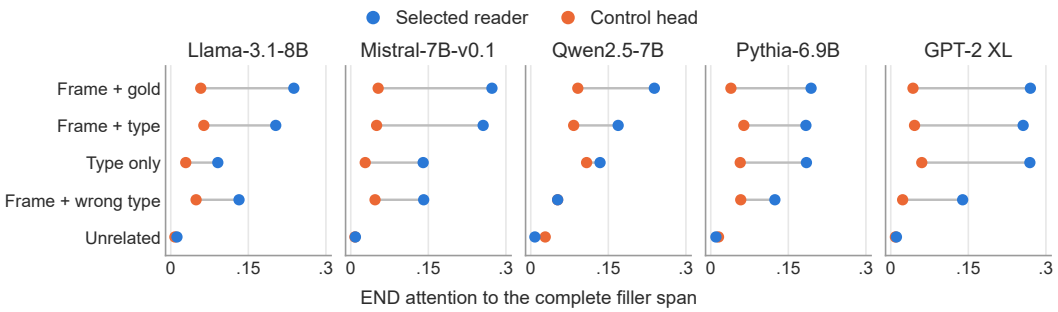


Figure B.5: **Reader attention by frame and type.** END attention summed over the complete filler span at the end of the prefix. "Type" denotes an incorrect answer of the right type. Readers are the ten middle-layer heads selected on the discovery half by golden-answer-span attention minus maximum `far_irr` context attention. All values, including control heads, are point estimates.

Table B.6: **Middle-block attention restoration.** Baseline changes are context minus bare-query values. Signed ratios use Eq. (B.5); brackets are 95% confidence intervals. For Qwen `far_irr`, the positive ratio removes a baseline gain.

		Baseline change		Signed ratio (%) [95% CI]	
Model	Condition	Gold logit	Margin	Gold logit	Margin
Llama-3.1-8B	`far_irr`	−1.450	−0.275	61 [57,65]	87 [67,120]
	`long_irr`	−1.882	−1.196	54 [50,58]	70 [64,78]
Mistral-7B-v0.1	`far_irr`	−0.708	−0.107	57 [52,62]	168 [−388,988]
	`long_irr`	−1.565	−0.731	60 [56,64]	79 [69,92]
Qwen2.5-7B	`far_irr`	+0.515	+0.330	72 [56,96]	57 [36,110]
	`long_irr`	−0.477	−0.788	−6 [−56,17]	17 [1,32]
Pythia-6.9B	`far_irr`	−0.525	−0.528	90 [77,109]	103 [87,130]
	`long_irr`	−1.227	−0.897	71 [65,78]	86 [74,100]
GPT-2 XL	`far_irr`	−0.376	−0.324	41 [21,53]	72 [54,93]
	`long_irr`	−1.724	−1.263	53 [48,58]	66 [58,75]

Table B.7: **The same restoration under golden context.** Benefit is the unedited gold-logit increase over the bare query; the removed fraction includes 95% confidence intervals.

Model	Gold-logit benefit	Benefit removed (%) [95% CI]
Llama-3.1-8B	+0.767	68 [61,79]
Mistral-7B-v0.1	+0.502	58 [48,69]
Qwen2.5-7B	+0.955	33 [23,43]
Pythia-6.9B	+0.690	45 [30,57]
GPT-2 XL	+0.683	87 [70,109]

C Intrusion Heads

C.1 Populations and Contextual Writes

C.1.1 Populations and Attention

We freeze three distinct head populations: I contains the twenty highest intrusion-scored heads; $I^{\star}$ contains twenty heads selected after excluding the top forty retrieval-scored heads; $I_{\text{only}} = I_{20} \setminus R_{20}$ is used only in §C.5. These sets are distinct from downstream $K{=}10$ interventions. Short and long denote `far_irr` and `long_irr`.

Table B.8: **Window restoration.** Signed gold-logit restoration ratio (%) after replacing both attention and MLP outputs at END within the early, middle, or late depth third. Brackets give 95% confidence intervals. Windows are restored in separate runs; their ratios do not partition the total effect. Qwen `far_irr` ratios measure removal of a context benefit.

Model	Condition	Early	Middle	Late
Llama-3.1-8B	`far_irr`	6 [2,9]	67 [64,70]	94 [91,97]
	`long_irr`	1 [−2,4]	60 [57,62]	88 [85,90]
Mistral-7B-v0.1	`far_irr`	2 [−5,8]	70 [66,75]	98 [93,103]
	`long_irr`	4 [−1,8]	74 [71,77]	90 [88,93]
Qwen2.5-7B	`far_irr`	17 [8,27]	62 [49,83]	95 [90,99]
	`long_irr`	22 [7,47]	25 [−4,49]	107 [102,116]
Pythia-6.9B	`far_irr`	12 [−2,24]	72 [62,84]	115 [96,141]
	`long_irr`	1 [−9,8]	56 [49,63]	87 [78,98]
GPT-2 XL	`far_irr`	−3 [−16,7]	56 [41,68]	81 [61,107]
	`long_irr`	0 [−7,5]	59 [54,64]	69 [62,76]

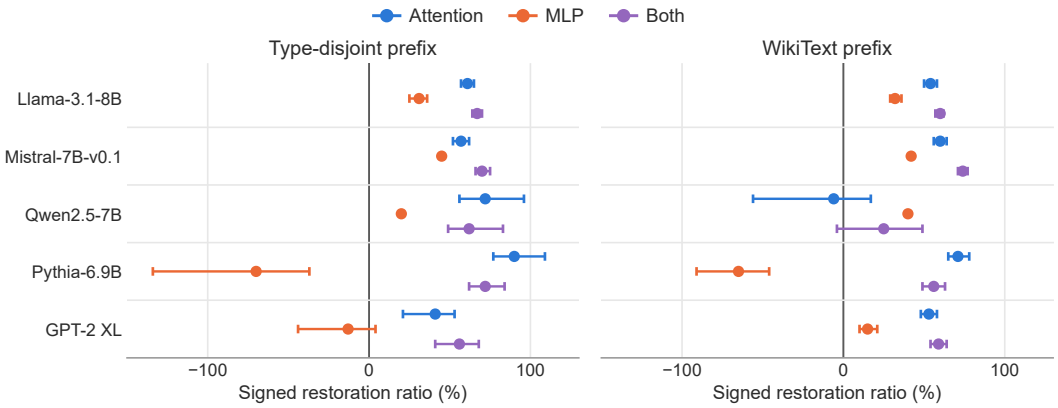


Figure B.6: **Sublayer restoration in the middle third.** Signed gold-logit restoration ratio. Bars are 95% intervals. Each point is a separate intervention, so ratios are not additive. Qwen `far_irr` ratios measure removal of a context benefit.

