SUPPRESSION IS NOT CORRECTION: CAUSAL TRACING OF RAG CORPUS POISONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) can be poisoned by malicious corpus passages that steer models toward answers chosen by attackers. Existing studies show that these attacks are effective, but do not explain how poisoned evidence changes the model internally or why removing its influence may fail to recover the correct answer. We present a causal tracing framework that follows retrieved evidence from document positions through the residual stream to the final answer score. Across five LLMs, we identify a six-layer residual evidence band whose state causally shifts the competition between attacker and correct answers. The same functional interface appears across models, while its downstream readout varies by architecture. Some models rely partly on a small set of attention heads, whereas others use more distributed pathways. Source and position controls show that the band carries answer-bearing document information rather than a generic signal that a document is present. On 106 answerable attacks, strong oracle steering suppresses the attacker answer in 92.5% of cases but recovers the correct answer in only 13.2%. Restoring clean residual states within the six-layer band recovers 71.7%, nearly matching the 72.6% recovery from restoring all layers. These results show that suppressing poisoned evidence is not sufficient for correction. Correct recovery requires restoring positive support for the right answer, and the residual band provides a compact causal interface for doing so.

1 INTRODUCTION

Retrieval-augmented generation (RAG) gives LLMs access to external evidence that can be updated and inspected directly (Lewis et al., 2020; Karpukhin et al., 2020). Yet the generator does not control who wrote these passages. Corpus poisoning exploits this gap by inserting malicious passages into the retrieval corpus (Zou et al., 2025; Shao et al., 2025; Wang et al., 2026). These passages are designed to be retrieved for a target question and to steer the model toward an answer chosen by the attacker. They compete directly with legitimate sources, so even a small amount of poisoned evidence can change the response.

Most studies measure this threat through attack success rate (ASR), while internal-state analyses also identify activation signatures of poisoning (Tan et al., 2025; Sun et al., 2025). ASR and activation separability alone do not establish how retrieved evidence causes an answer to change. A causal explanation must connect document positions to internal evidence states, their downstream readout, and the final answer.

Finding such a state does not yet tell us how to recover the correct answer. Weakening support for the attacker answer may leave another incorrect answer preferred. Recovery requires sufficient support for the correct answer, whether from retrieved evidence or parametric knowledge. To explain correction, we must examine what an intervention does to the evidence state, as well as where it acts.

We trace retrieved evidence from document positions through the residual stream to the final answer score (Figure 1). Across five LLMs, we identify a residual evidence band that spans six layers and causally shifts the competition between attacker and correct answers. Its location and downstream readout vary across models, but its functional role is shared. We then compare two oracle interventions that act on this interface but differ in how they supply clean evidence: steering applies a pooled direction, whereas restoration reinstates clean states at matching document positions. Both have access to paired clean evidence, so the comparison tests how the form of the intervention affects

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.

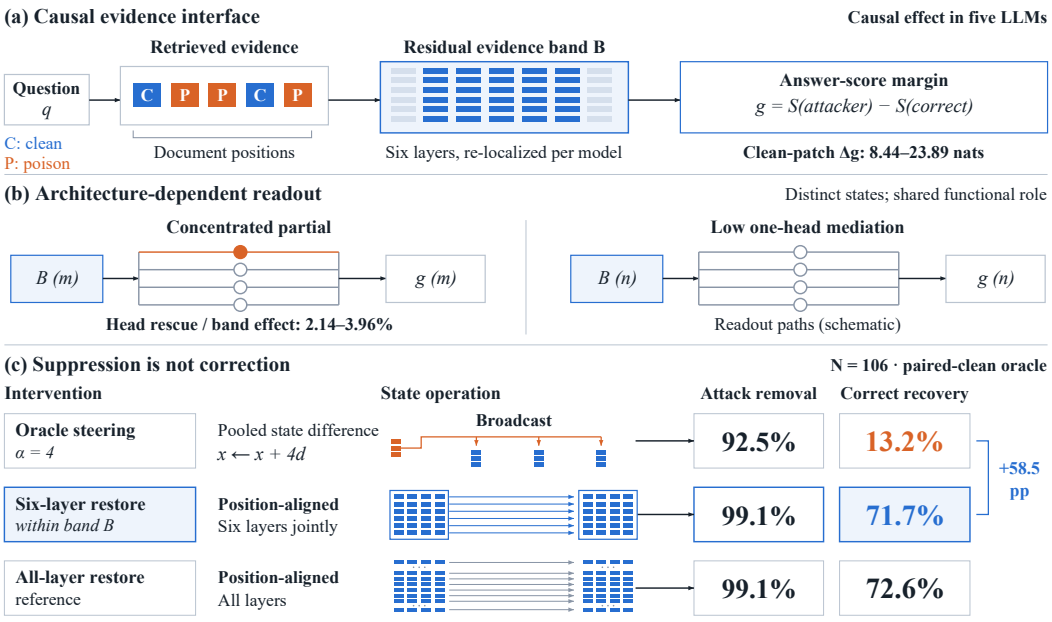


Figure 1: Overview. Retrieved evidence modifies a model-specific residual evidence band with architecture-dependent selected-head contributions. Readout paths are schematic. On the common 106 answerable attacks, position-aligned oracle restoration achieves substantially higher recovery than oracle steering, showing that suppression is not correction. Recovery is measured by reference-answer matching; a separate semantic audit is reported in Table S43.

recovery. Restoring clean states within six layers nearly matches recovery from restoring all layers, identifying the band as a compact causal interface for answer-bearing evidence.

Contributions.

- We identify a residual evidence band that causally shifts the competition between attacker and correct answers across five LLMs. Source and position controls show that this band carries answer-bearing document information rather than simply signaling document presence.
- The residual evidence interface appears across models, but models read it out in different ways. Selected attention heads show architecture-dependent partial mediation. This suggests that the shared structure is functional rather than tied to a fixed layer or head.
- On 106 human-adjudicated answerable attacks, strong oracle steering removes 92.5% of attacks but recovers only 13.2% of correct answers. Position-aligned restoration of clean residual states within the six-layer band recovers 71.7%, close to 72.6% from restoring all layers. This shows that correction requires reconstructing positive, answer-bearing evidence rather than merely suppressing attacker support.

2 Related Work

Corpus poisoning attacks. RAG conditions generation on retrieved evidence (Lewis et al., 2020; Karpukhin et al., 2020), motivating studies of retrieval timing, evidence quality, and faithfulness (Asai et al., 2024; Es et al., 2024; Liu et al., 2024). PoisonedRAG shows that a few poisoned passages can induce attacker-chosen answers; PoisonCraft and Joint-GCG jointly optimize retrieval and generation (Zou et al., 2025; Shao et al., 2025; Wang et al., 2026). RAGForensics traces poisoned texts through iterative retrieval and response-conditioned LLM judgments (Zhang et al., 2025); RAGOrigin combines retrieval similarity with conditional question and response likelihoods

to score document responsibility (Zhang et al., 2026). Our interventions connect document inspection to the generator's residual evidence interface and downstream readout.

Mechanistic and internal-state analysis. RevPRAG detects poisoning from activations; ReDeEP intervenes on copying heads and knowledge FFNs to study RAG hallucinations (Tan et al., 2025; Sun et al., 2025). Retrieval-head ablations establish a causal role in long-context factual retrieval (Wu et al., 2025). Building on causal tracing, activation patching, circuit discovery, and steering (Meng et al., 2022; 2023; Wang et al., 2023; Conmy et al., 2023; Belrose et al., 2023; Elhage et al., 2021; Li et al., 2023; Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024), we test poisoned evidence through source/position controls, held-out nested mediation, and correction-operator comparisons.

Robust RAG and control. Work on noisy, conflicting, or untrusted evidence develops robustness benchmarks, evidence-guided decoding, retrieval-stage sanitization, refusal, and abstention (Chen et al., 2024; Sui et al., 2026; Zhou et al., 2026; Yang et al., 2026; Liang et al., 2026; Xiang et al., 2024; Wang et al., 2025; Zhou et al., 2025). These methods change which evidence is admitted or how the model responds. Our analysis instead tests which retrieved document enters an evidence-to-answer pathway, then asks why deleting its negative influence need not restore a positive answer source.

3 METHODOLOGY

Threat model. We study targeted, inject-only corpus poisoning (Zou et al., 2025; Shao et al., 2025; Wang et al., 2026). A gray-box adversary chooses a target question and incorrect answer and can add a fixed number of passages using open-weight surrogates, but cannot alter model weights, decoding, the system prompt, legitimate documents, or the production retriever. Poison acts as retrieved *evidence*, not as an instruction override, prompt injection, or model backdoor.

Under this threat model, we treat RAG corpus poisoning as a causal internal process and apply one four-stage analysis uniformly to every model: locate candidate residual bands and readout sites, validate their link on held-out data, interpret what the band carries, and compare suppression with correction. This yields a topology-level account even when model coordinates differ.

Notation. For prompt P, let y_{inc} and y_{cor} be the complete attacker and reference answers. We define the *poisoned-answer gap* from teacher-forced token log probabilities,

$$g = S(y_{\mathrm{inc}} \mid P) - S(y_{\mathrm{cor}} \mid P),$$
$$S(y \mid P) = \sum_{j=1}^{|y|} \log p(y_j \mid P, y_{<j}). \tag{1}$$

We use no length normalization and retain the higher score of the original and leading-space forms. Thus $g > 0$ favors the attacker answer. Head-output interventions use the final prompt position unless noted otherwise; residual interventions use the specified token masks. For attention head (ℓ, h) at layer ℓ, write its last-token attention weights $a_{\ell,h,t}$ over prompt positions t, its value-weighted output $o_{\ell,h} \in \mathbb{R}^d$, and the residual stream $x_\ell \in \mathbb{R}^d$. Let $\mathcal{P}$ index the poisoned-passage token positions and $\mathcal{B}$ the *residual evidence band*, a contiguous six-layer region. We call its role in carrying document information into answer computation the *evidence interface*. An evidence state is the activation content at that interface for a particular model and prompt. We write $g\,|_{u \leftarrow s}$ for the gap recomputed after overwriting an internal quantity u with a source s.

Route identification. We rank candidate heads from last-token poison-attention mass, poison-versus-clean mass, and incorrect-answer-anchor contrasts. Causal validation then subtracts the value-weighted contribution $c^{\mathcal{P}}_{\ell,h}$ of all poison tokens from the candidate's last-token output, $\Delta g^{\mathrm{path}}_{\ell,h} = g - g\,|_{o_{\ell,h} \leftarrow o_{\ell,h} - c^{\mathcal{P}}_{\ell,h}}$. The reported poison-$\Delta$ share is the fraction of the positive poisoned-versus-noisy head-output projection attributable to poison tokens. Held-out cross-fit discovers heads on one sample half and fixes the top eight before evaluation on the other. For the four full-layer scans, a 200-split cross-fit ranks windows on one half and evaluates them on the other.

Held-out nested mediation. We test whether the selected head mediates part of its evidence band's effect. Forty discovery questions are set aside, leaving 60 for held-out testing. On baseline-attacked cases, paired clean residuals replace the six-layer band, after which we restore either the selected head's original poisoned-run last-token output or that of one of eight fixed same-layer controls. For sample i and head h,

$$\begin{aligned} T_i &= g_i - g_i\,|_{\mathcal{B}\leftarrow\text{clean}}, \\ R_i^h &= g_i\,|_{\mathcal{B}\leftarrow\text{clean},\, o_h\leftarrow o_h^{\text{poison}}} - g_i\,|_{\mathcal{B}\leftarrow\text{clean}}, \end{aligned} \tag{2}$$

Define the unrescued remainder as $U_i^h = T_i - R_i^h$. The resulting route decomposition and mediated fraction are

$$\overline{T} = \overline{R^h} + \overline{U^h}, \qquad m_h = \frac{\overline{R^h}}{\overline{T}}. \tag{3}$$

Here $\overline{R^h}$ is selected-head rescue and $\overline{U^h}$ is the band effect left unrescued by that head. We report m_h, the selected-minus-mean-control rescue, an exact paired sign test, and the selected head's rank among all nine. This nested intervention directly tests the band→head→answer-score pathway after discovery.

Causal validation by residual-source substitution. At retrieved-document token positions, we replace band-layer outputs with zero, neutral, noisy, clean, a cross-sample document-position mean (mean-across), or the same prompt's non-context background mean (mean-within). Writing $x_\ell \leftarrow s$ for one $\ell \in \mathcal{B}$, the *band-replacement gap reduction* and *selected-head content sensitivity* are

$$\Delta g(s) = \frac{1}{|\mathcal{B}|}\sum_{\ell\in\mathcal{B}} \left(g - g\,|_{\ell\leftarrow s}\right), \qquad \beta(s) = \frac{1}{|\mathcal{B}|}\sum_{\ell\in\mathcal{B}} \cos\!\left(o^{\text{sel}}\,|_{\ell\leftarrow s},\, o_0^{\text{sel}}\right), \tag{4}$$

Source and position sweeps average these single-layer effects within each question; nested mediation and oracle band restoration replace all six layers jointly. Here o_0^{sel} is the selected-head output on the unmodified poisoned run. The discriminating quantity is $1-\beta(\text{clean})$: a content-reading head moves when the band is swapped from poisoned to clean, whereas an inert head does not. Among eight same-layer and eight random-layer controls, this sensitivity ranks first in four models and outranks 15 of 16 controls in Llama 2; raw zero-source collapse ($\beta \to 0$) is shared by neighboring heads and is an upstream-state dependency check.

Source-direction decomposition. Mean-across averages poisoned-document residuals at matched context slots and token positions, while mean-within broadcasts the same prompt's non-context mean over document positions; together they provide a descriptive contrast between document-matched and background activation sources. Poison-free neutral/noisy/clean substitutions test source content, and a Llama 3 answer-only versus answer-removed contrast tests whether an answer mention carries a correction direction.

Direction-vector intervention: suppression versus correction. For each sample and band layer, the local oracle vector d_ℓ is the pooled clean-context document residual minus its poisoned-context counterpart; the correct-answer label is used only for evaluation. We add it at document positions, $x_\ell \leftarrow x_\ell + \alpha d_\ell$, with $\alpha = 2$ as the primary local point. The stronger $\alpha = 3, 4$ operators use this same oracle direction to suppress attacker support; they do not simply delete a retrieved passage. Over baseline-attacked cases, ASR reduction counts post-intervention attack failures, whereas recovery counts reference-answer matches in the generated response. Reporting both separates removing poison from re-establishing the answer. Table 5 compares stronger oracle steering and multi-vector controls with fallback/rerun and aligned band/all-layer oracle clean-state restoration. Directional steering, multi-vector controls, clean restoration, and clean reruns all use paired clean evidence, whereas no-context fallback does not. Significance uses paired sign tests on the poisoned-answer margin.

Human-adjudicated fair correction comparison. To compare operators only where correction is well-defined, two annotators independently and blindly label whether clean evidence supports the reference answer and whether the clean-context response is semantically correct; disagreements are adjudicated separately, and the strict paired subset requires both. Eligibility is semantic and human-adjudicated, while operator outcomes use the same fixed string-matching rule (agreement and counts in Experiments).

Route-based document attribution. Once a model-specific route is recovered, the fixed selected-head score assigns each retrieved document its mean routed attention and ranks documents. One route-extraction forward pass produces these scores. Gap effects are answer-score nats, not ASR points, and band- and head-site effects are not numerically interchangeable.

4 EXPERIMENTS

The experiments test the residual band's causal role and content, its downstream readout, and the distinction between suppression and correction. Document attribution then tests a practical use of the recovered readout.

4.1 SETUP

Models, data, and protocol. The causal suite is Llama 3-8B, Llama 2-7B, and Vicuna-7B/13B/33B; the wider screen adds Qwen2.5-7B, Mistral-7B, and DeepSeek-7B. We use NQ, HotpotQA, MS MARCO, and 2WikiMultiHopQA (2Wiki) (Kwiatkowski et al., 2019; Yang et al., 2018; Nguyen et al., 2016; Ho et al., 2020). A 30-setting, 3,000-trial reproduction across five models, three datasets, and LM-targeted/HotFlip attacks reaches 0.936 pooled ASR. Mechanism and correction use NQ LM-targeted evidence; held-out and re-localized replications report their split sizes with each result. Unless noted, runs use Contriever top-5 retrieval, five poison candidates, seed 12, native chat templates, greedy decoding, and dot-product scoring.

Attack and prompt construction. Following PoisonedRAG, GPT-4-1106-preview generated a short incorrect target and five supporting passages (27.7 words on average across the three datasets). The reproduction retains canonical LM-targeted and HotFlip retrieval prefixes; mechanism and correction runs remove the prefix, merge stored target-supporting passages with clean candidates, and rerank them. All conditions keep the question and short-answer-from-context wrapper fixed, including an "I don't know" instruction when evidence is absent; noisy slots use answer-free tail-pool documents.

Supplementary Table S9 gives model-specific coordinates and source-study denominators; Appendix A details configurations, protocols, interventions, and statistics.

4.2 A RESIDUAL EVIDENCE INTERFACE

Causal localization. Retrieved documents modify a residual evidence band that causally shifts the attacker-versus-correct answer score in all five causal-suite models. With coordinates fixed after 40 discovery questions, replacing all six layers jointly with clean states reduces the held-out margin by 8.44–23.89 nats (Supplementary Table S14). Across 200 held-out splits of the four full-layer scans, band stability is 0.73–0.94 and held-out percentiles are 0.876–0.947. These effects establish the upstream interface independently of how much any selected head mediates.

What the interface carries. A four-model replication ($n = 100$ per model) yields 240 attacked, source-complete units (Table 1). Clean residuals beat zero by 8.23 nats ($p = 1.92 \times 10^{-23}$); neutral and noisy residuals are also stronger, while the poisoned cross-sample mean preserves the attack. Across 272 mask-complete units, document replacement reduces the gap by 7.92 nats versus 2.39 at question positions and 4.11 at the answer prefix. The interface is therefore sensitive both to what the replacement contains and to where it is applied.