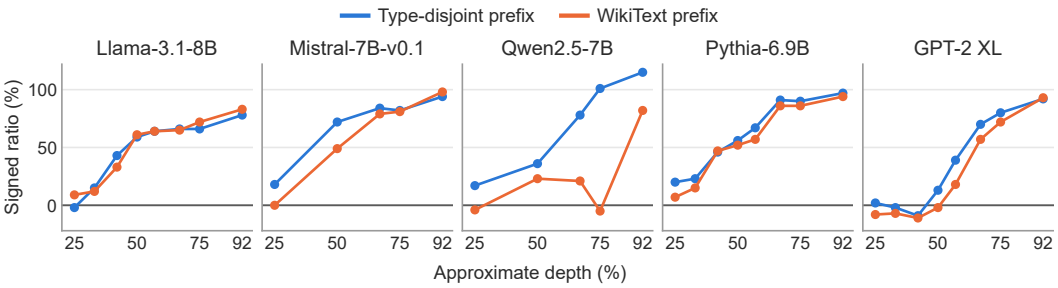


Figure B.7: **Single-layer END-residual restoration.** Signed gold-logit restoration ratio. Points are the reported layers nearest the indicated depth fractions; segments connect observations without smoothing. The denominator convention is Eq. (B.5).

C.1.2 Gold and distractor support

Literal-distractor attribution rises against both bare and token-exact neutral-filler references (Figure C.1). The gold contrast depends on the reference: the filler comparison retains a positive decrease for the long-prefix Qwen and GPT-2 XL conditions. Table C.2 gives the separate vocabulary-controlled readouts for I.

Table C.1: **Population stability and attention shifts.** Depth counts and stability describe I; inclusion and Jaccard use relation-stratified discovery resamples. Attention shifts describe $I^\star$ under short prefixes relative to bare queries and show selected source regions.

Model	Early/mid/late (I)	Inclusion ≥ 0.8	Jaccard	Δa_{prefix}	Δa_{BOS}	Δa_{rel}
Llama-3.1-8B	1/2/17	13/20	0.666	+0.09	−0.08	−0.01
Mistral-7B-v0.1	0/4/16	12/20	0.662	+0.10	−0.10	−0.00
Qwen2.5-7B	0/1/19	14/20	0.615	+0.23	−0.09	−0.09
Pythia-6.9B	0/6/14	10/20	0.487	+0.14	−0.09	−0.03
GPT-2 XL	2/8/10	12/20	0.587	+0.25	−0.11	−0.09

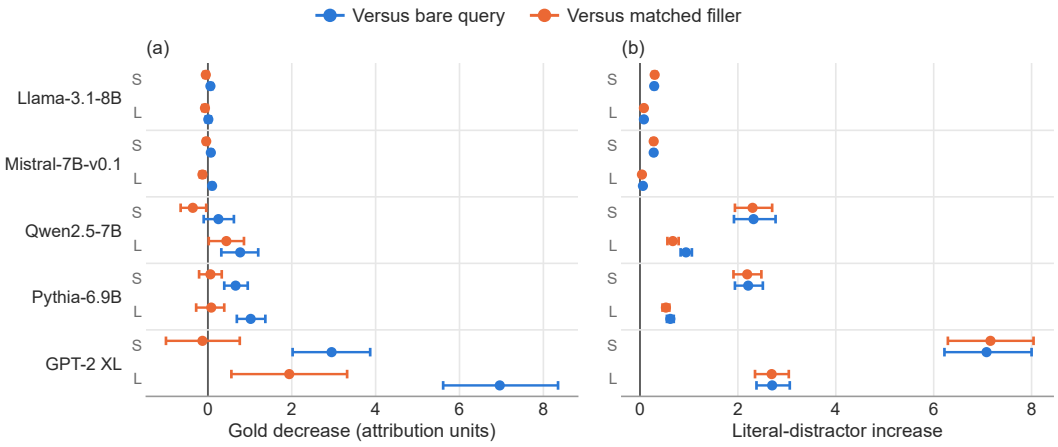


Figure C.1: **Reference-controlled attribution in $I^\star$.** Each row shows short (S) or long (L) prefixes. Gold decrease is reference minus irrelevant; distractor increase uses the reverse contrast. Readouts sum heads with fixed final-normalization weights. Bars are 95% intervals; axes are shared across models within each readout.

Table C.2: **Vocabulary-controlled attribution in I.** Irrelevant minus filler, after subtracting frequency- and norm-matched vocabulary controls. These point estimates use the unfiltered population; they are separate from the $I^\star$ contrasts above.

Model	Literal tokens	Distractor category	Gold	Query category	Other answers
Llama-3.1-8B	+0.2794	+0.1182	+0.0515	+0.0277	+0.0453
Mistral-7B-v0.1	+0.2881	+0.1080	+0.2178	+0.0158	+0.0138
Qwen2.5-7B	+0.1440	+0.0233	+0.0101	−0.0153	−0.0062
Pythia-6.9B	+0.2778	+0.0970	+0.0792	+0.0230	+0.0202
GPT-2 XL	+0.2428	+0.1349	−0.0092	+0.0121	+0.0089

C.2 Intrusion writes and interventions

C.2.1 Prefix and non-prefix contributions

The split $o_h = P_h + U_h$ separates prefix and remaining-source contributions. Prefix writes account for about 52%, 32%, 94%, 68%, and 74% of increased literal-distractor attribution, respectively. The net gold reduction lies in non-prefix terms; the END-position term is largest in Llama, Mistral, and Qwen (Table C.3).

Table C.3: **Source-resolved attribution in $I^\star$.** Short irrelevant prefix minus bare query, with means of head-summed readouts. P is the prefix contribution; ΔU partitions changes at the remaining sources. Entries are rounded independently.

Model	Target	Total	P	ΔU_{subj}	ΔU_{rel}	ΔU_{END}	ΔU_{BOS}
Llama-3.1-8B	Gold	-0.06	-0.00	$+0.04$	-0.03	-0.07	-0.00
	D_{lit}	$+0.29$	$+0.15$	$+0.03$	$+0.10$	$+0.02$	$+0.00$
Mistral-7B-v0.1	Gold	-0.07	-0.01	$+0.02$	$+0.05$	-0.12	$+0.00$
	D_{lit}	$+0.28$	$+0.09$	$+0.03$	$+0.15$	$+0.01$	$+0.00$
Qwen2.5-7B	Gold	-0.25	$+0.07$	-0.06	$+0.05$	-0.33	$+0.02$
	D_{lit}	$+2.32$	$+2.19$	$+0.04$	$+0.09$	$+0.01$	-0.01
Pythia-6.9B	Gold	-0.66	$+0.00$	-0.28	-0.23	-0.15	-0.01
	D_{lit}	$+2.21$	$+1.51$	$+0.17$	$+0.47$	$+0.07$	$+0.00$
GPT-2 XL	Gold	-2.95	-0.04	-1.08	-1.41	-0.37	-0.06
	D_{lit}	$+7.08$	$+5.22$	$+0.48$	$+1.30$	$+0.16$	-0.08

C.2.2 CLEAN-WRITE RESTORATION

Table C.4 replaces I's END writes with bare-query donors while retaining the prefix. It reports normalized movement toward the no-context gold logit. Figure C.2 varies replacement strength in a separately reported sweep.

Table C.4: **No-context write restoration and controls.** Point estimates. The ratio is $100(\ell_{\text{edited}} - \ell_{\text{ctx}})/(\ell_{\text{nctx}} - \ell_{\text{ctx}})$. Controls contain twenty random, layer-, score-decile-, or prefix-attention-decile-matched heads. A positive value denotes movement toward the bare-query logit, including removal of helpful-context gain when context raises that logit.

Model	Context	I (20)	Random	Layer	Score	Attention
Llama-3.1-8B	Short irrelevant	24.2	1.9	6.6	7.3	2.2
	Long irrelevant	22.5	-0.1	7.2	6.0	1.0
	Golden	15.6	-3.0	-1.3	6.2	4.2
Mistral-7B-v0.1	Short irrelevant	17.9	3.6	7.1	10.8	5.8
	Long irrelevant	22.6	3.7	6.1	8.0	3.5
	Golden	1.7	-0.2	8.1	16.4	1.2
Qwen2.5-7B	Short irrelevant	-12.0	-0.5	-2.1	6.2	9.5
	Long irrelevant	73.6	-5.8	5.9	12.5	1.0
	Golden	41.3	0.8	10.3	26.3	9.3
Pythia-6.9B	Short irrelevant	57.8	2.9	9.2	23.0	11.9
	Long irrelevant	31.7	2.1	4.7	14.3	2.9
	Golden	15.7	-5.6	-1.9	15.6	-0.5
GPT-2 XL	Short irrelevant	93.0	4.7	16.0	40.4	6.5
	Long irrelevant	32.1	2.2	6.0	13.5	1.5
	Golden	14.0	-1.0	3.6	16.6	5.5

C.2.3 PREFIX REMOVAL AND NON-PREFIX AMPLIFICATION

For fixed head inputs and prefix attention mass $a_h < 1$, the four replacements are

$$o_h^{\text{remove}} = U_h, \qquad o_h^{\text{amplify}} = P_h + \frac{U_h}{1 - a_h},$$
$$o_h^{\text{both}} = \frac{U_h}{1 - a_h}, \qquad o_h^{\text{gain}} = \frac{P_h + U_h}{1 - a_h}. \tag{C.1}$$

The first two terms separate the local effects of prefix removal and attention renormalization. Figure C.3 measures their immediate attributions; Table C.5 measures the resulting final predictions. In Mistral, whole-write gain increases gold log-probability more than the combined edit; its short-prefix distractor logit also increases.

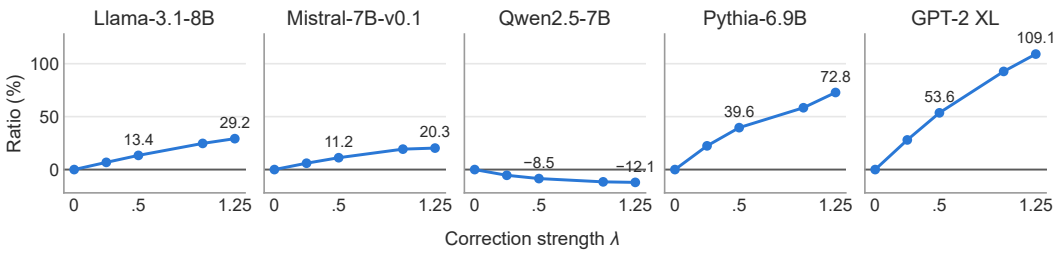


Figure C.2: **Correction-strength sweep.** Point estimates for $o_h^{\text{ctx}} + \lambda(o_h^{\text{nctx}} - o_h^{\text{ctx}})$. All five panels use the same ratio axis. The sweep is separate from Table C.4.

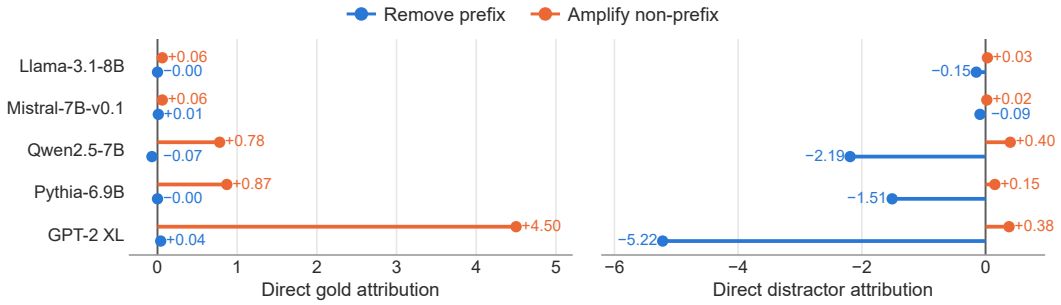


Figure C.3: **Immediate effects on edited writes.** Short-prefix direct-attribution changes summed over $I^\star$, averaged over held-out facts. Points show the individual removal and amplification terms; they are distinct from final prediction changes.