Table 1: Residual-source sensitivity ($N = 240$ attacked, source-complete units).

Band source	Δg
Zero	7.86
Poisoned mean-across	0.30
Same-prompt mean-within	11.69
Neutral documents	13.47
Noisy documents	14.85
Clean documents	**16.10**

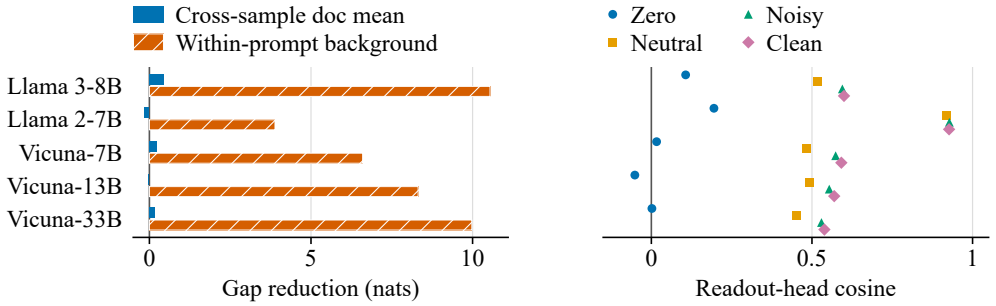
Answer content determines the direction of this source effect. On 40 Llama 3 questions whose clean evidence contains the answer, single-layer patches averaged over L7–12 give a 15.06-nat paired advantage for answer-only evidence over evidence without the answer (Table 2).

The residual band can therefore carry positive, answer-bearing evidence rather than merely encode whether a document is present.

A complementary five-model contrast finds that a cross-sample poisoned-document mean largely preserves the poisoned margin (-0.15 to 0.44 nats reduction), whereas a same-prompt non-context mean reduces it by 3.89–10.54 nats (Figure 2a). This source dependence is consistent across the causal suite.

Table 2: Answer-content control in Llama 3.

Source ($N = 40$)	Δg in nats
Answer-only	26.68
Evidence without answer	11.62
Neutral	9.58
Answer-only advantage	15.06



(a) Source direction **(b) Residual-head readout**

Figure 2: Replacement source and downstream readout across five models. (a) A cross-sample poisoned-document mean largely preserves the attack, whereas the same prompt's non-context background reduces the poisoned-answer gap. (b) Selected-head cosine after residual-band replacement. Nonzero document sources preserve more readout state than zeroing.

On the same 60 held-out Llama 3 questions, distributional and geometric checks place aligned clean/noisy sources within native residual and next-layer variation throughout L7–12 (Appendix C.8).

The band is source-sensitive and can carry an answer-bearing direction. We next test how this source-dependent state reaches the answer score.

4.3 Architecture-Dependent Readout

We now ask how this interface is read out to the answer score. The eight-model screen identifies candidate concentrated heads; the causal result rests on fixed-coordinate held-out interventions and fresh re-localizations.

Table 3: Eight-model readout screen (NQ, $n = 40$).

Model	Head	Poison Δ share	Path-sub. gap
Llama 3-8B	L17-H24	0.95	1.11
Vicuna-33B	L20-H45	0.91	0.76
Qwen2.5-7B	L19-H17	0.94	0.56
Vicuna-7B	L17-H26	0.91	0.49
Vicuna-13B	L18-H18	0.88	0.46
Mistral-7B	L19-H9	0.93	0.26
Llama 2-7B	L15-H30	0.14	0.01
DeepSeek-7B	L0-H13	0.36	0.03

Readout screening. The eight-model NQ screen ($n = 40$) ranks candidate heads by poison-attention contrasts and tests them by poison-path subtraction (Table 3). "Poison-grounding" denotes concentrated poison-token attribution; "causal readout" is reserved for selected heads whose subtraction changes the poisoned-answer margin. Path-subtraction gaps measure single-head effects, separately from residual-band reductions.

In six models, one head attributes 0.88–0.95 of its positive poisoned-versus-noisy output shift to poison tokens (at most 0.05 to clean), and subtracting that path reduces the poisoned-answer gap. Qwen2.5-7B and Mistral-7B extend this concentration beyond the Llama lineage. A/B held-out cross-fitting recovers 6–8 of the same top eight heads across folds in all six strong-screen models.

Table 4: Held-out post-discovery nested mediation on baseline-attacked cases.

Model	N	Mediated %	Sel.–ctrl	Sign p	Rank
Llama 3-8B	39	3.96	0.939	0.0095	1/9
Llama 2-7B	43	0.043	0.009	0.542	1/9
Vicuna-7B	42	2.14	0.328	0.088	1/9
Vicuna-13B	44	3.65	0.622	2.5×10^{-5}	1/9
Vicuna-33B	41	2.51	0.424	4.9×10^{-6}	1/9

DeepSeek-7B instead has a clean-dominant top head (0.50 versus 0.36 poison share), while Llama 2-7B assigns only 0.14 to poison tokens; neither exhibits concentrated one-head grounding.

Held-out causal validation. In the five-model suite, model-specific bands causally shift the answer score, and poison-to-clean content sensitivity ranks above all 16 controls in four models and 15/16 in Llama 2 (Figure 2b). This sensitivity identifies content-reading candidates, and the nested intervention then measures how much of the band effect each selected head mediates.

The held-out nested intervention fixes all coordinates and eight controls before the 60-question test (Table 4). Sel.–ctrl reports selected-head rescue minus the mean of eight fixed same-layer controls, in nats. The selected head ranks first among nine in every model. Four models show 2.14–3.96% mediation, whereas Llama 2 is near zero. The selected heads thus mediate part of the band's effect, with contributions that differ across models.

Re-localized pipeline transfer. For external validity, we rebuild Llama 3/HotpotQA with passage top-50 retrieval, 128-token chunks with 32-token overlap, a Contriever top-30 shortlist, and a reranker trained on 2,000 train and 500 disjoint validation queries; all 7,405 test queries are excluded (AP/AUROC 0.982/0.987). Poison remains in every final top five (4.8 chunks on average), yielding 0.82 ASR versus 0.05/0.06 for clean/noisy evidence.

Fresh localization on 40 discovery questions selects L17-H24 and L1–6. Among 53 attacked held-out cases, the band effect is 24.735 nats, one-head mediation is 3.00%, and selected-head rescue exceeds controls by 0.728 nats ($p = 0.0127$; rank 1/9). Positive mediation survives the pipeline while the band shifts from NQ's L7–12 to L1–6, supporting topology-level transfer.

Architectural variation and supporting diagnostics. Independent Qwen2.5-7B re-localizations find high-effect bands on NQ (L14–19, 10.60 nats) and HotpotQA (L11–16, 13.94 nats), with L19-H17 ranked first on both. The band-to-answer-score effect transfers, while mediation by the selected head remains unresolved, consistent with more distributed readout. At $N = 100$, Qwen has identical top-eight sets across folds and all 16 held-out effects are positive (poison share ≥ 0.836); Mistral retains $7/8$ overlap, while DeepSeek shows only $1/8$ overlap with mixed, often clean-dominant effects.

In a Llama 3 diagnostic, the best top-19 head prefix recovers 3.27 nats (59.7% of the reference readout patch; additivity ratio 1.41), indicating a broader non-additive head family. At answer token eight, a poison passage remains the most-attended document in 0.73–0.77 of prompts for the four models with concentrated partial readout versus 0.28 for Llama 2. Zeroing the selected head across 1, 4, 8, or all positions nevertheless removes at most 0.098 and recovers at most 0.049, confirming persistent readout without sufficient one-head correction.

Together, these interventions establish a re-localizable band-to-answer route with architecture-dependent selected-head mediation. Complete screening, cross-fit, control-head, nested-mediation, and re-localized replication results are reported in Supplementary Tables S2–S24.

4.4 Suppression Is Not Correction

Human-adjudicated fair comparison. Two annotators independently label 237 eligibility items and 320 responses, with disagreements separately adjudicated. Agreement is 0.911 ($\kappa = 0.831$) for evidence support, 0.983 ($\kappa = 0.966$) for clean-response correctness, and 0.947 ($\kappa = 0.929$)

for response class. Requiring supported evidence and a correct clean response yields the same 106 attacks for every operator.

Table 5: Operator comparison on 106 human-adjudicated attacks.

Operator	Removal (%)	Recovery (%)
Oracle all-layer restore	99.1	72.6
Oracle six-layer restore	99.1	**71.7**
No-context fallback	94.3	57.5
Oracle steering $\alpha = 3$	83.0	18.9
Oracle steering $\alpha = 4$	92.5	13.2
Multi-vector 2:1	70.8	22.6
Multi-vector 2:2	73.6	25.5

Using reference-answer matching on the common 106 events, six-layer oracle restoration removes 99.1% of attacks and recovers 71.7%, close to all-layer restoration's 72.6% (Table 5). Its paired recovery is higher by 58.49 percentage points than oracle steering at $\alpha = 4$ and by 14.15 points than no-context fallback, and lower by 0.94 points than all-layer restoration. Table S42 gives full paired counts over the 42 questions. The annotation protocol, agreement, per-model results, and paired clean-rerun analysis appear in Supplementary Tables S36–S45.

All listed interventions except no-context fallback use paired clean evidence. Steering broadcasts a pooled clean-minus-poison direction, reducing attacker support without explicitly reconstructing the document-specific residual state of a clean run. Oracle restoration instead reinstates residual states aligned to retrieved-document positions. Their different recovery rates show that the form in which clean evidence enters the model matters for correction, beyond access to that evidence alone.

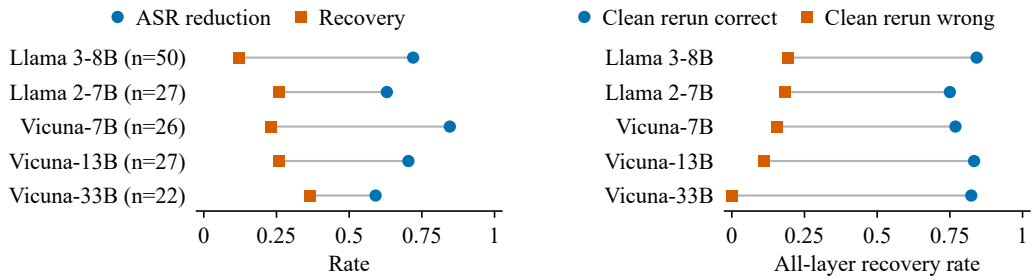


Figure 3: Suppression and recovery. (a) Oracle steering ($\alpha = 2$) on baseline-attacked cases. (b) All-layer oracle restoration by paired clean-rerun correctness.

Across the five models (50, 27, 26, 27, and 22 attacked cases, respectively), local oracle steering at $\alpha = 2$ reduces ASR by 0.59–0.85 but recovers only 0.12–0.36 of attacked cases ($p \leq 0.00845$ in every model; Figure 3a); stronger steering does not close the gap. The all-layer oracle removes 0.91–1.0 but restores only 0.44–0.64.

On 152 matched attacked model–question events, all-layer restoration recovers 75–84% when the clean-context rerun is correct, versus 0–19% otherwise (Figure 3b). Table S45 gives paired counts and patch–rerun tests ($p \geq 0.25$).

On 35 automatically eligible single-poison events, original-band restoration suppresses 34 attacks and matches the reference answer on 12. Adjacent-band restoration and oracle steering each also match 12, versus 17 for no-context generation (Appendix D.2, including eligibility and paired comparisons).

Semantic answer verification. A machine audit of all 742 stored responses keeps the 106 human-adjudicated eligibility events fixed and yields 75.47% recovery for six-layer restoration, 80.19% for all-layer restoration, and 13.21% for oracle steering at $\alpha = 4$ (Table S43). This endpoint accepts

aliases and semantic equivalents while excluding negated mentions and unresolved conflicting answers. The restoration–steering separation therefore also appears under semantic answer adoption.

4.5 Document Attribution from Internal Readout

Table 6: Matched document attribution on 143 attacked model–question events.

Method	Top-1 (%)	Top-2 recall (%)
Retrieval rank	85.31	90.91
RAGForensics (format repair)	30.05	56.60
RAGOrigin	**97.90**	**100.00**
Fixed-head mean attention	96.50	**100.00**

Table 6 compares fixed-head mean attention with RAGForensics and RAGOrigin on 143 attacked model–question events (88 questions) from NQ and HotpotQA, using Llama 3-8B and Vicuna-13B. The comparison fixes the observed answer and five consumed passages, including one poison passage. Both baselines use local backends and expand their native scopes within a shared top-100 cache; scores are then evaluated on the consumed five (Appendix E.1). RAGForensics uses a fixed format-repair rule and expected ranks under binary ties.

Mean attention reaches 96.50% Top-1 localization and 100% Top-2 recall, close to RAGOrigin's observed rates, with one additional target-model prefill given white-box access and a fixed head. It improves Top-1 over retrieval rank by 11.19 percentage points. In this local adaptation, RAGForensics usually flags the poison (99.30% set recall), but often flags clean passages as well: it selects 3.81 of five passages on average. The earlier localization study reaches AUC 0.963–1.000 in three models, with 8.5–12.8% pooled within-prompt extraction overhead (Tables S49 and S57; Figure S1).

5 Discussion

After re-localization in chunked/reranked HotpotQA and Qwen, the layer coordinates change, but the band-to-answer effect remains. Source and position controls show that these residual states carry answer-bearing document content.

With paired oracle evidence, the intervention's form determines how effectively that content supports correction. Steering broadcasts a pooled direction; restoration reinstates states at corresponding document positions. Their comparison separates removing attacker support from supplying positive support for the correct answer. The six-layer band provides a compact intervention site, while clean-context competence helps identify when that evidence can support recovery.

A fixed head also ranks consumed passages in one additional prefill, providing a direct way to prioritize documents for inspection. Future correction operators can retrieve trustworthy evidence and restore its residual states at the corresponding document positions.

6 Conclusion

RAG corpus poisoning acts through a residual evidence interface with architecture-dependent downstream readout. Source and position controls establish its answer-bearing role, and held-out interventions link that state to the answer score. On the common answerable cohort, position-aligned restoration within six layers approaches all-layer recovery and substantially exceeds pooled steering. The recovered readout also supports document inspection. These results distinguish suppression from correction: recovery requires positive support for the correct answer.

AI USE STATEMENT

Generative AI tools were used for language editing, literature discovery. All identified references were independently reviewed and verified by the authors. The authors made all decisions regarding the research methods, experiments, analyses, interpretations, and manuscript content, and take full responsibility for the final work.

ETHICS STATEMENT

This study analyzes corpus poisoning to understand failures of evidence use and to inform RAG evaluation and repair. Experiments use public question-answering datasets and controlled model and retrieval pipelines. The analysis distinguishes attack removal from correct-answer recovery so that a suppressed attack is not presented as a successful answer. Released materials should retain the original data and model licenses and document the dual-use nature of the inherited attack artifacts.

REPRODUCIBILITY STATEMENT

Appendix A specifies model versions, prompts, retrieval, activation alignment, seeds, and outcome definitions. Appendix C gives discovery splits, held-out controls, and complete causal replications. Appendix D gives human eligibility criteria and paired operator denominators. Code is available to reviewers during double-blind review. It will be released publicly with the camera-ready version.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of ICLR*, 2024.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of COLING*, 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 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*, 2020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

Xun Liang, Mengwei Wang, Yuefeng Ma, and Simin Niu. Safe RAG by RAG: Untying the bell that RAG rang with the RAG hand. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *Proceedings of ICLR*, 2023.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine reading COmprehension dataset. In *Proceedings of the NeurIPS Workshop on Cognitive Computation*, 2016.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL*, 2024.

Yangguang Shao, Xinjie Lin, Haozheng Luo, Chengshang Hou, Gang Xiong, Jiahao Yu, and Junzheng Shi. POISONCRAFT: Practical poisoning of retrieval-augmented generation for large language models. arXiv preprint arXiv:2505.06579, 2025.

Yize Sui, Yan Xu, Kun Hu, Jing Ren, and Wenjing Yang. ECD: Evidence-guided contrastive decoding in retrieval-augmented generation with accurate knowledge reference adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *Proceedings of ICLR*, 2025.

Xue Tan, Hao Luan, Mingyu Luo, Xiaoyan Sun, Ping Chen, and Jun Dai. RevPRAG: Revealing poisoning attacks in retrieval-augmented generation through LLM activation analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. arXiv preprint arXiv:2308.10248, 2023.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of ACL*, 2025.

Haowei Wang, Rupeng Zhang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, and Qing Wang. Joint-GCG: Unified gradient-based poisoning attacks on retrieval-augmented generation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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 *Proceedings of ICLR*, 2023.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9b77f07301b1ef1fe810aae96c12cb7b-Abstract-Conference.html.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. arXiv preprint arXiv:2405.15556, 2024.

Peiru Yang, Haoran Zheng, Yi Luo, Xinyi Liu, Jinrui Wang, Huili Wang, Xintian Li, Yongfeng Huang, and Tao Qi. ShieldRAG: Safeguarding retrieval-augmented generation from untrusted knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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 EMNLP*, 2018.

Baolei Zhang, Haoran Xin, Minghong Fang, Zhuqing Liu, Biao Yi, Tong Li, and Zheli Liu. Traceback of poisoning attacks to retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pp. 2085–2097. ACM, 2025. doi: 10.1145/3696410.3714756. URL https://doi.org/10.1145/3696410.3714756.

Baolei Zhang, Haoran Xin, Yuxi Chen, Zhuqing Liu, Biao Yi, Tong Li, Lihai Nie, Zheli Liu, and Minghong Fang. Who taught the lie? responsibility attribution for poisoned knowledge in retrieval-augmented generation. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 4203–4222. IEEE, 2026. doi: 10.1109/SP63933.2026.00053. URL https://doi.org/10.1109/SP63933.2026.00053.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. TrustRAG: Enhancing robustness and trustworthiness in retrieval-augmented generation. arXiv preprint arXiv:2501.00879, 2025.