Table C.5: **Four-write interventions in $I^\star$.** Each cell gives gold log-probability change / strongest literal-distractor logit change, with the gold 95% interval beneath. Changes are relative to the unedited run. Each column is a separate intervention; effects are not additive.

Model	Prefix	Remove prefix	Amplify non-prefix	Both	Whole-write gain
Llama-3.1-8B	Short	$+0.004 / -0.14$ $[-0.004, +0.014]$	$+0.002 / -0.01$ $[-0.001, +0.005]$	$+0.006 / -0.14$ $[-0.001, +0.014]$	$+0.002 / +0.02$ $[-0.004, +0.007]$
	Long	$-0.018 / -0.01$ $[-0.031, -0.005]$	$-0.002 / -0.00$ $[-0.019, +0.009]$	$-0.022 / -0.01$ $[-0.041, -0.005]$	$-0.033 / -0.34$ $[-0.068, +0.001]$
Mistral-7B-v0.1	Short	$-0.003 / -0.22$ $[-0.007, -0.000]$	$+0.014 / +0.00$ $[+0.010, +0.019]$	$+0.011 / -0.21$ $[+0.007, +0.016]$	$+0.017 / +0.06$ $[+0.012, +0.022]$
	Long	$-0.016 / -0.06$ $[-0.024, -0.008]$	$+0.029 / -0.01$ $[+0.022, +0.037]$	$+0.017 / -0.07$ $[+0.010, +0.025]$	$+0.038 / -0.15$ $[+0.020, +0.056]$
Qwen2.5-7B	Short	$-0.013 / -0.49$ $[-0.022, -0.003]$	$+0.012 / +0.03$ $[+0.005, +0.020]$	$+0.002 / -0.45$ $[-0.008, +0.014]$	$-0.086 / +0.50$ $[-0.264, +0.015]$
	Long	$-0.064 / -0.36$ $[-0.090, -0.036]$	$+0.035 / -0.07$ $[+0.015, +0.057]$	$-0.016 / -0.41$ $[-0.049, +0.016]$	$-0.831 / +0.05$ $[-1.404, -0.391]$
Pythia-6.9B	Short	$+0.026 / -0.34$ $[+0.017, +0.035]$	$+0.008 / -0.02$ $[+0.002, +0.014]$	$+0.032 / -0.35$ $[+0.021, +0.043]$	$-2.27 / +2.03$ $[-4.44, -0.59]$
	Long	$+0.051 / -0.12$ $[+0.031, +0.076]$	$+0.013 / -0.03$ $[-0.001, +0.030]$	$+0.064 / -0.14$ $[+0.034, +0.097]$	$-5.10 / +1.51$ $[-7.43, -3.03]$
GPT-2 XL	Short	$+0.044 / -0.57$ $[+0.016, +0.074]$	$+0.043 / -0.08$ $[+0.027, +0.064]$	$+0.080 / -0.62$ $[+0.043, +0.122]$	$+0.017 / +0.60$ $[-0.020, +0.064]$
	Long	$+0.156 / -0.30$ $[+0.104, +0.213]$	$+0.139 / -0.05$ $[+0.100, +0.182]$	$+0.277 / -0.34$ $[+0.197, +0.371]$	$-0.281 / +0.48$ $[-0.414, -0.157]$

C.2.4 Gold and distractor effects of the same intervention

Figure C.4 pairs the answer and distractor readouts from each forward pass. Masking increases gold log-probability in both prefix conditions for Mistral, Pythia, and GPT-2 XL; Llama's long-prefix

change is negative and Qwen's intervals include zero. Table C.7 evaluates golden context and matched heads.

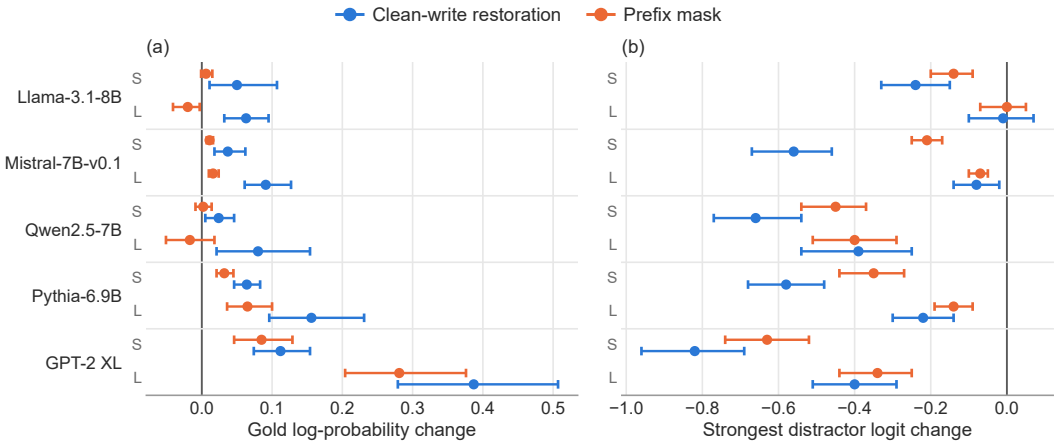


Figure C.4: **Same-pass outcomes in $I^\star$.** Restore uses bare-query END-write donors; mask blocks END-to-prefix attention without a donor. Changes are relative to the unedited irrelevant-context run, with 95% intervals. Short (S) and long (L) prefixes share axes. Gold log-probability and distractor logits are separate metrics.

Table C.6: **Joint improvement by fact.** Percentages with increased gold log-probability and decreased distractor probability / strongest distractor logit. Entries use the same passes as Figure C.4.

	Short		Long	
Model	Restore	Mask	Restore	Mask
Llama-3.1-8B	72 / 59	46 / 47	67 / 37	41 / 30
Mistral-7B-v0.1	79 / 84	81 / 52	83 / 48	73 / 32
Qwen2.5-7B	50 / 44	44 / 28	44 / 34	28 / 14
Pythia-6.9B	79 / 77	74 / 74	67 / 63	70 / 60
GPT-2 XL	88 / 84	84 / 84	77 / 71	80 / 71

Table C.7: **Golden-context effects and matched controls.** Gold log-probability changes for the same interventions. $R_{\text{only}} = R_{20} \setminus I_{20}$; control heads match the layer and discovery prefix-attention mass of $I^\star$. All entries are point estimates.

Model	Golden restore $I^\star$	Golden mask $I^\star$	Golden mask R_{only}	Short mask control	Long mask control
Llama-3.1-8B	−0.023	−0.024	−0.172	−0.020	−0.046
Mistral-7B-v0.1	−0.003	−0.008	−0.089	+0.001	−0.005
Qwen2.5-7B	−0.019	−0.021	−0.155	+0.001	−0.002
Pythia-6.9B	−0.041	−0.007	−0.075	+0.015	+0.025
GPT-2 XL	−0.012	−0.012	−0.189	+0.005	+0.033

When irrelevant and golden prefixes occur together, whole-context masking and oracle masking target different spans (Figure C.5). The whole-context $I^\star$ mask reduces gold log-probability in Llama, Mistral, and Qwen, while the oracle restricts the edit to the known irrelevant span.

C.3 Direct prefix access and the copier-to-query test

The tested copier-to-query substitution changes gold log-probability by -0.005 to $+0.001$. Whole-query donors and aligned filler-prefix K/V replacements provide broader comparisons (Figure C.6). K/V substitution recomputes attention and improves gold log-probability in both prefix conditions for

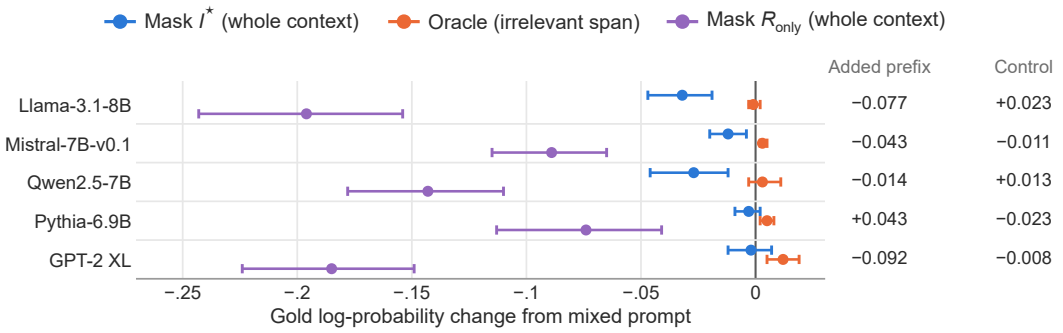


Figure C.5: **Useful evidence with distraction.** The irrelevant prefix precedes the golden prefix and query. Points and 95% intervals compare edits with the unedited mixed prompt; $I^\star$ and R_{only} masks block all context, while the oracle uses the irrelevant-span boundary. The added-prefix column compares mixed with golden alone. The matched-control column is a point estimate relative to mixed.

Mistral, Pythia, and GPT-2 XL. The experiment isolates the specified copier contribution; the relative query-norm changes are 0.008–0.035 (Table C.8).

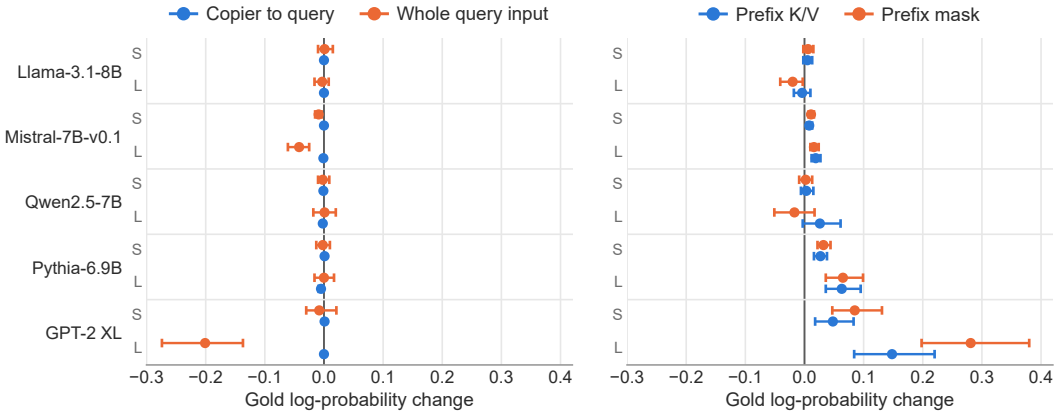


Figure C.6: **Query-input and prefix-access interventions.** Gold log-probability changes with 95% intervals. Query substitutions use bare-query donors; prefix K/V uses token-length-matched filler. END is unpatched in the K/V intervention. Both panels use the same axis. Small changes in total prefix attention do not imply fixed source-wise attention.

C.4 Entrainment heads

In Llama-3.1-8B, neither the published 36-head country–capital entrainment set (Niu et al., 2025) nor its 29-head reproduction overlaps the top twenty intrusion heads. The uniform expected overlap is 0.70 (Table C.9). Their depth profiles differ: 17 of 20 intrusion heads and 3 of 36 published entrainment heads lie in the last third.

C.5 Contextual retrieval

Retrieval ranking uses attention to the golden answer position weighted by the gold-directed write. The global top-twenty rankings overlap in 3, 5, 6, 5, and 3 heads, respectively. Table C.10 compares the resulting I_{only} and R_{only} sets, not the twenty-head $I^\star$ population. Their sizes are 14–17 after exclusion. Intrusion-only masking increases the gold logit in Llama, Mistral, Pythia, and GPT-2 XL under both irrelevant conditions; Qwen's changes are negative. Golden-context changes are reported alongside the irrelevant-context changes in the same logit units.

Table C.8: **Query and attention changes.** Relative query-norm changes and mean END-to-prefix attention-mass changes under the same substitutions. The copier contribution is small relative to the full query, and prefix K/V substitution changes what the heads read with little change in total prefix attention. Values are point estimates.

		Copier→query		Prefix K/V
Model	Prefix	$\|\Delta q\|/\|q\|$	Δa_{prefix}	Δa_{prefix}
Llama-3.1-8B	Short	0.008	+0.000	+0.002
	Long	0.010	+0.001	−0.001
Mistral-7B-v0.1	Short	0.012	+0.001	+0.001
	Long	0.013	+0.002	−0.001
Qwen2.5-7B	Short	0.018	+0.001	−0.000
	Long	0.023	+0.001	−0.002
Pythia-6.9B	Short	0.025	+0.001	+0.001
	Long	0.027	+0.005	+0.002
GPT-2 XL	Short	0.025	+0.000	+0.000
	Long	0.035	+0.001	+0.005

Table C.9: **Overlap with the published entrainment set.** Shared heads at each cutoff of the Llama-3.1-8B intrusion ranking; expected counts use uniformly sampled head sets.