Youchao Zhou, Heyan Huang, Yicheng Liu, Rui Dai, Xinglin Wang, Xingchen Zhang, Shumin Shi, and Yang Deng. Do retrieval augmented language models know when they don't know? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *Proceedings of the 34th USENIX Security Symposium*, pp. 3827–3844, 2025.

SUPPLEMENTARY MATERIAL

ADDITIONAL FIGURES

This section provides additional figures supporting the main-text analyses.

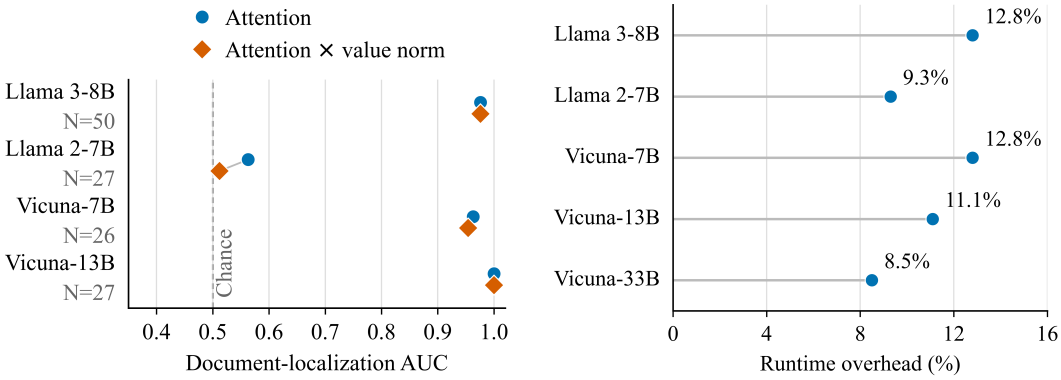


Figure S1: Route-based document attribution. Left: attention and attention × value-norm document-localization AUCs on baseline-attacked cases; document ranking is accurate in three concentrated-readout models, while Llama 2's selected-head scores are near chance. Right: pooled within-prompt overhead for one extraction pass across 1,000 prompts.

This supplement provides the reproducibility protocol and detailed evidence for the paper's causal account of RAG corpus poisoning. It first documents attack reproduction, route discovery, held-out nested mediation, cross-model and retrieval-pipeline replications, activation-realism controls, and source decomposition. It then gives the full suppression–correction results, including multi-operator interventions and the human-adjudicated $N = 106$ common-denominator comparison. Finally, it evaluates route-based document attribution, including document localization and deployment cost, together with additional evaluations of decodability, detection, defense tradeoffs, retrieval stress, adaptive dilution, and calibration.

Unless stated otherwise, experiments use seed 12.

A REPRODUCIBILITY PROTOCOL

A.1 MODELS AND DATA

The five-model causal suite uses Meta-Llama-3-8B-Instruct, Llama-2-7B-Chat, Vicuna-7B-v1.3, Vicuna-13B-v1.3, and Vicuna-33B. The wider screen additionally uses Qwen2.5-7B, Mistral-7B, and DeepSeek-7B. NQ supplies single-hop factoid questions; HotpotQA and 2WikiMultiHopQA stress multi-hop evidence aggregation; MS MARCO supplies web-retrieval questions with a different answer and passage distribution. All are existing, public datasets. Poison targets and LM-targeted passages are inherited from the public PoisonedRAG artifacts rather than resampled by us. In that protocol, GPT-4-1106-preview jointly generated a short incorrect target and five target-supporting passages; the passages average 27.7 words over NQ, HotpotQA, and MS MARCO.

All causal attention runs use eager attention so that document-level weights and head outputs can be captured. Decoding uses temperature 0, no sampling, and at most 150 new tokens. Seed 12 controls example selection, discovery splits, calibration, and bootstrap resampling.

A.2 Retrieval, Prompting, and Poison Construction

We use Contriever with dot-product scoring, retrieve the top five passages, and inject five poison passages per target question unless an ablation states otherwise. Clean and injected candidates are merged and reranked by the same retriever. The canonical instruction template is:

> You are a helpful assistant, below is a query from a user and some relevant contexts. Answer the question given the information in those contexts. Your answer should be short and concise. If you cannot find the answer to the question, just say "I don't know."
>
> `Contexts: [context]`
> `Query: [question]`
> `Answer:`

Each model's native chat template is applied after this wrapper. In the mechanism and correction experiments, targeted attack success requires an attacker-answer match and no canonical-reference match in the generated response. Operator recovery counts the reference answer appearing in the response, evaluated separately from the attack-success flag; a response containing both strings retains its reference match.

The 3,000-trial attack reproduction uses the canonical LM-targeted protocol, which prepends the question to each precomputed PoisonedRAG adversarial passage. The mechanism, correction, and five-model routing/last-token detection suites instead use the stored target-supporting passages without that textual prefix, merge them with clean candidates, and rerank them; this isolates evidence use from prefix optimization. The adaptive dilution experiment uses a separately constructed prefix-matched baseline, described below. HotFlip starts from the question plus target-supporting text and performs 30 retrieval-embedding updates. At each update, one prefix token is selected and the best of 100 candidate tokens is retained; the adversarial prefix is 30 tokens. HotFlip therefore optimizes retrievability, not the internal route. The noisy condition substitutes answer-free documents from a 30-document tail pool for poison slots.

The retrieval-depth/poison-density factorial uses the same 100 seed-12 NQ questions in every cell, varies top-$k \in \{5, 10\}$ and injected passages $m \in \{1, 5\}$, and decodes at most 32 new tokens. The model-specific head is fixed at the coordinate discovered under the canonical setting, and the score is always the maximum, across documents, of mean per-token attention (`attention_mean`). The $k = 5, m = 5$ cell reuses the canonical NQ routing evaluation; the other 15 cells use the identical prompt, retriever, and scoring implementation. The factorial tests robustness of fixed route coordinates under changes in retrieval depth and poison density rather than re-localizing the mechanism.

A.3 Positions, Discovery, and Interventions

Poison positions are all tokens in retrieved passages labeled as poisoned by the experiment metadata (`source=poison`). Residual substitution replaces complete decoder-layer outputs at aligned document positions. Cross-prompt clean substitution pairs the same question under poisoned and clean top-5 contexts by context slot and relative token position; alignments below 0.90 are recorded as unavailable. Evidence bands are contiguous six-layer windows. Head and band discovery use disjoint sample halves, and final metrics use fixed coordinates or held-out rows.

Controlled band sources are zero, length-matched neutral, noisy, clean, mean-across, and mean-within. Mean-across averages poisoned-document residuals over samples at the same layer, context slot, and relative token position. Mean-within averages the evaluated prompt's non-context residuals at a layer and broadcasts that vector over all replaced document positions. Neither is an across-layer average. The local clean direction is constructed per sample and layer as the pooled clean-context document residual minus its poisoned-context counterpart. We add this raw, non-unit-normalized vector at document positions. It therefore uses paired clean evidence and is an oracle source; the correct-answer label is used for evaluation, not to fit a token-level answer vector. Leave-one-out and learned probe directions exclude the evaluated sample.

The correction experiment uses band-only and all-layer oracle clean-state restoration, two multi-vector mixtures that combine the same-example oracle direction with a leave-one-out learned direction using coefficient pairs $(2, 1)$ and $(2, 2)$, strong local steering with $\alpha \in \{3, 4\}$, no-context parametric fallback, and a clean-context rerun. Clean-state restoration, local steering, multi-vector

mixtures, and the clean-context rerun use paired clean evidence. They are oracle interventions; no-context fallback uses parametric knowledge.

The held-out nested experiment excludes the 40 questions used for coordinate discovery and evaluates the remaining 60 questions. For every baseline-attacked case, paired clean residuals first replace the fixed six-layer band. We then restore the original poisoned-run last-token output of either the selected head or one of eight seed-12 same-layer controls. Total effect is the baseline poisoned gap minus the clean-band-patched gap; rescue is the restored-head gap minus the clean-band-patched gap; the mediated fraction is mean rescue divided by mean total effect. All nine heads are evaluated on the same attacked sample IDs. Selected-minus-control rescue additionally uses an exact paired sign test.

For the human fair comparison, two independent human annotators, identified as Annotator A and Annotator B, receive separately blinded annotation forms. They label clean-evidence support, clean-response semantic correctness, contradiction, response semantic class, and explicit answer negations. The forms omit model, operator, automatic score, and sample identifiers. Labels are completed independently; disagreements are resolved by adjudication only after both annotations are fixed. The strict fair subset requires both clean evidence that supports the correct answer and a semantically correct clean response. Human labels determine eligibility only; operator removal and recovery retain the fixed automatic string-matching rules.

A.4 Calibration, Metrics, and Statistical Tests

Abstention thresholds are selected from clean-correct examples at a target clean-coverage cost of 0.25 unless stated otherwise. The canonical last-token evaluation uses a fixed clean calibration; document-route audits use leave-one-out calibration. Neutralization is measured only over baseline-attacked poisoned examples. AUROC and AUPRC use all clean and poisoned rows. We report attack success rate (ASR), poisoned-answer gap reduction, correct-answer recovery, clean-coverage cost, neutralization, document poison precision, AUROC/AUPRC, FPR at TPR 0.9, and selective risk–coverage. Paired changes in the poisoned-answer margin use sign tests. All reported deterministic model settings use one generation run per example.

Complete-answer scoring implementation. For each candidate answer, the scorer teacher-forces the complete continuation and sums its token log probabilities without length normalization, retaining the higher score of the original and leading-space forms. Head interventions stay at the final position of the original prompt throughout this scoring pass; residual interventions retain their fixed document-token masks. In 360 stored records, including 328 with multi-token candidates, the saved gap equals the attacker-answer sum minus the reference-answer sum. For example, the four-token reference *Ashrita Furman* has $S(y_{\mathrm{cor}}) = -21.984375$, while the four-token attacker answer *Evelyn Smith* has $S(y_{\mathrm{inc}}) = -0.548340$. Their gap is $-0.548340 - (-21.984375) = 21.436035$ nats, matching the stored nested-intervention baseline. Generated-response outcomes are evaluated separately by the matching rules above.

Code availability. Anonymous source code, model configurations, and dependency specifications accompany the submission. Model weights and public datasets remain under their original licenses and are not redistributed.

B Attack Reproduction and Model Coordinates

B.1 Head Stability at $N = 100$

We separately expand Qwen2.5-7B, Mistral-7B, and DeepSeek-7B to the same ordered 100-question, seed-12 NQ cohort. This is an expanded replication of the screen in Table S2: the eight-model screen establishes breadth, while these three fully expanded models provide higher-powered causal and split-stability evidence. The retrieval and prompting protocol remains Contriever dot-product retrieval, five retrieved contexts, five injected passages, the canonical instruction template above, greedy decoding, and a 12-token response cap. For each clean, noisy, and poisoned prompt, we capture last-query attention for every head. We rank heads using poison mass, poison-minus-clean mass, incorrect-answer-anchor mass, poisoned-minus-noisy incorrect-anchor mass, and their

Table S1: Open-weight PoisonedRAG reproduction. Each row pools the five causal-suite models (100 trials/model); pooled ASR is 0.936 over 3,000 trials.

Dataset	Attack	Trials	Successes	ASR
HotpotQA	LM-targeted	500	483	0.966
HotpotQA	HotFlip	500	480	0.960
MS MARCO	LM-targeted	500	457	0.914
MS MARCO	HotFlip	500	445	0.890
NQ	LM-targeted	500	464	0.928
NQ	HotFlip	500	479	0.958

Table S2: Eight-model NQ screen ($n = 40$). Six models exhibit a concentrated poison-grounding head; Llama 2 and DeepSeek delimit the single-head concentration claim.

Model	Head	Poison share	Head gap
Llama 3-8B	L17-H24	0.95	1.11
Vicuna-33B	L20-H45	0.91	0.76
Qwen2.5-7B	L19-H17	0.94	0.56
Vicuna-7B	L17-H26	0.91	0.49
Vicuna-13B	L18-H18	0.88	0.46
Mistral-7B	L19-H9	0.93	0.26
Llama 2-7B	L15-H30	0.14	0.01
DeepSeek-7B	L0-H13	0.36	0.03

equal-rank consensus, each on all examples, baseline-attacked examples, and poison-flip examples. Candidate stability uses 2,000 bootstrap resamples containing 80% of the relevant rows. The 32 consensus candidates are then causally evaluated on baseline-attacked examples by (i) zeroing the last-token head output and (ii) subtracting the poison-token contribution to that output. An equal-rank consensus over mean subtraction effect, mean zeroing effect, and positive-effect rate selects four descriptive full-cohort heads. Source shares decompose the positive poisoned-minus-noisy head-output change. This design combines a broad architecture survey with higher-powered causal tests at the three selected expansion points.

The leakage-controlled result is a bidirectional cross-fit. Sorted sample ids are alternated into disjoint 50-example halves A and B. Fold AB discovers on A, fixes its first eight consensus heads in discovery order, and evaluates those heads only on baseline-attacked rows in B; fold BA reverses the halves. The held-out causal measurements never reorder the selected heads. Thus the full-cohort table is a higher-powered descriptive estimate, whereas the two fixed folds provide the held-out coordinate-stability test.

Table S3: Four-head outputs of the three $N = 100$ full-cohort screens. Sub. and Zero are mean reductions in the poisoned wrong-minus-correct log-probability gap; Sub. $P+$ is the positive subtraction-effect rate. Qwen and Mistral retain concentrated poison routing, whereas DeepSeek's smaller effects and clean-dominant source shares provide an architecture-specific contrast that sharpens the scope of the concentrated-head result.

Model	Head	Att. N	Sub. Δg	Zero Δg	Sub. $P+$	Poison share	Clean share
Qwen2.5-7B	L19-H17	76	1.140	1.134	0.840	0.919	0.079
Qwen2.5-7B	L22-H24	76	0.927	0.884	0.747	0.931	0.051
Qwen2.5-7B	L23-H11	76	0.882	0.819	0.720	0.891	0.043
Qwen2.5-7B	L19-H22	76	0.864	0.855	0.720	0.905	0.030
Mistral-7B	L19-H9	69	0.271	0.249	0.812	0.940	0.022
Mistral-7B	L19-H16	69	0.248	0.202	0.768	0.773	0.020
Mistral-7B	L18-H30	69	0.177	0.164	0.710	0.711	0.038
Mistral-7B	L31-H6	69	0.106	0.106	0.580	0.566	0.085
DeepSeek-7B	L22-H11	27	0.029	0.074	0.708	0.294	0.412
DeepSeek-7B	L16-H11	27	0.089	0.154	0.500	0.465	0.451
DeepSeek-7B	L19-H24	27	0.061	0.101	0.500	0.424	0.405
DeepSeek-7B	L24-H16	27	0.037	0.080	0.583	0.321	0.402

Table S4: All fixed Qwen2.5-7B heads in both 50/50 cross-fit directions. Rank is selected on the discovery half. The top-eight sets overlap $8/8$; every held-out head has positive mean subtraction and zeroing effects, and poison share is at least 0.836.

Fold	Rank	Head	Att. N	Sub. Δg	Zero Δg	Sub. $P+$	Poison share	Clean share
AB	1	L22-H24	35	1.328	1.232	0.765	0.930	0.072
AB	2	L23-H11	35	1.343	1.196	0.794	0.900	0.063
AB	3	L19-H17	35	1.595	1.569	0.853	0.912	0.091
AB	4	L21-H14	35	1.071	1.051	0.676	0.865	0.070
AB	5	L22-H25	35	1.031	0.956	0.588	0.837	0.049
AB	6	L22-H1	35	1.185	1.110	0.765	0.928	0.078
AB	7	L25-H25	35	1.091	1.035	0.676	0.865	0.075
AB	8	L24-H21	35	1.276	1.158	0.765	0.908	0.101
BA	1	L19-H17	41	0.759	0.719	0.829	0.924	0.069
BA	2	L23-H11	41	0.465	0.446	0.683	0.883	0.027
BA	3	L22-H24	41	0.552	0.513	0.732	0.932	0.034
BA	4	L22-H1	41	0.317	0.285	0.683	0.940	0.049
BA	5	L22-H25	41	0.161	0.189	0.561	0.886	0.033
BA	6	L25-H25	41	0.210	0.195	0.585	0.836	0.073
BA	7	L21-H14	41	0.154	0.264	0.561	0.891	0.047
BA	8	L24-H21	41	0.392	0.350	0.707	0.919	0.073

Table S5: All fixed Mistral-7B heads in both cross-fit directions.

Fold	Rank	Head	Att. N	Sub. Δg	Zero Δg	Sub. $P+$	Poison share	Clean share
AB	1	L19-H9	31	0.386	0.376	0.806	0.954	0.008
AB	2	L19-H16	31	0.176	0.119	0.710	0.754	0.017
AB	3	L19-H17	31	0.003	0.039	0.355	0.758	0.041
AB	4	L28-H0	31	-0.060	-0.051	0.355	0.756	0.064
AB	5	L16-H29	31	-0.019	-0.012	0.516	0.774	0.048
AB	6	L18-H30	31	0.149	0.176	0.677	0.733	0.046
AB	7	L25-H29	31	0.080	0.034	0.419	0.689	0.096
AB	8	L17-H26	31	0.038	0.045	0.613	0.786	0.039
BA	1	L28-H0	38	-0.096	-0.080	0.342	0.683	0.068
BA	2	L19-H9	38	0.177	0.146	0.816	0.929	0.033
BA	3	L18-H22	38	-0.002	0.026	0.474	0.595	0.020
BA	4	L19-H16	38	0.306	0.269	0.816	0.789	0.023
BA	5	L16-H29	38	0.034	-0.052	0.526	0.694	0.073
BA	6	L25-H29	38	0.036	0.002	0.526	0.625	0.081
BA	7	L18-H30	38	0.199	0.153	0.737	0.693	0.032
BA	8	L19-H17	38	-0.175	-0.177	0.500	0.829	0.028

Table S5 shows a strong model-level replication: the head sets overlap $7/8$, AB's L19-H9 has subtraction effect $+0.386$ and poison share 0.954, and BA retains L19-H9/L19-H16/L18-H30 effects of $+0.177/+0.306/+0.199$. The narrower qualification is top-1 ordering: BA's discovery-rank-1 L28-H0 transfers at -0.096. Thus the set is highly stable while the winner among several related heads has one-direction sampling variance.

Table S6: All fixed DeepSeek-7B heads in both cross-fit directions. One coordinate overlaps between top-eight sets ($1/8$), poison share is often below clean share, and signed subtraction effects vary across AB. DeepSeek therefore provides an architecture contrast for the conditions under which single-head concentration emerges.

Fold	Rank	Head	Att. N	Sub. Δg	Zero Δg	Sub. $P+$	Poison share	Clean share
AB	1	L20-H15	13	-0.041	-0.051	0.545	0.338	0.603
AB	2	L16-H16	13	0.037	0.111	0.545	0.380	0.607
AB	3	L17-H22	13	-0.040	-0.057	0.545	0.308	0.676
AB	4	L13-H18	13	-0.057	-0.012	0.545	0.395	0.811
AB	5	L15-H3	13	-0.017	-0.049	0.455	0.314	0.670
AB	6	L16-H11	13	-0.173	-0.123	0.364	0.405	0.569
AB	7	L18-H24	13	-0.037	-0.107	0.273	0.373	0.617
AB	8	L26-H11	13	0.009	-0.010	0.636	0.301	0.558
BA	1	L16-H16	14	0.038	0.001	0.308	0.527	0.370
BA	2	L19-H24	14	-0.049	-0.040	0.231	0.453	0.326
BA	3	L18-H8	14	0.021	0.041	0.462	0.363	0.542
BA	4	L22-H5	14	0.033	0.042	0.538	0.254	0.272
BA	5	L24-H3	14	0.008	0.008	0.538	0.365	0.303
BA	6	L16-H31	14	0.065	0.102	0.615	0.633	0.343
BA	7	L24-H16	14	-0.004	0.023	0.462	0.301	0.390
BA	8	L26-H20	14	0.031	0.147	0.385	0.345	0.261

Table S7: Held-out source-masking controls for each discovery-selected rank-1 head. Each intervention subtracts the top eight positive-contribution tokens from the named source.

Model	Fold/head	Poison top-doc	Anchor overlap	Poison-mask Δg	Backgr.-mask Δg	Clean-mask Δg	Elig. N
Qwen2.5-7B	AB/L22-H24	0.943	0.971	1.356	1.173	1.288	34
Qwen2.5-7B	BA/L19-H17	0.976	0.927	0.593	0.301	0.524	41
Mistral-7B	AB/L19-H9	1.000	1.000	0.373	0.000	0.002	31
Mistral-7B	BA/L28-H0	0.895	0.868	-0.091	0.014	-0.002	38
DeepSeek-7B	AB/L20-H15	0.385	0.462	-0.044	-0.001	0.001	11
DeepSeek-7B	BA/L16-H16	0.786	0.857	0.038	-0.056	0.027	13

Table S7 shows sharp poison specificity in Mistral AB. Mistral BA's -0.091 row pertains only to the fold's unstable top-1 ordering; the same fold contains three positive fixed heads in Table S5. Qwen's poison mask is largest in both folds, with broader source sensitivity also visible; this is why source masking complements, rather than replaces, the primary full-head causal endpoint. The displayed Mistral-AB background-mask value 0.000 is -0.000295 before three-decimal rounding; all comparisons and statistics use the unrounded values.

Table S8: Full-cohort discovery stability over 2,000 80%-size bootstrap resamples. Jaccard is for the 32-head candidate set; frequencies correspond to the preferred-subset rank-average top four. Together with cross-fold overlap of $8/8$, $7/8$, and $1/8$, respectively, the table identifies stable Qwen/Mistral localization and a distinct DeepSeek routing regime.

Model	Attacked	Flip	Preferred subset	Rank-avg. Jaccard	All Jaccard range	Top-4 frequencies
Qwen2.5-7B	76	72	Poison flip	0.787	0.786–0.836	1.00/1.00/1.00/0.55
Mistral-7B	69	68	Poison flip	0.849	0.733–0.849	1.00/0.97/0.86/0.26
DeepSeek-7B	27	25	Poison flip	0.375	0.323–0.789	0.60/0.43/0.29/0.11

At $N = 100$, Qwen shows a large and stable localized route: all 16 held-out head evaluations are positive, and the two discovery halves select the same eight coordinates. Mistral retains a concentrated poison-grounding family with $7/8$ coordinate overlap. Although the top-ranked BA head changes on held-out data, three other BA heads and the AB winner retain positive effects; the head family is stable even though its top-1 ordering is not. DeepSeek has a smaller attacked denominator, mixed source decomposition, and low coordinate overlap, indicating a different routing regime.

Table S9: Selected coordinates and source-study denominators. Held-out mediation uses the baseline-attacked subsets of 60 post-discovery questions, with effect estimates reported separately in Tables S14 and S15.

Model	Selected head	Band	Route N	Source N
Llama 3-8B	L17-H24	L7–12	68	80
Llama 2-7B	L15-H30	L5–10	35	40
Vicuna-7B	L17-H26	L7–12	35	40
Vicuna-13B	L18-H18	L5–10	35	40
Vicuna-33B	L20-H45	L0–5	35	40

Table S10: Full-layer zero-source sweep used to localize residual evidence bands. The Llama 3 sweep contains 15,360 rows over layers, residual/MLP components, and masks; all top-ten rows are residual rather than MLP interventions. Values are poisoned-gap reductions in nats.

Model	Layers	N	Top residual	Δg
Llama 3-8B	32	80	L8	12.03
Llama 2-7B	32	40	L10	1.94
Vicuna-7B	32	40	L7	3.94
Vicuna-13B	40	40	L4	3.34
Vicuna-33B	60	40	L9 / L1	3.3 / 3.0

Each model contributes 100 questions divided into disjoint 50/50 folds; the full-cohort analysis evaluates 32 candidate heads, and each cross-fit direction evaluates the eight heads selected on the opposite fold.

C Causal Route Evidence

C.1 Layer and Component Sweeps

The Llama 2 maximum is smaller than the other models, matching its weaker single-head screen. Vicuna-33B has two shallow peaks, and its selected L0–5 band covers the stronger front-network region. This supports coordinate-varying residual-band localization, not a fixed layer coordinate.

C.2 Head and Band Controls

For each model, the fixed selected head is compared with eight same-layer and eight random-layer heads. These comparisons measure evidence sensitivity at candidate readout sites. The held-out nested test below measures their causal mediation. Evidence sensitivity is $1 - \cos(o_{\text{clean}}, o_{\text{poison}})$; zero-collapse depth is $\cos_{\text{clean}} - \cos_{\text{zero}}$.

Table S11: Evidence-sensitivity control. Poison-to-clean output divergence ranks the selected head above every one of 16 controls in four models and above 15 of 16 (94th percentile) in Llama 2. This establishes evidence sensitivity in all five models; causal mediation is evaluated separately below.

Model	Selected	Same-layer	Pctl./16
Llama 3-8B	0.400	0.093	1.00
Llama 2-7B	0.073	0.016	0.94
Vicuna-7B	0.409	0.158	1.00
Vicuna-13B	0.431	0.157	1.00
Vicuna-33B	0.461	0.117	1.00

On 200 seed-12 half-splits, a six-layer band selected on half A is evaluated on half B.

Two additional random-source controls separate full-mask from single-slot replacement. A full-length random source is matched to the entire retrieved-document mask, avoiding a trivial token-count mismatch. A second control restricts replacement to a single retrieved slot with strict length

Table S12: Zero-source sanity check. Raw collapse is comparable or deeper for neighboring heads, so it supports an upstream-state dependency but is not used as the head-specificity metric.

Model	Readout collapse	Same-layer collapse
Llama 3-8B	0.493	0.432
Llama 2-7B	0.732	0.783
Vicuna-7B	0.575	0.842
Vicuna-13B	0.621	0.695
Vicuna-33B	0.537	0.638

Table S13: Band cross-fit. The selected six-layer band is recovered with 0.73–0.94 stability; held-out percentiles are 0.876–0.947. Llama 3 was scanned over a narrower layer subset and is not included in this full-layer control.

Model	Band	In Δg	Out Δg	Held-out pct.
Llama 2-7B	5–10	10.48	4.05	0.947
Vicuna-7B	7–12	15.90	7.46	0.876
Vicuna-13B	5–10	15.46	6.15	0.907
Vicuna-33B	0–5	15.50	6.56	0.940

matching; its near-zero effect shows that one arbitrary slot is not a sufficient carrier, but does not rule out a route distributed over several poisoned documents. These controls are not used to rank heads.

C.3 Held-Out Nested Mediation

The nested experiment fixes the model coordinate and eight same-layer control heads using the first 40 discovery questions, then evaluates only the remaining 60 questions. The analyzed denominator is the baseline-attacked subset of those 60. Restoring a head after clean-band replacement asks a strict causal question: how much of the band intervention's poisoned-margin reduction is rescued by reinstating that one poisoned-run head state? Tables S14 and S15 report every model-level summary.

Table S14: Held-out nested mediation. Effects are poisoned-answer-gap changes in nats; mediated fraction is mean selected-head rescue divided by mean total clean-band effect.

Model	Attacked N	Total effect	Selected rescue	Mediated %
Llama 3-8B	39	23.891	0.946	3.958
Llama 2-7B	43	8.440	0.004	0.043
Vicuna-7B	42	15.172	0.324	2.138
Vicuna-13B	44	16.992	0.620	3.649
Vicuna-33B	41	17.022	0.427	2.508

Table S15: Nested head specificity. Every attacked sample contributes nine rows (selected plus eight fixed same-layer controls). Sign p is the two-sided exact paired sign test against the per-sample mean control.

Model	Rows	Mean control rescue	Sel.–ctrl	Sign p	Rank
Llama 3-8B	351	0.006	0.939	0.00948	1/9
Llama 2-7B	387	-0.005	0.009	0.542	1/9
Vicuna-7B	378	-0.004	0.328	0.0884	1/9
Vicuna-13B	396	-0.002	0.622	2.54×10^{-5}	1/9
Vicuna-33B	369	0.003	0.424	4.87×10^{-6}	1/9

Table S15 shows that the selected head ranks first in all five models. Vicuna-7B does not meet the 0.05 sign-test threshold ($p = 0.0884$).

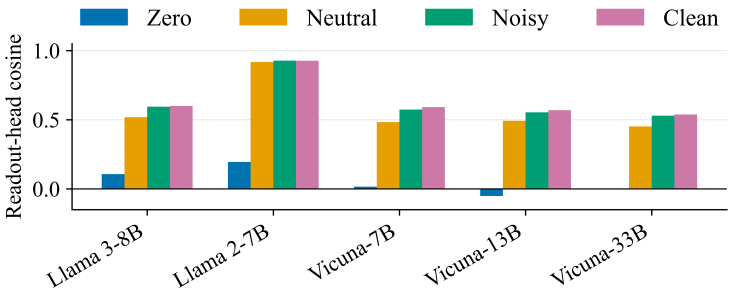


Figure S2: Evidence-band readout sensitivity across the five causal-suite models. Bars show cosine similarity between the selected-head output on the poisoned run and its output after replacing the evidence band with each source. Nonzero document sources preserve substantially more of the readout state than zeroing. The held-out nested test in Appendix C.3 quantifies partial band-to-head-to-answer-score mediation.

The 40 discovery questions and 60 test questions are disjoint, and all nine heads are evaluated on the same attacked test cases. Total effect decomposes into rescued and unrescued effects, so the nested experiment tests the selected coordinate without reusing discovery examples.

C.4 QWEN NQ REPLICATION

We repeat the complete mechanism chain on Qwen2.5-7B rather than transferring a Llama/Vicuna coordinate. The experiment uses 100 NQ questions with `torch.bfloat16` and eager attention. The first 40 questions are used for head and band discovery, and the remaining 60 form the held-out nested-mediation set.

Table S16: Qwen2.5-7B NQ causal-route replication. The architecture-specific six-layer band is strong. These results support an architecture-specific band-to-answer-score route.

Mechanism component	Result
Evaluation cohort	100 questions; poisoned ASR 0.51; clean ASR 0.03
Backend	`bfloat16` + eager attention
Layer/source sweep	28 layers $\times$ 2 sources; $n = 40$ per condition
Selected coordinate	L19-H17; evidence band L14–19
Discovery band effect	10.604 nats
Nested denominator / rows	33 attacked / 297 rows
Total effect	11.794
Selected rescue	0.071
Mediated %	0.606
Sel.–ctrl	0.104
Sign p / selected rank	0.163 / 1 of 9

Table S17: Qwen NQ discovery-set mean clean-patch gap reduction for every layer. The six-layer selector ranks L14–19 first (10.604 mean), ahead of L15–20 (10.581), before the 60-question nested test is evaluated.

Layer	Δg	Layer	Δg	Layer	Δg	Layer	Δg
0	10.065	7	10.106	14	10.344	21	10.566
1	9.996	8	10.209	15	10.769	22	6.062
2	9.888	9	10.276	16	10.721	23	4.163
3	10.225	10	10.338	17	10.688	24	3.153
4	10.127	11	10.262	18	10.652	25	2.705
5	10.010	12	10.193	19	10.452	26	1.705
6	10.047	13	10.127	20	10.202	27	0.014

The selected head ranks first among all nine candidates. Qwen recovers the localized residual evidence band on NQ.

C.5 QWEN HOTPOTQA REPLICATION

We next repeat the entire Qwen chain on HotpotQA, re-localizing rather than transferring either the NQ band or the NQ head. The experiment uses 100 questions under the same `bfloat16`, eager-attention configuration. The first 40 questions are used for head and band discovery; the remaining 60 form the held-out nested-mediation set. Poisoned ASR is 0.89 over all 100 questions, compared with 0.04 under clean evidence and 0.05 under noisy evidence. The discovery and post-discovery partitions contain 31/40 and 58/60 attacked cases, respectively.

The discovery screen ranks L19-H17 first: subtracting its poisoned-token path reduces the poisoned-answer gap by 0.412 nats on average, with a positive reduction on 0.903 of attacked examples. Poison tokens account for 0.875 of its positive output shift, and a poison document is the top positive document on all 31 attacked discovery examples. We then evaluate every decoder layer under paired clean-document replacement and zeroing, using the same 40 discovery questions for each layer and source. The six-layer selector chooses L11–16 with a 13.941-nat mean clean-patch effect on attacked discovery cases.

Table S18: Qwen2.5-7B HotpotQA causal-route replication. A large separately localized evidence-band effect coexists with near-zero one-head rescue on a multi-hop dataset.

Mechanism component	Result
Evaluation cohort	100 questions; clean/noisy/poisoned ASR 0.04/0.05/0.89
Discovery / test	40 / 60 questions; 31 / 58 attacked
Head screen	L19-H17; path subtraction 0.412; poison share 0.875
Layer/source sweep	28 layers $\times$ 2 sources; $n = 40$ per condition
Selected evidence band	L11–16; discovery effect 13.941 nats
Nested denominator / rows	58 attacked / 522 rows
Total effect	15.738
Selected rescue	0.043
Mediated %	0.273
Sel.–ctrl	0.069
Sign p / selected rank	0.896 / 1 of 9

Across NQ and HotpotQA, Qwen exhibits a high-effect residual evidence band, while the downstream effect is not captured by one localized head.

Table S19: Qwen/HotpotQA layer/source sweep. Entries are mean poisoned-gap reductions on the 31 attacked discovery cases for paired clean-document replacement and zeroing. The selector uses only the clean column and ranks L11–16 first; the next windows are L12–17 (13.914) and L10–15 (13.905).

Layer	Clean	Zero	Layer	Clean	Zero
0	11.807	9.528	14	14.400	10.765
1	12.111	6.033	15	14.225	7.177
2	11.957	5.282	16	13.940	7.897
3	12.202	8.218	17	13.645	8.662
4	12.554	9.988	18	13.480	9.157
5	12.494	10.679	19	13.723	8.993
6	12.677	9.613	20	13.375	9.677
7	12.848	9.687	21	12.766	8.327
8	13.415	10.795	22	7.781	4.471
9	13.675	11.948	23	5.563	2.754
10	13.720	11.558	24	4.508	2.355
11	13.811	9.969	25	4.176	2.042
12	13.556	10.671	26	2.218	0.544
13	13.716	6.901	27	0.000	0.000

Table S20: All nine fixed Qwen/HotpotQA nested-restoration controls on the 58 attacked post-discovery cases. H17 ranks first.

Restored L19 head	Rescue	Mediated %
H17 (selected)	0.043	0.273
H18	−0.015	−0.097
H9	−0.032	−0.203
H25	−0.038	−0.242
H19	0.023	0.149
H12	−0.023	−0.144
H4	−0.041	−0.261
H13	−0.013	−0.080
H0	−0.067	−0.424

C.6 Chunked and Reranked HotpotQA

This experiment changes both the dataset and the retrieval pipeline. Contriever first retrieves 50 passages. Each passage is split into 128-token chunks with 32-token overlap; Contriever retains a 30-chunk shortlist, and a learned joint question–chunk reranker selects the final five contexts. The reranker is trained only from HotpotQA train qrels: 2,000 train queries and 500 disjoint validation queries, with two positive and four cross-query negative chunks per query. It encodes each question–chunk pair jointly with Contriever and fits a standardized logistic classifier. All 7,405 HotpotQA test queries are excluded from fitting and validation. This experiment tests route persistence under this learned question–chunk reranking pipeline.