Top intrusion k	5	10	20	40	60	100	150
Shared heads	0	0	0	1	3	5	7
Expected	0.18	0.35	0.70	1.41	2.11	3.52	5.27

Table C.10: **Contextual retrieval and intrusion.** Gold-logit change under END-to-prefix masking, relative to the unedited condition. Both exclusive populations contain n heads per model; random sets match this size and the intrusion-only layer distribution.

Model	Masked population	Golden	Short irrelevant	Long irrelevant
Llama-3.1-8B	Intrusion-only	−0.056	+0.052	+0.070
($n = 17$)	Retrieval-only	−0.494	+0.027	+0.028
	Random	+0.042	−0.002	−0.002
Mistral-7B-v0.1	Intrusion-only	−0.083	+0.001	+0.054
($n = 15$)	Retrieval-only	−0.276	+0.022	+0.031
	Random	−0.015	−0.009	−0.029
Qwen2.5-7B	Intrusion-only	−0.233	−0.107	−0.030
($n = 14$)	Retrieval-only	−0.370	−0.004	+0.010
	Random	+0.016	−0.007	+0.001
Pythia-6.9B	Intrusion-only	−0.108	+0.159	+0.087
($n = 15$)	Retrieval-only	−0.068	+0.068	+0.035
	Random	+0.026	+0.050	+0.004
GPT-2 XL	Intrusion-only	−0.034	+0.150	+0.369
($n = 17$)	Retrieval-only	−0.299	+0.022	+0.087
	Random	−0.002	−0.018	+0.010

D Downstream Evaluation

We compare clean accuracy C, distracted accuracy D, and edited accuracy E for oracle residual patching and the K10/K20 head masks. K10 denotes ten intrusion-ranked heads after retrieval-head exclusion; K20 denotes the full top-20 ranking, including overlapping heads. These are separate from the 20-head disjoint population and the exclusive-set comparisons in Appendix C.

Oracle patching replaces the END-position `resid_pre` at each decoding step with a matched clean-query state. The reported layers are $L = 30$ for Llama-3.1-8B, Mistral-7B-v0.1 and Pythia-6.9B, and $L = 26$ for Qwen2.5-7B. The larger models use layer-swept settings. This intervention provides a clean-state recovery reference.

Table D.1: Cross-task overlap of intrusion-head sets. The reference is the same model's factual-recall ten-head set; each entry is $|H^{(10)}_{\text{factual}} \cap H^{(10)}_{\text{task}}|/10$. *Reference set.

Task	Llama	Mistral	Qwen	Pythia
Factual recall*	10/10	10/10	10/10	10/10
IOI (retrieval)	10/10	10/10	9/10	10/10
GSM-DC	10/10	9/10	9/10	10/10
GSM-IC	10/10	9/10	9/10	10/10
HotpotQA	10/10	10/10	10/10	10/10
PubMedQA	10/10	10/10	10/10	10/10

D.1 GSM-DC

Figure D.1 reports GSM-DC accuracy by model (Yang et al., 2025). The clean baseline varies across k; comparisons are therefore made within each plotted condition. The figure supplies clean, distracted and patch outcomes.

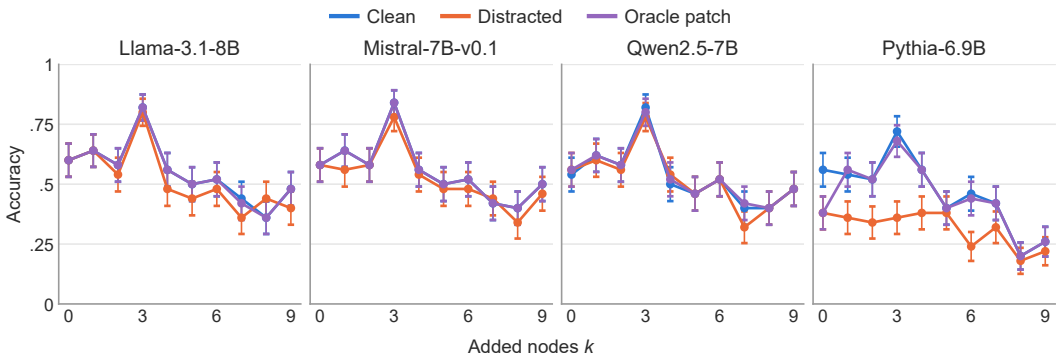


Figure D.1: GSM-DC accuracy at the ten plotted settings $k = 0, \ldots, 9$.

D.2 GSM-IC: Single-Token Evaluation

The single-token evaluation measures argmax correctness at the first answer token on GSM-IC (Shi et al., 2023). Patched accuracy equals or exceeds clean accuracy in each row (Table D.2).

Table D.2: Single-token accuracy. Gain is $100(E - D)$ percentage points (pp). Recovery uses the clean–distracted gap; a value above 100% corresponds to $E > C$.

Model	Clean	Distracted	Patch	Patch gain (pp)	Recovery (%)
Llama-3.1-70B	.578	.408	.656	+24.8	145.9
Qwen2.5-72B	.900	.840	.932	+9.2	153.3
Qwen2.5-32B	.888	.810	.924	+11.4	146.2
Qwen2.5-7B	.688	.638	.730	+9.2	184.0
Mistral-7B-v0.1	.456	.352	.456	+10.4	100.0

D.3 GSM-IC: Chain-of-Thought Evaluation

K10 improves accuracy by 13 and 12 percentage points for Mistral and Llama, respectively; K20 improves it by 4 and 7 points (Table D.3).

Accuracy changes and recovery. For every observed condition we report $\Delta\text{acc} = 100(E - D)$ in percentage points. Conditional loss recovery is

$$R = 100\frac{E - D}{C - D}, \qquad C > D. \tag{D.1}$$

Table D.3: GSM-IC chain-of-thought accuracy. Relative to the distractor-induced loss, K10/K20 recover 46.4%/14.3% on Mistral and 70.6%/41.2% on Llama. These percentages use $C - D$, not the gain from oracle patching.

Model	Clean	Distracted	Patch	K10	K20
Mistral-7B-v0.1	.66	.38	.66	.51	.42
Llama-3.1-8B	.64	.47	.67	.59	.54

When $C \leq D$, the condition remains in the absolute-effect analysis but has no loss-recovery value. Table D.4 compares all three interventions on the same observed cells within each task. Each cell receives equal weight.

Table D.4: Common-cell summaries: n counts cells with all three intervention results; n_+ further requires $C > D$.

		Mean change (pp)				Mean loss recovery (%)		
Task	n	Patch	K10	K20	n_+	Patch	K10	K20
GSM-IC CoT	2	+24.00	+12.50	+5.50	2	108.82	58.51	27.73
HotpotQA	15	+23.07	+14.53	+9.27	13	85.51	48.54	29.53
PubMedQA	11	+15.64	+5.45	+3.09	9	99.26	45.31	17.67

D.4 HotpotQA

HotpotQA (Yang et al., 2018) uses two supporting passages and k added non-supporting paragraphs. At $k = 8$, K10 gains 29, 34, 30 and 20 percentage points on Llama, Mistral, Qwen and Pythia, respectively. Figure D.2 shows the full response across k, including unchanged and adverse effects.

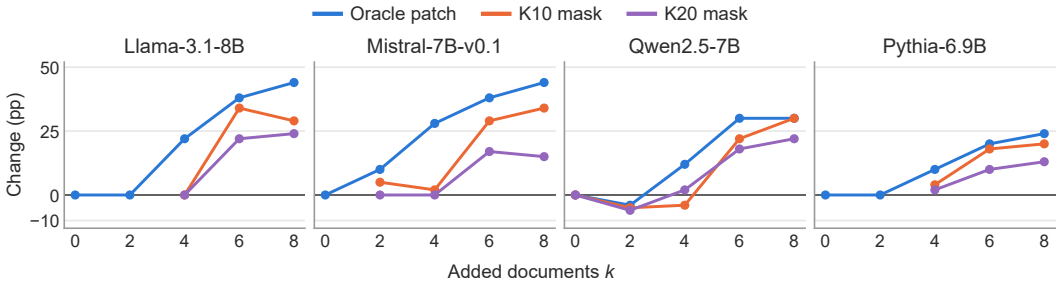


Figure D.2: Edited-minus-distracted exact-match accuracy on HotpotQA. Lines connect observed conditions. All panels use the same scale, and zero denotes no change.

D.5 PubMedQA

PubMedQA (Jin et al., 2019) appends k irrelevant abstracts to the supporting context. At $k = 4$, K10 increases accuracy by 7, 21 and 26 percentage points on Llama, Mistral and Qwen. The same comparison includes a -9 point effect on Pythia at $k = 1$, where the added abstract had improved baseline accuracy.

E Base and Instruction-Tuned Checkpoint Comparison

We compare Llama-3.1-8B and Llama-3.1-8B-Instruct on identical raw-continuation prompts, without a chat template. The shared architecture permits comparison of the same layer/head indices. Table E.1 reports late-block ranking correlations and set overlaps for this checkpoint comparison; it is separate from the global head-set analysis in Appendix C.

Table D.5: HotpotQA exact-match accuracy.

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

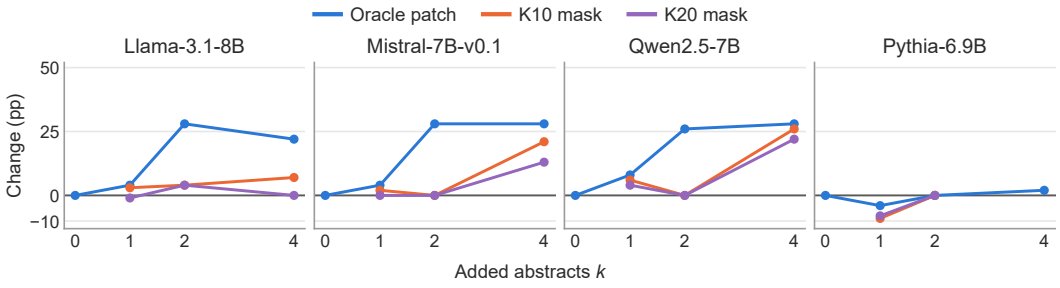


Figure D.3: Edited-minus-distracted accuracy on PubMedQA, with the same scale as the HotpotQA panels. At Pythia $k = 1$, $C = .62 < D = .66$; K10/K20 produce .57/.58, corresponding to $-9/-8$ pp, not positive loss recovery.

Head-mask effects. Table E.2 reports the class-conditional interventions for the two checkpoints. The short-prefix K10 changes are +.009 and +.006 logits; the long-prefix changes are +.015 and $-.017$. K20 yields positive irrelevant-context changes in both checkpoints. Bare-query invariance follows from the absence of context edges.

Oracle patching across tasks. With residual patching at $L = 30$, the instruction-tuned checkpoint recovers 94.7–100.0% of the clean–distracted gap on the four task summaries (Table E.3). This comparison evaluates oracle state substitution.

F Sensitivity to Head Count and Retrieval Penalty

We vary the head count $K \in \{5, 8, 10, 12, 15, 20\}$ and introduce a retrieval penalty $\alpha_{\text{sel}} \in \{0, .5, 1, 2\}$ in the ranking

$$s_{\text{sel}}(h) = z(s_{\text{int}}(h)) - \alpha_{\text{sel}}\, z(s_{\text{ret}}(h))\,. \tag{F.1}$$

Here z standardizes each score across candidate heads. The baseline ranking uses $\alpha_{\text{sel}} = 0$; this penalty is separate from any restoration-strength parameter. We plot short-irrelevant gold-logit changes and the accompanying golden-context scores as separate quantities.

Table D.6: PubMedQA accuracy. At $k = 2$, K10 differs from the distracted baseline on Llama (.54 to .58) and is unchanged on the other three models.