The held-out reranker validation contains 3,000 pairs and reaches 0.950 accuracy, 0.982 average precision, 0.987 AUROC, and 0.188 log loss. We then select 100 seed-12 test IDs. The first 40 are used for head and band discovery; the remaining 60 are untouched until nested mediation. All conditions use the canonical instruction template, five injected passages, top-5 final retrieval, and greedy decoding up to 150 tokens. The pipeline remains intentionally attack-active: poison survives into the final top five on 100/100 questions, with 4.8 poison chunks on average. Clean/noisy/poisoned ASR is 0.05/0.06/0.82.

Table S21: Leakage-controlled chunking and learned-reranking protocol. The reranker is trained without HotpotQA test queries.

Protocol item	Value or result
Reranker fit	2,000 train queries; 12,000 pairs; train qrels only
Reranker validation	500 disjoint queries; 3,000 pairs; AP/AUROC 0.982/0.987
Excluded test split	7,405 queries; no train/validation overlap
Retrieval pipeline	passage top-50 $\rightarrow$ chunk128/overlap32 $\rightarrow$ shortlist30 $\rightarrow$ top-5
Readout cohort	100 IDs; first 40 discovery, final 60 post-discovery
Poison survival	1.00 any-final rate; 4.8 poison chunks in final top-5
Readout ASR	clean 0.05; noisy 0.06; poisoned 0.82

The discovery screen evaluates 32 candidates on 29 attacked discovery questions. L17-H24 ranks first: poisoned-path subtraction reduces the target margin by 0.640 nats, with a positive effect on 0.897 of cases. Poison tokens account for 0.807 of its positive output shift, versus 0.001 for clean tokens; a poison document and sentence rank first on every attacked discovery question. We then evaluate all 32 layers under clean-document replacement and zeroing on the same 40 discovery questions. The six-layer selector chooses L1–6 with a 21.368-nat attacked-case clean-patch effect.

Table S22: Chunked/reranked Llama 3–HotpotQA causal-route replication. Both the residual band and concentrated partial head mediation reproduce.

Mechanism component	Result
Head screen	L17-H24; path subtraction 0.640; poison share 0.807
Layer/source sweep	32 layers $\times$ 2 sources; $n = 40$ per condition
Selected evidence band	L1–6; discovery effect 21.368 nats
Post-discovery denominator	53 attacked of 60; 477 rows across nine heads
Total effect	24.735 nats
Selected rescue	0.742 nats
Mediated %	2.998
Sel.–mean-control	0.728 nats
Sign p / selected rank	0.0127 / 1 of 9

Table S23: Chunked/reranked layer/source sweep. Entries are mean gap reductions on the 29 attacked discovery cases. The selector ranks L1–6 first (21.368), narrowly ahead of L2–7 (21.349) and L0–5 (21.307); the broad early plateau and near-zero final layer support conservation of the broad early-layer topology.

Layer	Clean	Zero	Layer	Clean	Zero
0	21.157	14.277	16	13.189	7.828
1	21.338	13.184	17	10.962	6.394
2	21.016	13.413	18	10.454	7.555
3	21.218	15.523	19	10.266	7.060
4	21.573	13.965	20	9.704	7.609
5	21.538	14.968	21	9.491	8.984
6	21.525	15.441	22	7.995	6.899
7	21.221	19.373	23	5.953	4.717
8	20.620	18.129	24	3.603	2.334
9	20.422	18.003	25	2.488	4.089
10	19.394	16.930	26	1.735	1.787
11	18.674	15.368	27	0.686	2.201
12	18.838	16.853	28	0.822	2.451
13	16.919	15.144	29	0.767	1.804
14	16.465	14.372	30	0.109	1.079
15	14.721	11.635	31	−0.005	−0.005

Table S24: All nine fixed same-layer nested controls on the 53 attacked post-discovery cases. H24 ranks first.

Restored L17 head	Rescue	Mediated %	Sel.–head
H24 (selected)	0.742	2.998	–
H15	0.011	0.046	0.730
H8	0.001	0.003	0.741
H21	0.138	0.560	0.603
H16	−0.001	−0.002	0.742
H29	−0.000	−0.001	0.742
H11	−0.001	−0.002	0.742
H4	−0.035	−0.143	0.777
H12	−0.003	−0.014	0.745

The selected head's rescue exceeds the control mean on 0.679 of paired cases, with exact sign $p = 0.0127$. The coarse rank-randomization test is necessarily $1/9 = 0.111$ with only eight controls; we therefore report it alongside, not instead of, the paired sign test. Finally, the train-only reranker is deliberately lightweight and poison survival remains high. Under this attack-active pipeline, the broad causal-route topology persists through realistic chunking and learned reranking while its evidence-band coordinates shift.

C.7 Nonzero Source Controls

This replication starts from 100 NQ questions per model and fixes the selected six-layer bands. Source-complete attacked denominators are 54, 62, 61, and 63 for Llama 3, Llama 2, Vicuna-7B, and Vicuna-13B ($N = 240$ pooled). Neutral construction is available for 83/86/86/86 complete examples respectively, and the common attacked set is used across all six source types.

Table S25: Mean poisoned-gap reduction by band source in the four-model $N = 100$ replication. Mean-across is a poisoned document-position mean and therefore preserves the attack; clean, neutral, and noisy nonzero sources produce larger reductions than destructive zeroing.

Model	N	Zero	Mean-across	Mean-within	Noisy	Neutral	Clean
Llama 3-8B	54	16.93	0.49	18.21	22.15	19.01	23.58
Llama 2-7B	62	3.18	0.14	5.74	9.22	7.71	9.69
Vicuna-7B	61	6.56	0.25	10.89	14.23	13.58	15.74
Vicuna-13B	63	5.96	0.33	12.75	14.72	14.28	16.34
Pooled	240	7.86	0.30	11.69	14.85	13.47	16.10

Mask-complete attacked denominators are 63/70/68/71 ($N = 272$ pooled). The same band and zero source are applied at document, question, or answer-prefix positions.

Table S26: Paired clean-minus-zero contrasts. Every model and the pooled analysis show that a semantically meaningful nonzero source is more effective than zeroing, directly rejecting a zero-ablation-artifact explanation.

Model	Paired N	Clean−zero	Exact sign p
Llama 3-8B	54	6.648	1.38×10^{-7}
Llama 2-7B	62	6.505	0.00316
Vicuna-7B	61	9.181	7.67×10^{-6}
Vicuna-13B	63	10.379	1.66×10^{-12}
Pooled	240	**8.234**	1.92×10^{-23}

Table S27: Mean poisoned-gap reduction by intervention mask. The pooled effect is largest at retrieved-document positions, the hypothesized evidence-interface site.

Model	N	Document	Question	Answer prefix
Llama 3-8B	63	16.44	3.05	2.88
Llama 2-7B	70	3.17	0.97	1.03
Vicuna-7B	68	6.66	2.66	6.92
Vicuna-13B	71	6.24	2.96	5.56
Pooled	272	**7.92**	2.39	4.11

Length-matched random-position masking is undefined here because eligible background spans are shorter than the required document spans. We therefore report the document/question/answer-prefix comparison without imputing values for this unavailable control; the nonzero-source experiments are unaffected.

C.8 Activation Realism

The source-sensitivity experiments above establish causal efficacy. This complementary analysis asks a separate question: does each fixed replacement operator create an activation state farther from ordinary prompt states than clean/noisy prompt changes do? We reuse Llama 3-8B, the selected NQ L7–12 evidence band, and exactly the 60 post-discovery IDs used by nested mediation. No layer, source, or sample is reselected. For every sample, layer, and source, 32 seed-12 random projections estimate sliced Wasserstein distance (SWD); we also record normalized L2 shift, cosine distance, log norm deviation, and the sign-flip rate of the next layer's `gate_proj`. The five sources are zero, same-prompt background mean (mean-within), cross-sample poisoned-document mean (mean-across), aligned noisy residuals, and aligned clean residuals.

Let P, C, N denote the native poison, clean, and noisy document-position states and X the patched state. The SWD ratio is the minimum SWD from X to P, C, N, divided by the mean of the three pairwise native SWDs. For L2, cosine distance, absolute log norm deviation, and gate-pattern change, the denominator is the larger of the corresponding C-versus-P and N-versus-P changes. A ratio at or below one therefore places the intervention within that sample's observed clean/noisy envelope. Ratios diagnose realism only; causal efficacy remains the quantity reported in Tables S25–S26.

Table S28: Pooled residual-shift realism. Each sample is first averaged over L7–12. Full-coverage clean/noisy substitutions remain within the natural L2 and cosine envelopes. Synthetic operators have heavy-tailed means because some natural denominators are nearly zero; Table S30 reports robust medians.

Source	R_2 mean	$R_{\cos}$ mean
Zero	1.33×10^{7}	8.68×10^{5}
Mean-within	1.77×10^{7}	7.77×10^{5}
Mean-across	1.08×10^{7}	4.02×10^{5}
Noisy	0.863	0.994
Clean	0.862	0.994

Table S29: Pooled norm and next-layer gate-pattern realism under the same estimator as Table S28. The two natural nonzero substitutions remain within both envelopes, including the immediate downstream computation.

Source	R_{norm} mean	R_{gate} mean
Zero	2.46×10^{8}	3.61×10^{6}
Mean-within	1.01×10^{6}	3.09×10^{6}
Mean-across	6.65×10^{6}	3.02×10^{6}
Noisy	0.689	0.864
Clean	0.599	0.862

Table S30: Pooled sample-level medians. Mean-across retains the cross-sample operator's partial position support (median coverage 0.040) and is not used as a full-state realism control. The clean/noisy SWD ratio is zero because the patched state exactly equals one observed native state at every aligned position.

Source	Coverage	R_{SWD}	R_2	$R_{\cos}$	R_{norm}	R_{gate}
Zero	1.000	5.322	0.808	1.323	1761.0	1.070
Mean-within	1.000	7.236	1.063	1.157	7.172	0.903
Mean-across	0.040	1.535	0.691	0.617	22.05	0.938
Noisy	1.000	0.000	0.999	0.999	0.780	0.997
Clean	1.000	0.000	0.997	1.000	0.698	0.998

Table S31: Layerwise means for clean/noisy (C/N) full-coverage substitutions. The result is stable throughout the selected six-layer band rather than being driven by one layer.

Layer	R_2 C/N	$R_{\cos}$ C/N	R_{norm} C/N	R_{gate} C/N
L7	0.862/0.863	0.994/0.994	0.586/0.718	0.860/0.861
L8	0.861/0.863	0.993/0.994	0.565/0.717	0.862/0.865
L9	0.862/0.863	0.994/0.995	0.604/0.689	0.861/0.864
L10	0.862/0.863	0.994/0.995	0.601/0.651	0.866/0.867
L11	0.863/0.863	0.993/0.995	0.627/0.665	0.860/0.860
L12	0.862/0.863	0.994/0.993	0.610/0.698	0.865/0.866

Table S32: Band-replacement gap reductions in nats. Mean-across stays near zero (-0.15–0.44); $\Delta_{\mathrm{wi-ac}}$ is mean-within minus mean-across and ranges from 4.04 to 10.10 nats.

Model	Neutral	Noisy	Clean	$\Delta_{\mathrm{wi-ac}}$
Llama 3-8B	11.89	15.47	16.53	10.10
Llama 2-7B	6.40	7.18	7.76	4.04
Vicuna-7B	10.04	11.03	11.69	6.36
Vicuna-13B	10.47	10.82	11.72	8.34
Vicuna-33B	11.31	11.08	11.44	9.80

Table S33: Llama 3 semantic-source diagnostic on all 40 questions whose clean evidence contains the answer. Each question averages the six single-layer effects in L7–12. Answer-only advantage is the paired difference from evidence without the answer.

Llama 3 semantic source	N	Δg
Answer-only	40	26.68
Evidence without answer	40	11.62
Neutral	40	9.58
Answer-only advantage	40	15.06

The 1,800 observations span all 60 questions, six layers, and five source conditions. Clean and noisy substitutions stay within ordinary activation variation across the selected band; the heavy-tailed zero and synthetic-mean ratios are not used as realism evidence.

C.9 What the Band Carries

Neutral controls cover 68/80 Llama 3 samples and 35/40 samples for each smaller causal-suite run; unavailable alignments are excluded rather than imputed.

C.10 Route Across Generated Tokens

This experiment fixes the selected head and uses the 60 post-discovery NQ questions in each model. Clean and poisoned prompts greedily generate at most eight tokens. At every generated query position we capture the selected head's pre-output-projection state and its attention over the five document spans. Paired N includes only questions that generated that step in both conditions; it decreases after early end-of-sequence tokens. P/C is the ratio between the poisoned and clean condition means on that paired set. "Top poison" is the fraction of poisoned prompts for which a poison document has maximum document attention mass.

On baseline-attacked cases only, a second experiment zeros the selected head's pre-output-projection slice at the first 1, 4, 8, or all answer-scoring query positions and during the corresponding initial autoregressive calls ("all" uses all calls up to the 32-token decoding cap). The margin outcome is the baseline wrong-minus-correct answer log probability minus its intervened value. Generation outcomes use the same fixed attack-removal and correct-recovery definitions as the other intervention audits.

Across all five models, the same 60 post-discovery questions are used for clean and poisoned trajectories and for the four intervention horizons. The result extends route persistence beyond the first decoded token: a concentrated readout remains visible over generation, but this single-head interface is not sufficient for generation-level correction.

C.11 Multi-Head Diagnostics

Llama 3 multi-head circuit. The best prefix is $K = 19$ at 3.27 nats, or 0.597 of the 5.483-nat reference patch, with an additivity ratio of about 1.41. Pairwise coverage is 1/171, so this is a prefix bound rather than a minimal-circuit claim.

Table S34: Generated-token trajectory. Poison remains the maximum-mass document through step 8 in the four models with concentrated partial readout; Llama 2 has lower selected-head document concentration. Ratios are descriptive paired condition means, not classifier scores.

Model	Metric / step	1	2	3	4	5	6	7	8
Llama 3-8B	Paired N	60	60	60	57	52	45	43	43
	Context mass P/C	1.02	1.09	0.97	1.12	1.09	1.05	1.15	1.04
	Max-doc mass P/C	0.89	0.89	0.78	0.88	0.98	0.89	0.96	0.87
	Top poison	0.77	0.75	0.72	0.68	0.63	0.69	0.70	0.77
	State norm P/C	1.07	0.87	0.90	1.05	1.05	0.98	1.14	1.08
Llama 2-7B	Paired N	60	60	60	60	60	58	56	53
	Context mass P/C	0.92	1.21	0.93	0.93	0.81	0.78	0.92	1.02
	Max-doc mass P/C	0.90	1.12	0.83	0.87	0.83	0.76	0.87	0.96
	Top poison	0.37	0.43	0.33	0.30	0.32	0.26	0.25	0.28
	State norm P/C	1.06	0.97	0.73	0.78	0.80	0.70	0.88	0.85
Vicuna-7B	Paired N	60	60	60	55	54	52	52	48
	Context mass P/C	1.08	0.98	1.28	1.07	1.25	1.34	1.27	1.45
	Max-doc mass P/C	0.97	0.93	1.26	1.00	1.09	1.22	1.23	1.29
	Top poison	0.77	0.73	0.78	0.65	0.59	0.62	0.63	0.77
	State norm P/C	0.99	1.03	1.20	1.08	1.15	1.14	1.15	1.29
Vicuna-13B	Paired N	60	60	60	56	55	54	52	51
	Context mass P/C	1.08	1.09	1.28	1.10	1.22	1.34	1.21	1.28
	Max-doc mass P/C	0.91	0.94	1.13	0.91	0.99	1.16	1.07	1.14
	Top poison	0.82	0.75	0.80	0.77	0.73	0.69	0.71	0.75
	State norm P/C	0.96	1.05	1.11	1.08	1.06	1.15	1.03	1.17
Vicuna-33B	Paired N	60	60	60	57	56	53	53	52
	Context mass P/C	1.05	1.09	1.19	1.16	1.28	1.30	1.21	1.18
	Max-doc mass P/C	0.98	1.01	1.08	1.05	1.16	1.17	1.06	1.01
	Top poison	0.80	0.80	0.80	0.67	0.68	0.66	0.66	0.73
	State norm P/C	1.11	1.09	1.06	1.04	1.16	1.03	1.06	1.06

Table S35: Multi-token selected-head intervention. Margin effects in the four localized models largely saturate by four steps; attack removal remains at most 0.098 and recovery at most 0.049.

Model	Horizon	N_a	Gap reduction	Attack removal	Recovery
Llama 3-8B	1	39	0.905	0.026	0.000
	4	39	1.076	0.026	0.000
	8	39	1.078	0.026	0.000
	All	39	1.078	0.026	0.000
Llama 2-7B	1	43	−0.011	0.023	0.000
	4	43	0.017	0.023	0.000
	8	43	0.019	0.023	0.000
	All	43	0.022	0.023	0.000
Vicuna-7B	1	42	0.326	0.024	0.000
	4	42	0.378	0.024	0.000
	8	42	0.379	0.024	0.000
	All	42	0.379	0.024	0.000
Vicuna-13B	1	44	0.289	0.023	0.000
	4	44	0.331	0.023	0.000
	8	44	0.333	0.023	0.000
	All	44	0.336	0.023	0.000
Vicuna-33B	1	41	0.564	0.049	0.000
	4	41	0.657	0.073	0.000
	8	41	0.654	0.049	0.000
	All	41	0.658	0.098	0.049

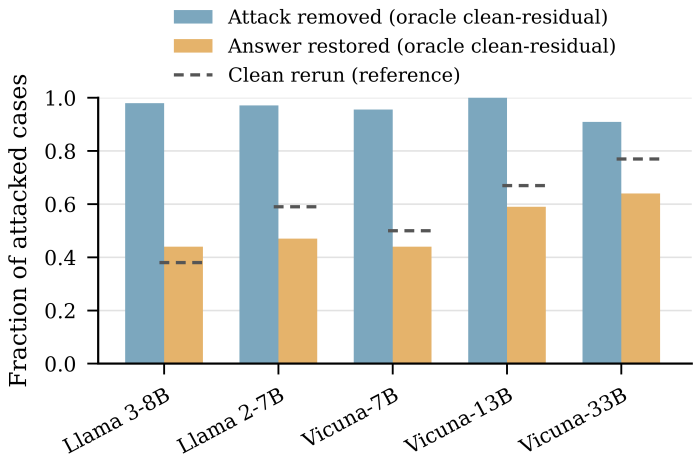


Figure S3: Suppression and recovery on the original attacked sets. The standardized all-layer clean-residual oracle removes 0.91–1.00 of attacks (blue) but restores only 0.44–0.64 of answers (orange); dashed lines show clean-context competence. The human-adjudicated common-denominator comparison below is the primary operator result.

Decode-time readout patch. Clean substitution at L17-H24 shifts the answer margin by $+5.48$ nats, establishing the reference scale for the selected readout.

L13-H18 intervention diagnostic. L13-H18 replacement moves ASR from 1.00 to 0.00 within its group, but final-position replacement changes the margin by only $+0.45$ and perturbs L17-H24 by about 5.1%. Across 50 generation cases it suppresses 13, of which 8 become IDK, separating refusal or commitment disruption from correct-answer readout.

Cross-model cases. The qualitative analysis contains 24 rows over 9 cases: nine rows are refusal disruptions and four are answer corrections; the remainder persist or change without correction. Together these diagnostics motivate band-level control and distinguish answer readout from generation commitment.

D Suppression and Correction

For compactness, each cell in Tables S36 and S37 is "ASR reduction / correct-answer recovery" on the same baseline-attacked set. The clean-context rerun is a reference and therefore has no intervention ASR-reduction entry.

Table S36: Oracle clean-state restoration and no-context fallback on the original attacked sets. All-layer restoration removes 0.91–1.00 of attacks and recovers 0.44–0.64 of answers. The clean rerun supplies an oracle reference; Table S39 controls answerability.

Model	Band patch	All-layer patch	No context	Clean rerun
Llama 3-8B	0.96/0.36	0.98/0.44	0.98/0.64	–/0.38
Llama 2-7B	0.97/0.47	0.97/0.47	0.94/0.51	–/0.59
Vicuna-7B	0.96/0.43	0.96/0.44	0.97/0.35	–/0.50
Vicuna-13B	0.96/0.59	1.00/0.59	1.00/0.59	–/0.67
Vicuna-33B	1.00/0.73	0.91/0.64	1.00/0.64	–/0.77

D.1 Human-Adjudicated Comparison

Annotator A and Annotator B are two independent human annotators whose identities are anonymized. They independently complete 237 eligibility items and 320 response-annotation items

Table S37: Local multi-vector and strong steering operators. Stronger deletion steering does not monotonically improve correct-answer recovery.

Model	Multi-vector 2:1	Multi-vector 2:2	Oracle steer 3	Oracle steer 4
Llama 3-8B	0.76/0.14	0.78/0.16	0.82/0.10	0.90/0.08
Llama 2-7B	0.51/0.24	0.59/0.29	0.71/0.29	0.87/0.19
Vicuna-7B	0.76/0.13	0.87/0.16	0.87/0.07	0.94/0.06
Vicuna-13B	0.78/0.30	0.89/0.37	0.89/0.33	0.93/0.22
Vicuna-33B	0.64/0.36	0.64/0.36	0.77/0.36	0.91/0.41

before labels are compared. Direct agreement covers 189 eligibility items and 284 response items; the remaining 48 and 36 items, respectively, are adjudicated separately. Annotation and adjudication records are aligned by content without exposing sample identifiers.

Table S38: Inter-annotator agreement before adjudication. The two fields that define fair-subset eligibility and the six-way response semantic class show 0.911–0.983 raw agreement and $\kappa = 0.831$–0.966. Lower κ on rare contradiction/negation labels reflects class prevalence; those labels do not define the fair subset.

Task	Field	N	Agreement	κ	Unclear A/B
Eligibility	Evidence supports answer	237	0.911	0.831	4/12
Eligibility	Clean response correct	237	0.983	0.966	1/3
Eligibility	Contradiction present	237	0.895	0.290	4/5
Response	Semantic class	320	0.947	0.929	11/13
Response	Negates correct answer	320	0.950	0.370	0/11
Response	Negates attacker answer	320	0.956	0.350	0/10

After adjudication, evidence support is yes/no/unclear for 128/103/6 items; clean-response semantic correctness is yes/no/unclear for 140/96/1; and contradiction is yes/no/unclear for 6/228/3. The 320 response labels are 88 attacker, 106 correct, 3 both, 49 refusal, 63 other, and 11 unclear. Requiring both evidence-support=yes and clean-response-correct=yes, then pairing all operators on identical attacked sample IDs, produces the strict $N = 106$ subset.

Table S39: Pooled operator comparison on the human-adjudicated fair subset. Every row uses the same 106 attacked cases. All interventions except no-context fallback use paired clean evidence.

Operator	Removal n/N (rate)	Recovery n/N (rate)
Oracle all-layer restore	105/106 (0.991)	77/106 (0.726)
Oracle six-layer restore	105/106 (0.991)	76/106 (0.717)
No-context fallback	100/106 (0.943)	61/106 (0.575)
Oracle steering $\alpha = 3$	88/106 (0.830)	20/106 (0.189)
Oracle steering $\alpha = 4$	98/106 (0.925)	14/106 (0.132)
Multi-vector 2:1	75/106 (0.708)	24/106 (0.226)
Multi-vector 2:2	78/106 (0.736)	27/106 (0.255)

The next two tables provide the per-model decomposition. Each cell is attack removal / correct-answer recovery; denominators are fixed within model and operator.

Paired operator differences on the human-fair subset. The same 106 eligible model–question events span 42 NQ questions. We retain the original eligibility and recovery flags and compute six-layer restoration minus each comparator. Table S42 gives the full paired contingency counts and recovery differences.

On cases where human judgment establishes that the correct answer is supported and achievable, six-layer oracle clean-state restoration nearly matches all-layer oracle restoration in pooled recovery. The same denominator retains a large gap between attack removal and answer recovery under oracle steering, identifying the evidence band as a compact oracle interface for reinstating clean document state.

Table S40: Oracle restoration and no-context fallback on the human-fair subset (removal/recovery). Per-model values complement the pooled comparison.

Model	N	All-layer	Evidence band	No context
Llama 3-8B	21	1.000/0.762	1.000/0.619	0.952/0.619
Llama 2-7B	33	1.000/0.667	1.000/0.697	0.879/0.515
Vicuna-7B	31	0.968/0.710	0.968/0.677	0.968/0.484
Vicuna-13B	12	1.000/0.750	1.000/0.833	1.000/0.750
Vicuna-33B	9	1.000/0.889	1.000/1.000	1.000/0.778

Table S41: Human-fair per-model oracle steering and multi-vector results (removal/recovery). Increasing steering strength consistently improves or preserves removal but does not reconstruct the correct answer; pooled recovery falls from 0.189 to 0.132.

Model	N	Oracle steer 3	Oracle steer 4	Multi-vector 2:1	Multi-vector 2:2
Llama 3-8B	21	0.857/0.000	0.905/0.048	0.762/0.190	0.762/0.190
Llama 2-7B	33	0.727/0.242	0.879/0.121	0.515/0.242	0.515/0.242
Vicuna-7B	31	0.839/0.065	0.935/0.097	0.806/0.129	0.806/0.129
Vicuna-13B	12	1.000/0.417	1.000/0.167	0.750/0.250	1.000/0.417
Vicuna-33B	9	0.889/0.556	1.000/0.444	0.889/0.556	0.889/0.667

Table S42: Paired recovery on the common $N = 106$ subset. Counts denote both operators recovering, six-layer restoration only, comparator only, and neither. Differences are six-layer restoration minus the comparator, in percentage points.

Comparator	Both	Band only	Comparator only	Neither	Δ
No context	47	29	14	16	$+14.15$
All-layer restoration	70	6	7	23	-0.94
Oracle steering $\alpha = 4$	11	65	3	27	$+58.49$

Machine semantic audit of generated answers. We retain the same 106-event eligibility set and audit all 742 stored responses from the seven operators in Table 5. Exact responses shared by operators within a model–question event are judged once, yielding 542 items. Three language-model judging agents read disjoint, shuffled batches containing the question, reference, attacker answer, and complete response, with operator identities and original outcome flags withheld. A second machine pass reviews 79 uncertain, string-discrepant, or dual-string items while operator identities remain hidden. Judgments assess the answer adopted relative to the stored reference, accepting aliases and semantic equivalents; negated mentions, unresolved conflicting answers, and incomplete date precision do not count as semantic recovery. These are machine-produced response labels; the original human eligibility labels remain fixed.

Table S43 reports both endpoints. Among the 18 outputs containing both candidate strings, 16 adopt the reference and two do not. Across all 742 outputs, semantic review recognizes 36 correct equivalents missed by reference matching and excludes 14 reference matches that do not establish adoption. Six-layer restoration achieves 75.47% semantic recovery, versus 80.19% for all-layer restoration and 13.21% for strong steering. Six-layer minus steering is $+62.26$ percentage points, minus no context is $+16.04$, and minus all-layer restoration is -4.72. The original reference-matching results and the semantic audit both distinguish restoration from pooled steering.

Table S43: Stored-response audit on the same $N = 106$ events for every operator. Reference match is the original endpoint reported in the main comparison; semantic recovery requires adopting the reference or an equivalent answer. Unresolved responses are excluded from the recovery numerator and retained in the denominator. No target-model generations are added.

Operator	Reference match	Semantic recovery	Unresolved
Six-layer restoration	76 (71.70%)	80 (75.47%)	8
All-layer restoration	77 (72.64%)	85 (80.19%)	7
No-context fallback	61 (57.55%)	63 (59.43%)	4
Oracle steering $\alpha = 3$	20 (18.87%)	22 (20.75%)	3
Oracle steering $\alpha = 4$	14 (13.21%)	14 (13.21%)	4
Multi-vector $(2, 1)$	24 (22.64%)	27 (25.47%)	2
Multi-vector $(2, 2)$	27 (25.47%)	30 (28.30%)	2

Table S45 reports patch recovery conditional on whether the clean rerun is correct or wrong. Oracle patching recovers 0.75–0.84 of clean-answerable cases but only 0.00–0.19 of the rest, and McNemar detects no significant patch–vs–clean-rerun difference in any model ($p \geq 0.25$); output-correctness agreement is 0.78–0.86. This is not an equivalence test: it shows that observed patch recovery is concentrated on samples that the clean-context run can answer.

In separate $N = 100$ Llama 2 and Vicuna-7B replications at $\alpha = 2$, attacked denominators are 70 and 68, ASR reductions are 0.500 and 0.765, and recoveries are 0.214 and 0.132. These larger runs reproduce the suppression–recovery asymmetry; they do not enlarge the other three models' causal samples. At the same-example oracle steering point ($\alpha = 2$), positive poisoned-margin reductions occur in 0.94, 0.926, 0.962, 0.889, and 0.773 of attacked cases for Llama 3, Llama 2, Vicuna-7B, Vicuna-13B, and Vicuna-33B. Exact paired sign-test p values are 1.85×10^{-11}, 2.82×10^{-6}, 4.02×10^{-7}, 2.46×10^{-5}, and 0.00845, respectively.

D.2 Sparse-Poisoning Operator Comparison

We reuse the historical coordinates in the single-poison NQ setting of Appendix E.1. All 116 model–question events are evaluated; the common automatically eligible attacked subset contains 35 events from 21 questions (18 Llama 3 and 17 Vicuna-13B). The original six-layer bands are zero-based L7–12 and L5–10, respectively. The adjacent comparison bands are L13–18 and L11–16, each six layers wide with no overlap with the original band. Clean donor states come from the first five nonpoison passages in the same retrieval snapshot. Restoration, adjacent-band restoration, and pooled oracle steering at $\alpha = 4$ use the same aligned donor; clean-context and no-context runs use the same question, prompt wrapper, and 150-token greedy decoding cap. The historical bands and heads remain fixed throughout this comparison.

Table S44: Recovery summary for the strongest oracle and clean-context operators. Attacked denominators are 50, 70, 68, 27, and 22.

Model	Local	Band clean	All clean	No ctx.	Clean rerun
Llama 3-8B	0.16	0.36	0.44	0.64	0.38
Llama 2-7B	0.29	0.47	0.47	0.51	0.59
Vicuna-7B	0.16	0.43	0.44	0.35	0.50
Vicuna-13B	0.37	0.59	0.59	0.59	0.67
Vicuna-33B	0.36	0.73	0.64	0.64	0.77

Table S45: Paired all-layer clean patch vs. clean-context rerun on matched attacked samples. The four counts report both recover / patch only / clean rerun only / neither.

Model	N	Both/Patch/Rerun/Neither	Patch rec.|rerun correct	Patch rec.|rerun wrong	McNemar p
Llama 3-8B	50	16/6/3/25	0.84	0.19	0.50
Llama 2-7B	27	12/2/4/9	0.75	0.18	0.68
Vicuna-7B	26	10/2/3/11	0.77	0.15	1.00
Vicuna-13B	27	15/1/3/8	0.83	0.11	0.62
Vicuna-33B	22	14/0/3/5	0.82	0.00	0.25

Table S46: Five paired operators under single-poison mixed retrieval. Suppression counts responses no longer satisfying the targeted attack criterion; recovery follows the fixed reference-matching rule. All rows use the same 35 automatically eligible events.

Operator	Attack suppression	Reference-answer recovery
Original six-layer restoration	34/35	12/35
Adjacent six-layer restoration	34/35	12/35
Oracle steering $\alpha = 4$	25/35	12/35
Clean-context generation	33/35	9/35
No-context generation	33/35	17/35

Original-band restoration suppresses the attack on 34/35 events and recovers the reference on 12/35, preserving the separation between suppression and recovery. Relative to steering, its suppression difference is $+25.71$ percentage points. The three intervention operators each recover 12 events overall; original-band and steering recovery agree on seven successes, with five successes unique to each. Original-band and adjacent-band restoration have the same 7/5/5/18 both/original-only/adjacent-only/neither counts. Per-model recovery counts for original band, adjacent band, steering, clean context, and no context are 6/9/6/6/11 for Llama 3 and 6/3/6/3/6 for Vicuna-13B. Table S47 gives paired recovery differences.

Table S47: Paired recovery differences on the same 35 events.

Original band minus	Recovery difference (pp)
Oracle steering $\alpha = 4$	0.00
Adjacent six-layer restoration	0.00
Clean-context generation	8.57
No-context generation	-14.29

This cohort uses automatic attack eligibility, whereas the $N = 106$ comparison additionally requires human-adjudicated clean-evidence support and clean-response correctness. The two recovery rates therefore refer to different eligibility sets. Within the sparse set, original-band restoration recovers 7/9 events with a reference-matching clean-context response and 5/26 of the remaining events. Excluding the one clear nonadoption event leaves 34 events and unchanged recovery counts for all five operators; excluding the two ambiguous events as well leaves 32, with recovery counts 12/12/11/9/17 in the table's operator order. These audits retain the original outputs and show the suppression–recovery distinction across the reported eligibility variants.

Table S48: Representative best or diagnostic settings from the lightweight correction operators. The highest recovery incurs 0.48 collapse; no family meets both the recovery and clean-cost criteria.

Family / best setting	Rec.	Persist.	Collapse	Clean cost
Head-anchor projection / 0.5–1.5	0.00	0.98	0.02	0.00
Poison-doc attention mask / 0.4–0.6	0.02	0.06	0.92	1.00
L12 residual projection / 1.5	0.08	0.84	0.08	0.05
Contrastive steering / 3.0	0.04	0.84	0.12	0.11
Token-level redaction / 0.3	0.34	0.18	0.48	0.26
Contiguous-span redaction / $\tau = 0.4,\ k = 2$	0.12	0.70	0.18	0.11
Relative-threshold redaction / 1.5	0.08	0.76	0.16	0.16

D.3 Lightweight Correction Operators

Before the multi-operator comparison, five inexpensive correction families were evaluated on 50 attacked Llama 3/NQ examples against fixed recovery and clean-cost criteria.

Together, these negative results motivate comparison with the stronger oracle clean-restoration operators reported in Table 5.

E Route Localization

The remaining sections evaluate route-based document attribution through document localization, prompt-level detection, abstention, deployment cost, retrieval stress, and adaptive dilution.

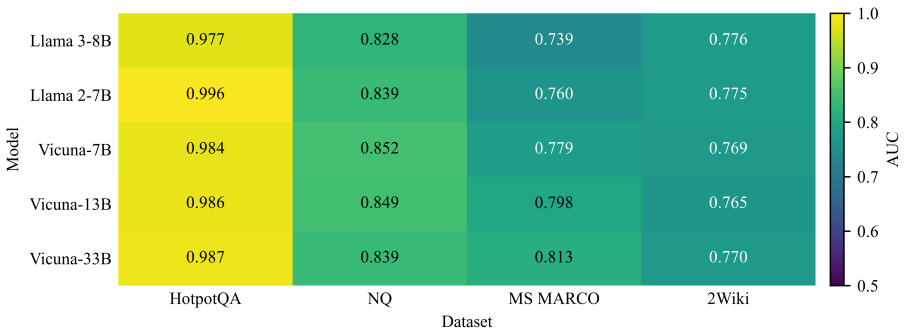


Figure S4: Token-count-residualized clean/poison classification AUC over 20 model–dataset settings. Mean AUC is 0.844, the floor is 0.739, and HotpotQA spans 0.977–0.996. A final-layer/$k = 10$ variant has mean 0.819 and floor 0.684. Each setting has 42–83 clean and 55–96 poisoned examples.

On each training fold, this 20-setting experiment removes linear token-count information. The attack-family and lexical-confound experiment below additionally controls the attacker-answer string and attack template while changing the prompt wrapper; retrieval rank and document count remain uncontrolled. These probes measure prompt-level clean/poison separability, whereas residual-source interventions provide the causal evidence.