Model	Condition	$k = 0$	$k = 1$	$k = 2$	$k = 4$
Llama-3.1-8B	Clean	0.82	0.82	0.82	0.82
	Distracted	0.82	0.78	0.54	0.60
	Oracle patch	0.82	0.82	0.82	0.82
	K10 mask	—	0.81	0.58	0.67
	K20 mask	—	0.77	0.58	0.60
Mistral-7B	Clean	0.86	0.86	0.86	0.86
	Distracted	0.86	0.82	0.58	0.58
	Oracle patch	0.86	0.86	0.86	0.86
	K10 mask	—	0.84	0.58	0.79
	K20 mask	—	0.82	0.58	0.71
Qwen2.5-7B	Clean	0.74	0.74	0.74	0.74
	Distracted	0.74	0.66	0.48	0.44
	Oracle patch	0.74	0.74	0.74	0.72
	K10 mask	—	0.72	0.48	0.70
	K20 mask	—	0.70	0.48	0.66
Pythia-6.9B	Clean	0.62	0.62	0.62	0.62
	Distracted	0.62	0.66	0.62	0.60
	Oracle patch	0.62	0.62	0.62	0.62
	K10 mask	—	0.57	0.62	—
	K20 mask	—	0.58	0.62	—

Table E.1: Checkpoint-level population comparison. Correlations compare per-head scores; Jaccard values compare selected memberships. These statistics do not by themselves establish transfer of a frozen base-model mask.

Metric	Base	Instruct / comparison
$\lvert R \cap I \rvert$, top-20 late-block sets	5/20	5/20
Spearman, intrusion-score rankings	—	.924
Spearman, retrieval-score rankings	—	.962
Jaccard, top-20 retrieval sets	—	.818
Jaccard, top-20 intrusion sets	—	.481
Jaccard, disjoint top-10 sets	—	.429

Table E.2: Gold-logit changes, knockout minus unedited baseline. The golden column uses this same sign convention.

Head set	Checkpoint	Short $\Delta\ell_g$	Long $\Delta\ell_g$	Golden $\Delta\ell_g$	Bare
Disjoint K10	Base	+.009	+.015	+.010	.000
	Instruct	+.006	−.017	+.039	.000
Full K20	Base	+.149	+.151	+.161	.000
	Instruct	+.035	+.048	+.113	.000

Table E.3: Task summaries for oracle patching at $L = 30$, averaged over conditions with $C > D$. The base GSM-IC denominator is approximately zero.

Task	Base recovery (%)	Instruct recovery (%)
GSM-IC, single-token	Undefined	94.7
GSM-DC	87.5	100.0
HotpotQA	89.8	97.7
PubMedQA	100.0	100.0

F.1 Head-Count Sweep

The disjoint and full rankings yield different response curves (Figure F.1). The disjoint-set short-prefix change at $K = 20$ is +.054, +.042, +.038 and +.209 logits for Llama, Mistral, Qwen and Pythia. The complete curves retain the changes of direction between adjacent values of K.

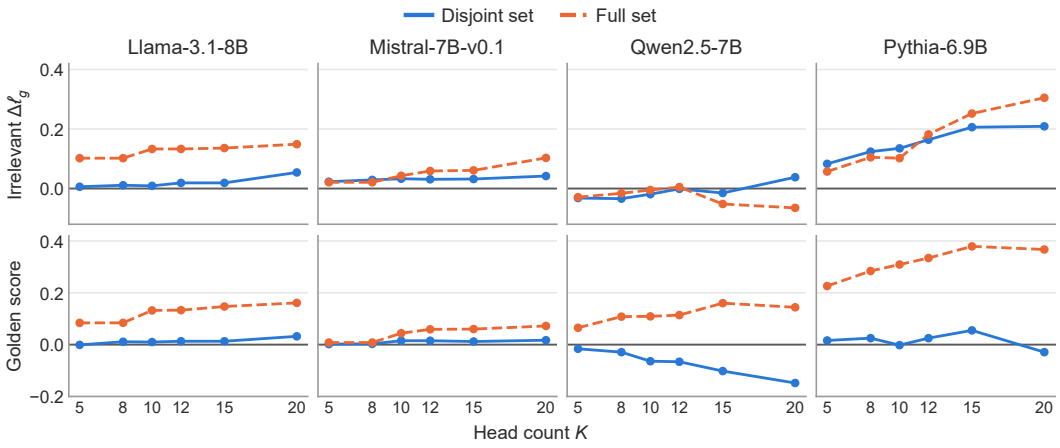


Figure F.1: Head-count sweep at zero retrieval penalty. Within each model, the upper panel shows short-irrelevant $\Delta\ell_g$ (edited minus baseline); the lower panel shows the golden-context score.

F.2 Joint Head-Count and Penalty Sweep

Figure F.2 presents every setting. Zero penalty gives the largest $K = 10$ short-prefix change on Llama, Mistral and Pythia; Qwen's largest value occurs at penalty .5. The curves describe sensitivity over the evaluated settings, with no smoothing or removal of negative outcomes.

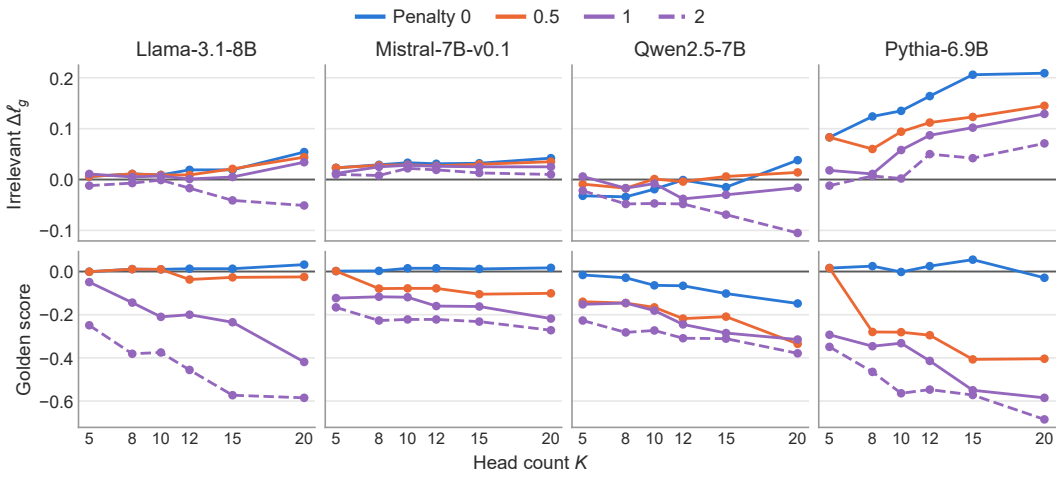


Figure F.2: Complete $K \times \alpha_{\text{sel}}$ grid for the disjoint ranking. Blue, orange and purple distinguish penalties; the dashed purple curve denotes penalty 2. Upper panels show short-prefix $\Delta\ell_g$; lower panels show the golden-context score. These are point estimates.