Table S49: Document localization reaches high precision in three models. Llama 2 has lower selected-head localization, and answer-direction readout-head projection does not improve token attribution. This analysis covers the four models shown.

Model	Attack N	Attention doc AUC	Readout token AUC
Llama 3-8B	50	0.976	0.42–0.46
Llama 2-7B	27	0.51–0.56	0.42–0.46
Vicuna-7B	26	0.963	0.42–0.46
Vicuna-13B	27	1.000	0.42–0.46

The localization result is document-level. Attention/value-norm routing identifies which retrieved document carries poison, while the readout-head token projection is at or below chance. The route therefore supports abstention or document review rather than automatic token redaction.

E.1 Direct Traceback Comparison

Matched inputs and fixed coordinates. We evaluate Meta-Llama-3-8B-Instruct and Vicuna-13B-v1.3 on 58 NQ and 100 HotpotQA questions (316 model–question events). The 58 NQ questions are disjoint from the historical 40-question coordinate-discovery set; two of the requested 60 have empty original top-five passages and are excluded jointly across methods. We insert the first original PoisonedRAG adversarial passage, retain natural Contriever order, and cache the top 100 candidates. Targets use the canonical prompt and greedy decoding with a 150-token cap. Eligibility requires an observed response containing the normalized attacker answer but not the canonical reference, and a consumed top-five context containing one poison and four clean passages. This yields 143 events from 88 questions. For each event, we hold the question, observed answer, and consumed passages fixed across methods; candidate IDs are neutralized and evaluation labels are withheld. Our heads are fixed at zero-based L17-H24 for Llama 3 and L18-H18 for Vicuna-13B, with no retuning on either dataset. A document's score is its mean attention from that head at the final prompt position.

Baseline implementation. We adapt the authors' implementations of RAGForensics (Zhang et al., 2025) and RAGOrigin (Zhang et al., 2026), retaining judging prompts, scope construction, and scoring rules. All judge and proxy calls use local Llama-3-8B-Instruct; expanded-context generation uses the corresponding target checkpoint. RAGOrigin separately standardizes answer likelihood, question likelihood, and retrieval similarity, averages the three features, and uses two-cluster KMeans with seed 42. Each likelihood feature is a mean continuation log probability, distinct from the summed-answer gap used in our causal analysis. Native scope expansion uses the shared top-100 cache, and evaluation projects the resulting scores onto the original consumed five. Nine RAGForensics and two RAGOrigin scopes reach the cache limit before their native stopping criterion; all 143 original five-document predictions are complete.

The strict RAGForensics output parser produces complete consumed-five predictions on 116 events. The format-repair variant reuses every valid cached decision, normalizes an unambiguous standalone `Label: Yes/No` line, and otherwise permits one fixed format-only follow-up. It completes all 143 events and changes none of the original five binary decisions on the strict 116. Binary ties are evaluated by the analytic expectation over uniformly ordered tied documents. Thus Top-1 measures one-document inspection success, while native-set recall measures whether the poison is included anywhere in the flagged set.

Localization. Table S50 reports the strict comparison and all four dataset–model cells. On the strict 116, Top-2 recall is 100% for mean attention and RAGOrigin, 57.27% for RAGForensics, and 92.24% for retrieval rank. On the full 143, mean attention and RAGOrigin both succeed on 135 events, with three mean-only and five Origin-only successes. Their paired Top-1 difference is -1.40 percentage points. The differences from retrieval rank and RAGForensics are $+11.19$ and $+66.46$ points. When poison has retrieval rank 2–5, mean attention succeeds on 21/21 events and Origin on 19/21 (14 questions); at rank 1, the counts are 117/122 and 121/122. Document-max attention and attention–value-norm variants have Top-1 rates of 90.91% and 89.51%, respectively, and Top-2 recall of 100% and 99.30%.

Native prediction sets and response audit. RAGForensics flags the poison in 142/143 events (99.30% set recall), with 403/572 clean passages flagged (70.45% false-positive rate) and a mean set size of 3.81/5. RAGOrigin flags all 143 poisons, with 84/572 clean passages flagged (14.69%) and mean size 1.59/5. Broad binary flagging explains the distinction between RAGForensics' high set recall and its lower Top-1 ranking precision under this local backend. An AI-assisted reading of all 143 stored responses identifies five clear nonadoption cases and two ambiguous cases under the original string rule. These exclusions are reported as separate sensitivity analyses in Table S50; the original labels remain unchanged. The audit evaluates response wording against stored references and is separate from the human eligibility annotations used for the $N = 106$ operator comparison.

Table S50: Top-1 localization (%) with matched denominators. Dataset–model and audit rows use the format-repair protocol. Mean denotes fixed-head mean attention; RF denotes RAGForensics. The strict and full cohorts contain 77 and 88 distinct questions.

Protocol / data	Model	N	Mean	Origin	RF	Retrieval
Strict common	Both	116	95.69	97.41	30.36	86.21
Format repair, all	Both	143	96.50	97.90	30.05	85.31
NQ	Llama 3-8B	18	100.00	94.44	32.69	61.11
NQ	Vicuna-13B	17	100.00	94.12	38.43	58.82
HotpotQA	Llama 3-8B	51	94.12	98.04	28.01	94.12
HotpotQA	Vicuna-13B	57	96.49	100.00	28.54	92.98
Audit: exclude clear	Both	138	97.10	97.83	30.31	85.51
Audit: also exclude ambiguous	Both	136	97.06	97.79	30.37	86.03

Attribution examples. For an NQ question about the filming location of *War for the Planet of the Apes*, the poison passage appears fifth in retrieval order. Mean attention ranks it first for both models; RAGOrigin retains it in the top two, with its standardized retrieval component raising another passage to first. A HotpotQA question about a *SuperMansion* actress and her *Workaholics* role illustrates the complementary case: the response combines an actress name from a clean passage with a role from the poison. Mean attention ranks the name-bearing clean passage first, whereas RAGOrigin ranks the poison first. These examples connect the observed differences to the methods' signals: internal evidence use for mean attention and combined question, answer, and retrieval responsibility for RAGOrigin.

Additional model calls. Table S51 counts the recorded logical inference workload per event within the shared candidate cache, including requests served from cache. It excludes shared observed-answer generation and offline head selection. RAGForensics' format-only follow-ups have median zero and mean 0.36 calls per event. During the repair run, 893 unique new generations were needed, including 52 format follow-ups; this incremental cache workload is separate from the recorded logical call counts. Our median route-extraction time is 0.123 seconds under the concurrent experiment load. The call types in the table are reported separately, without converting them into a common FLOP or speed ratio.

Table S51: Additional calls for the 143-event traceback comparison within the shared candidate cache. Likelihood forwards, prefills, and generations have different computational costs.

Method	Median additional logical calls per event
Mean attention	1 target-model prefill with white-box head access
RAGForensics	16 judge generations; format-only follow-ups counted separately
RAGOrigin	20 proxy likelihood forwards, 1 judge generation, and 1 target generation

F Detection and Abstention

F.1 Detector Comparison

Table S52 uses the same clean and poisoned examples for both methods. The RevPRAG-inspired detector uses the layer-averaged last-query representation (`all_mean`) and scores each prompt by the difference between its mean cosine similarity to the 20 nearest poisoned and clean support examples. The route score is the maximum, across documents, of mean per-token attention. This RevPRAG-inspired detector is evaluated under the same open-weight protocol as the route score rather than reproducing the original pipeline.

Each NQ/HotpotQA/MS MARCO setting has 100 clean and 100 poisoned rows; each 2Wiki setting has 200 of each. Clean-correct denominators range from 42 to 83 and attack-active denominators from 42 to 92. Route AUROC spans 0.624–0.826, while FPR at TPR 0.9 spans 0.31–0.85. Selective

Table S52: Detector head-to-head on NQ and HotpotQA.

	AUPRC		AUROC	
Model	RevPRAG-inspired	Route	RevPRAG-inspired	Route
NQ				
Llama 2-7B	0.845	0.625	0.821	0.694
Llama 3-8B	0.882	0.712	0.876	0.766
Vicuna-7B	0.900	0.710	0.875	0.783
Vicuna-13B	0.902	0.772	0.879	0.826
Vicuna-33B	0.899	0.736	0.876	0.809
HotpotQA				
Llama 2-7B	0.996	0.663	0.994	0.747
Llama 3-8B	0.994	0.589	0.992	0.691
Vicuna-7B	0.999	0.676	0.999	0.785
Vicuna-13B	0.999	0.669	0.999	0.783
Vicuna-33B	0.999	0.605	0.999	0.707

Table S53: Detector head-to-head on MS MARCO and 2Wiki. Across all settings, the RevPRAG-inspired detector wins 20/20 AUROC and 20/20 neutralization comparisons at the 0.25 budget. The route score contributes causal interpretation and localization.

	AUPRC		AUROC	
Model	RevPRAG-inspired	Route	RevPRAG-inspired	Route
MS MARCO				
Llama 2-7B	0.834	0.577	0.805	0.627
Llama 3-8B	0.823	0.705	0.789	0.748
Vicuna-7B	0.838	0.695	0.823	0.727
Vicuna-13B	0.838	0.683	0.811	0.738
Vicuna-33B	0.846	0.707	0.831	0.733
2Wiki				
Llama 2-7B	0.817	0.800	0.764	0.738
Llama 3-8B	0.825	0.602	0.791	0.650
Vicuna-7B	0.816	0.648	0.781	0.624
Vicuna-13B	0.823	0.683	0.783	0.678
Vicuna-33B	0.812	0.726	0.774	0.705
All 20 (mean)	–	–	0.863	0.728

risk decreases with abstention in most settings but is nearly flat on 2Wiki. Calibration is dataset dependent: a threshold set to 0.25 clean cost on one dataset reaches 0.93 clean cost in the worst off-diagonal transfer.

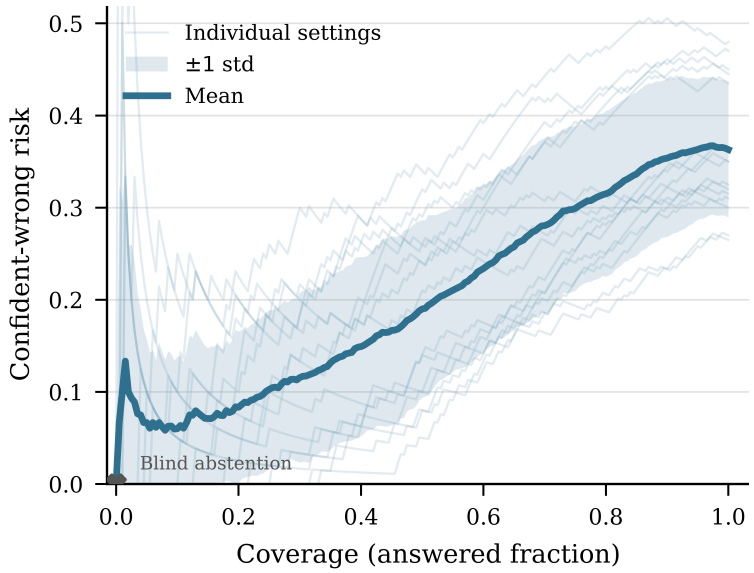


Figure S5: Route selective prediction over five models and three datasets (100 clean/poisoned queries per setting); raw curves are unsmoothed.

The strongest token-count-residualized configuration of the RevPRAG-inspired detector still has mean AUROC 0.844; its advantage is therefore not removed by the principal length control. In contrast, normalized document attention share (`attention_share`) and contribution-share metrics are direction-inverted (mean AUROC about 0.36), which is why the fixed route score uses maximum document-level mean attention.

F.2 Detector Audits

Activation centroids (4 models). The band centroid gives raw AUC 0.844–0.886 and neutralization 0.852–1.000. After document-count control, residual AUC is 0.675–0.702 and the exact Llama 3 caliper result is 0.540, showing that the raw separation is strong but partly length-confounded.

Matched and ReDeEP-inspired detectors (12 settings). The supervised matched route detector reaches AUC 0.666–0.802, wins 8/12 cells at budget, and has mean $\Delta = 0.052$, although two cells fall by 0.500 and 0.244. The concentration-only heuristic reaches AUC 0.154–0.497; its centroid variant reaches 0.392–0.852; adding length reaches 0.660–0.969. Thus distributional route features carry signal, while length is also a calibration dependency. These ReDeEP-inspired variants evaluate comparable representation signals under our protocol rather than reproducing the original training and evaluation pipeline.

RevPRAG-inspired detector support analysis (4 models). This variant uses residual-band activations averaged over document tokens. Across k = 1–20, raw AUC is 0.703–0.860 and residualized AUC is 0.704–0.784. The activation- distance signal survives residualization, with a model-dependent best k; this is a lightweight variant evaluated under our protocol.

Cross-dataset transfer. Across 48 train/test cells, held-out AUC is 0.497–0.895. The length-aware route distribution neutralizes 0.612 versus 0.396 for the length-free variant, with poison precision 0.971.

Last-token operating points. The prespecified final-layer/$k = 10$ evaluation covers 12 settings and 921 attacked examples, neutralizing 0.800 at 0.243 clean cost. A stronger exploratory variant reaches 0.943 at 0.277 cost; all primary comparisons retain the prespecified setting.

Attack-family and lexical confound controls. We fix the all-layer/$k = 5$, fold-local token-count-residualized configuration and train the leave-one-question-out support score on either LM-targeted or HotFlip examples. Each row below reports the range over the two training directions. "Corrected" replaces each exact attacker-target occurrence inside retrieved poison text by the correct answer ($n = 82$–97 where replacement is possible). "Wrong mention" appends the target as an unrelated glossary entry to clean evidence; "Alt." changes the prompt wrapper while preserving evidence.

Table S54: AUC controls for the fixed all-layer/$k = 5$ residualized probe. Within denotes same-family evaluation; Cross trains on one attack family and evaluates the other.

Setting	Within	Cross	Corrected	Wrong mention	Alt.
Vicuna-7B / NQ	0.991–0.993	0.971–0.980	0.989–0.990	0.527–0.529	0.820–0.840
Vicuna-7B / HotpotQA	0.992–0.995	0.950–0.984	0.956–0.990	0.514–0.523	0.801–0.831
Vicuna-7B / MS MARCO	0.938–0.969	0.845–0.846	0.925–0.973	0.511–0.515	0.688–0.725
Llama 2-7B / NQ	0.994–1.00	0.915–0.991	0.994–1.00	0.520–0.533	0.781–0.849

Clean correct-answer mentions are likewise near chance (AUC 0.510–0.551), and direct wrong-vs.-correct mention discrimination is 0.477–0.499. The fixed residualized final-layer/$k = 10$ control gives within-family AUC 0.901–1.000, cross-family 0.684–0.993, corrected 0.847–1.000, and wrong-mention 0.502–0.534. The all-layer result is therefore not an exact-answer-string detector, but target correction retains the rest of the adversarial passage distribution; neither probe establishes causal mediation.

Joint-GCG document evaluation. Across ten settings and 356 attacked examples, the fixed maximum document-level mean-attention score reaches 0.713 neutralization at 0.242 clean cost and 0.921 poison precision among flagged attacked cases. The four models with concentrated partial readout reach 0.788 / 0.243 / 0.962; Llama 2 has lower selected-head localization.

Table S55: Joint-GCG document evaluation over HotpotQA and MS MARCO using the original cosine-similarity Contriever protocol. Poison precision is measured among flagged attack-active cases; Top-doc is poison localization over all 200 poisoned queries per model.

Model	Attack N	Neut.	Clean cost	Poison prec.	Top-doc
Llama 3-8B	68	0.853	0.243	1.000	0.675
Llama 2-7B	87	0.483	0.236	0.714	0.505
Vicuna-7B	72	0.792	0.244	0.930	0.620
Vicuna-13B	66	0.788	0.243	0.923	0.645
Vicuna-33B	63	0.714	0.243	1.000	0.670
All five	356	0.713	0.242	0.921	0.623

G Defense Comparisons

Under the separate Llama 3 baseline evaluation ($n = 100$ on each of NQ, HotpotQA, and MS MARCO), no-defense confident-wrong ASR is 0.85–0.87. Astute RAG reaches 0.39–0.69, TrustRAG reaches 0.16–0.22, and route abstention reaches 0.00–0.16 at the selected budget. The keyword-based RobustRAG adaptation does not improve the saturated-poisoning setting. Blind abstention reaches zero only at full clean cost. Route abstention uses one generator forward pass plus a lightweight probe; TrustRAG uses document clustering and three generator calls. The controlled profile below directly measures route-probe cost without inferring a cross-system speedup from incomparable baseline implementations.

Table S56: Cross-model defense-family comparison at the selected operating points. Steering/abstention neutralizes 0.481–0.692, while simply appending clean retrievals reaches only 0.053–0.294.

Model	Defense	N_a	Neut.	Rec.	Clean cost
Llama 3-8B	steer/abstain	50	0.600	0.120	0.386
Llama 3-8B	more clean	30	0.133	0.100	0.000
Llama 2-7B	steer/abstain	27	0.481	0.259	0.308
Llama 2-7B	more clean	19	0.105	0.105	0.000
Vicuna-7B	steer/abstain	26	0.692	0.231	0.318
Vicuna-7B	more clean	19	0.053	0.053	0.091
Vicuna-13B	steer/abstain	27	0.519	0.259	0.345
Vicuna-13B	more clean	17	0.294	0.294	0.125

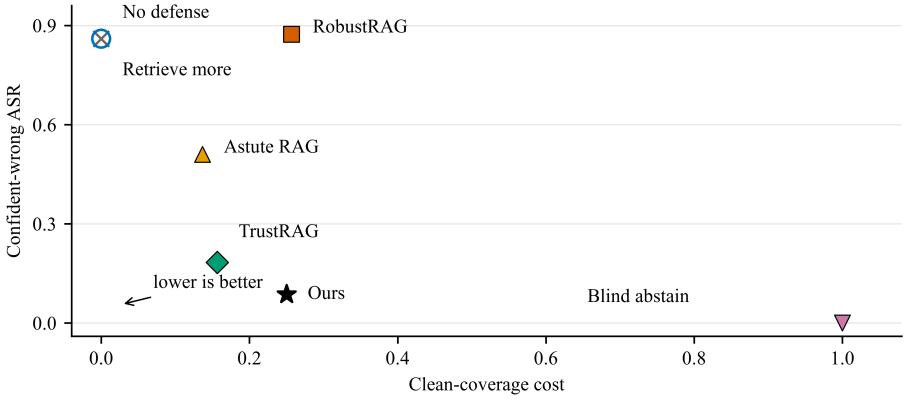


Figure S6: Confident-wrong ASR versus clean cost under mean 4.8/5 poison documents (Llama 3, three datasets, $n = 100$). Lower-left is better.

H Routing Cost

We profile the fixed selected-head document route on the same 100 seed-12 matched NQ questions for each causal-suite model. Each question produces one clean and one poisoned top-5 prompt. After one warm-up question per split, CUDA is synchronized around greedy generation (at most 32 new tokens) and around route extraction. A model-level pre-hook counts forwards, and allocated memory is recorded immediately before each operation and at its peak. Candidate retrieval and reranking are timed separately and excluded from the combined generator-plus-route time so that the table isolates the generator-side route computation. Profiling uses one RTX 4090 per model; Vicuna-33B uses mixed GPU/CPU inference with an 18-GiB GPU-memory allocation. Absolute generation time is therefore hardware-, response-length-, and placement-dependent, so the within-prompt overhead ratios are the appropriate comparison.

Each model contributes 100 clean and 100 poisoned profiles. Selected-head document routing uses exactly one forward pass and adds modest within-prompt overhead in these measurements.

I Retrieval Depth and Poison Density

This experiment fixes the five model-specific heads and the maximum document-level mean-attention score. Every cell uses the same NQ questions ($n = 100$), seed 12, dot-product Contriever, the canonical prompt, and 32-token greedy decoding. "Target output" is target-string match in this evaluation, not canonical reproduction ASR. Table S59 reports behavior and retrieval; Table S60 reports the fixed route score. No head or layer is re-selected.

Reducing m from five to one lowers target output in every model by 0.30–0.45 at $k = 5$ and 0.21–0.37 at $k = 10$; increasing retrieval depth has no monotone behavioral effect. The fixed route AUROC drops by 0.095–0.122 and 0.051–0.166, respectively. Even so, the four models with con-

Table S57: Per-prompt generator and selected-head document-routing time. Means are computed over the indicated prompts; Route/gen. averages the per-prompt ratio rather than dividing aggregate means. The route adds 8.5–12.8% in the pooled within-prompt comparison.

Model	Split	N	Gen. (s)	Route (s)	Route/gen. (%)	Combined (×)
Llama 3-8B	Clean	100	14.901	0.783	7.1	1.071
	Poisoned	100	6.276	0.675	18.6	1.186
	Pooled	200	10.589	0.729	12.8	1.128
Llama 2-7B	Clean	100	3.135	0.192	6.9	1.069
	Poisoned	100	0.971	0.074	11.8	1.118
	Pooled	200	2.053	0.133	9.3	1.093
Vicuna-7B	Clean	100	2.267	0.186	11.3	1.113
	Poisoned	100	0.787	0.068	14.3	1.143
	Pooled	200	1.527	0.127	12.8	1.128
Vicuna-13B	Clean	100	4.402	0.289	9.3	1.093
	Poisoned	100	1.451	0.109	12.9	1.129
	Pooled	200	2.927	0.199	11.1	1.111
Vicuna-33B	Clean	100	336.243	13.615	5.3	1.053
	Poisoned	100	230.031	12.861	11.8	1.118
	Pooled	200	283.137	13.238	8.5	1.085

Table S58: Forward counts and maximum allocated-memory peaks. "Incr." is the maximum increase above allocation immediately before the operation. Route extraction is exactly one forward in all 1,000 profiled prompts and adds at most 1.30 GiB in the observed runs.

Model	Split	Gen. forwards	Route forwards	Gen. peak / incr. (GiB)	Route peak / incr. (GiB)
Llama 3-8B	Clean	21.56	1.00	14.79 / 1.62	14.47 / 1.30
	Poisoned	10.11	1.00	14.47 / 1.29	14.30 / 1.13
	Pooled	15.84	1.00	14.79 / 1.62	14.47 / 1.30
Llama 2-7B	Clean	24.92	1.00	13.76 / 0.79	13.65 / 0.68
	Poisoned	16.12	1.00	13.20 / 0.23	13.14 / 0.17
	Pooled	20.52	1.00	13.76 / 0.79	13.65 / 0.68
Vicuna-7B	Clean	19.55	1.00	13.75 / 0.79	13.64 / 0.68
	Poisoned	14.73	1.00	13.19 / 0.22	13.13 / 0.17
	Pooled	17.14	1.00	13.75 / 0.79	13.64 / 0.68
Vicuna-13B	Clean	21.74	1.00	25.62 / 0.96	25.50 / 0.84
	Poisoned	14.99	1.00	24.91 / 0.26	24.87 / 0.21
	Pooled	18.36	1.00	25.62 / 0.96	25.50 / 0.84
Vicuna-33B	Clean	24.82	1.00	17.97 / 1.21	17.85 / 1.09
	Poisoned	17.91	1.00	17.25 / 0.50	17.21 / 0.45
	Pooled	21.36	1.00	17.97 / 1.21	17.85 / 1.09

Table S59: Behavior and retrieval in the five-model factorial. m is the number of injected passages before reranking; Poison docs is their mean count in the retrieved top-k; Retrieved is the fraction of queries retrieving at least one; Top-doc is unconditional poison localization.

Model	k	m	Target	Poison docs	Retrieved	Top-doc
Llama 3-8B	5	5	0.60	2.40	0.85	0.76
	5	1	0.30	0.54	0.54	0.41
	10	5	0.51	3.23	0.94	0.78
	10	1	0.30	0.66	0.66	0.44
Llama 2-7B	5	5	0.68	2.40	0.85	0.67
	5	1	0.28	0.54	0.54	0.31
	10	5	0.65	3.23	0.94	0.66
	10	1	0.29	0.66	0.66	0.18
Vicuna-7B	5	5	0.68	2.40	0.85	0.80
	5	1	0.30	0.54	0.54	0.41
	10	5	0.68	3.23	0.94	0.80
	10	1	0.37	0.66	0.66	0.41
Vicuna-13B	5	5	0.71	2.40	0.85	0.82
	5	1	0.26	0.54	0.54	0.42
	10	5	0.61	3.23	0.94	0.81
	10	1	0.29	0.66	0.66	0.42
Vicuna-33B	5	5	0.62	2.40	0.85	0.80
	5	1	0.22	0.54	0.54	0.47
	10	5	0.55	3.23	0.94	0.86
	10	1	0.18	0.66	0.66	0.50

Table S60: Fixed-score detection and abstention. Neut. is neutralization among attack-active queries at the leave-one-out 0.25 clean-cost target; Cost is the realized clean-correct cost.

Model	k	m	AUROC	AUPRC	Neut.	Cost
Llama 3-8B	5	5	0.77	0.71	0.72	0.24
	5	1	0.67	0.66	0.90	0.24
	10	5	0.75	0.66	0.67	0.23
	10	1	0.69	0.68	0.90	0.23
Llama 2-7B	5	5	0.69	0.62	0.43	0.24
	5	1	0.58	0.57	0.32	0.24
	10	5	0.69	0.63	0.55	0.24
	10	1	0.53	0.54	0.28	0.24
Vicuna-7B	5	5	0.78	0.71	0.72	0.24
	5	1	0.67	0.66	0.77	0.24
	10	5	0.76	0.69	0.75	0.24
	10	1	0.65	0.66	0.78	0.24
Vicuna-13B	5	5	0.83	0.77	0.86	0.24
	5	1	0.70	0.71	0.88	0.24
	10	5	0.77	0.71	0.77	0.24
	10	1	0.68	0.69	0.83	0.24
Vicuna-33B	5	5	0.81	0.74	0.87	0.23
	5	1	0.69	0.67	0.91	0.23
	10	5	0.78	0.72	0.76	0.25
	10	1	0.68	0.70	0.83	0.25

centrated partial readout neutralize 0.767–0.909 of their low-density attacked cases ($N_a = 18$–37) at realized cost 0.234–0.246. Llama 2 reaches only 0.276–0.321. Unconditional low-density Top-doc is 0.41–0.50 in the four localized models versus 0.18–0.31 in Llama 2; the corresponding poison-retrieval rates are 0.54/0.66. Thus the route retains detection signal outside saturated poisoning when concentrated partial readout is present, with scores also varying by poison density and model.

J Adaptive Dilution

The matched baseline and adaptive passage share the same query prefix and false assertion; the adaptive version appends four generic, answer-free filler sentences. Each condition uses $n = 100$, top-5 retrieval, five injected passages, seed 12, and the fixed model-specific head. Mean, sum, and share columns report neutralization at the same 0.25 clean-cost calibration. "Target output" is the probe's attacker-target string match and is not the canonical PoisonedRAG reproduction ASR.

Table S61: Matched baseline → adaptive dilution on HotpotQA. Vicuna-33B retains a 0.72 target-output rate while mean neutralization falls from 0.207 to 0.014; its sum signal increases.

Setting	Target output base→adaptive	Mean neut.	Sum neut.	Share neut.
Llama 3-8B	0.82→0.85	0.122→0.000	0.000→0.000	0.000→0.000
Llama 2-7B	0.87→0.84	0.448→0.488	0.000→0.726	0.000→0.012
Vicuna-7B	0.83→0.80	0.578→0.000	0.012→0.000	0.000→0.000
Vicuna-13B	0.76→0.77	0.592→0.013	0.171→0.169	0.000→0.000
Vicuna-33B	0.82→0.72	0.207→0.014	0.061→0.319	0.000→0.069

Table S62: Matched baseline → adaptive dilution on MS MARCO. Across the ten settings (five models × two datasets), adaptive target-output rate is 0.72–0.87 and per-token mean neutralization falls in nine. Routed attention mass remains: sum neutralization recovers for Llama 2 and Vicuna-33B and varies across the other models.

Setting	Target output base→adaptive	Mean neut.	Sum neut.	Share neut.
Llama 3-8B	0.85→0.82	0.341→0.012	0.047→0.073	0.047→0.061
Llama 2-7B	0.91→0.87	0.604→0.310	0.033→0.609	0.033→0.195
Vicuna-7B	0.84→0.86	0.821→0.012	0.155→0.070	0.119→0.058
Vicuna-13B	0.86→0.82	0.512→0.012	0.163→0.146	0.116→0.061
Vicuna-33B	0.84→0.78	0.667→0.077	0.155→0.231	0.095→0.167

In the Llama 3/HotpotQA dilution diagnostic, filler increases poison-document token count by 3.168×, lowers maximum per-token attention to 0.377×, and increases total poison attention to 1.201×; $1.201/3.168 = 0.379$, matching the observed mean ratio. The matched five-model, two-dataset grid removes the prefix confound. The attack therefore evades a normalization. Because causal interventions are not rerun here, surviving attention mass is evidence of route-state persistence, not proof that the same causal mediation remains intact.

J.1 Length-Conditioned Gate

The length-conditioned gate was prespecified before adaptive labels were examined. On matched-clean documents only, we regress log attention mass on log document-token count, robust-standardize that residual and logit attention share, and define the length-conditioned score

$$\begin{aligned} u &= \mathrm{resid}[\log \mathrm{mass} \mid \log \mathrm{tokens}], \\ v &= \mathrm{logit}(\mathrm{share}), \\ s_{\mathrm{len}} &= \big(z_{\mathrm{robust}}(u) + z_{\mathrm{robust}}(v)\big)/\sqrt{2}. \end{aligned} \tag{5}$$

Five folds use only clean documents; poison and adaptive labels are evaluation-only. The threshold is the matched-clean 75th percentile and is then fixed for the paired matched/diluted comparison.

Table S63: Paired result for the prespecified length-conditioned gate. Median paired change is -0.0197 and only 4/10 cells are nonnegative, so clean-only linear length correction does not repair adaptive dilution. HP is HotpotQA and MS is MS MARCO.

Model	Data	Joint attacked N	Matched→diluted neut.	Paired Δ
Llama 3-8B	HP	79	0.000→0.000	0.000
Llama 2-7B	HP	81	0.049→0.420	+0.370
Vicuna-7B	HP	75	0.013→0.000	−0.013
Vicuna-13B	HP	72	0.028→0.000	−0.028
Vicuna-33B	HP	68	0.059→0.132	+0.074
Llama 3-8B	MS	77	0.078→0.052	−0.026
Llama 2-7B	MS	85	0.082→0.341	+0.259
Vicuna-7B	MS	81	0.284→0.062	−0.222
Vicuna-13B	MS	78	0.231→0.051	−0.179
Vicuna-33B	MS	73	0.260→0.123	−0.137

J.2 Context-Attention Mass

Because the length-conditioned gate degrades under dilution, we test one mechanism-guided statistic: total selected-head attention mass summed over all five retrieved documents. It removes the vulnerable per-document token-count normalization rather than fitting another adaptive classifier. The statistic is exploratory; a separate prespecified NQ experiment evaluates its held-out behavior.

Table S64: Exploratory mass-aware result. Total context-attention mass improves 9/10 paired cells with median gain $+0.221$, including large gains in Llama 2 and Vicuna-33B. Thresholds still use matched-clean data only.

Model	Data	Joint attacked N	Matched→diluted neut.	Paired Δ
Llama 3-8B	HP	79	0.380→0.367	−0.013
Llama 2-7B	HP	81	0.210→1.000	+0.790
Vicuna-7B	HP	75	0.827→0.973	+0.147
Vicuna-13B	HP	72	0.361→0.569	+0.208
Vicuna-33B	HP	68	0.471→0.882	+0.412
Llama 3-8B	MS	77	0.532→0.649	+0.117
Llama 2-7B	MS	85	0.035→0.765	+0.729
Vicuna-7B	MS	81	0.481→0.716	+0.235
Vicuna-13B	MS	78	0.372→0.385	+0.013
Vicuna-33B	MS	73	0.274→0.712	+0.438

J.3 Five-Model NQ Confirmation

The held-out NQ protocol was prespecified before any routing outcomes were inspected. It fixes the five causal-suite models, the same 100 seed-12 NQ questions, Contriever top-5 retrieval, five poisoned passages, 32-token greedy decoding, and the matched/diluted construction above. The primary score is total selected-head attention mass over all five retrieved documents; fixed comparators are maximum per-document attention mean and maximum document attention mass.

For each model and metric, five folds are assigned deterministically from the seed, model, and question ID. The 75th-percentile threshold ("higher" quantile) uses only matched clean-correct rows outside the held-out fold and is transferred unchanged to matched and diluted held-out prompts. The prespecified clean-correct cost criterion is at most 0.30.

Two of the 500 clean top-5 retrieval slots have empty corpus text. Depending on the tokenizer, these correspond to an omitted span, a zero-token span, or one attributed separator token. We define an empty document to have zero document mass. The primary analysis keeps all 100 questions; a sensitivity analysis excludes the corresponding two questions from every model and arm before fold assignment. Their near-identical results in Table S65 rule out dependence on this normalization choice.

Table S65: Pooled NQ context-mass result and input-quality sensitivity.

(a) Neutralization and paired differences

Analysis	Diluted neut.	Mass−mean Δ	Diluted−matched Δ
Primary analysis	0.442	+0.375	+0.304
Exclude empty-doc IDs	0.424	+0.356	+0.295

(b) Sample sizes and clean-correct cost

Analysis	N/model	Joint att.	Clean cost
Primary analysis	100	405	0.252
Exclude empty-doc IDs	98	396	0.238

Table S66: Per-model prespecified context-mass evaluation. M/D denote matched/diluted. Four models improve under dilution; Llama 3 changes little. Document localization remains 0.97–0.98 in every diluted cell.

(a) Intervention and clean-correct cost

Model	Attacked M/D	Joint N	Mass neut. M→D	Paired Δ	Clean cost
Llama 3-8B	85/82	78	0.051→0.038	−0.013	0.269
Llama 2-7B	88/89	87	0.011→0.885	+0.874	0.283
Vicuna-7B	89/85	81	0.481→0.728	+0.247	0.250
Vicuna-13B	89/88	84	0.143→0.333	+0.190	0.232
Vicuna-33B	85/80	75	0.000→0.160	+0.160	0.230

(b) Detection and document localization

Model	D AUROC	D AUPRC	D top-poison
Llama 3-8B	0.443	0.427	0.970
Llama 2-7B	0.906	0.878	0.980
Vicuna-7B	0.850	0.802	0.970
Vicuna-13B	0.704	0.605	0.980
Vicuna-33B	0.689	0.589	0.980

Table S67: All 15 predeclared paired model–metric comparisons. Per-token attention mean degrades under dilution in every model; total context mass recovers signal in Llama 2 and Vicuna-33B and improves two additional Vicuna models.

Model	Metric	Joint N	Matched neut.	Diluted neut.	Paired Δ
Llama 3-8B	Attn. mean	78	0.128	0.000	-0.128
	Max-doc mass	78	0.000	0.000	0.000
	Context mass	78	0.051	0.038	-0.013
Llama 2-7B	Attn. mean	87	0.414	0.287	-0.126
	Max-doc mass	87	0.000	0.782	+0.782
	Context mass	87	0.011	0.885	+0.874
Vicuna-7B	Attn. mean	81	0.432	0.000	-0.432
	Max-doc mass	81	0.000	0.012	+0.012
	Context mass	81	0.481	0.728	+0.247
Vicuna-13B	Attn. mean	84	0.595	0.012	-0.583
	Max-doc mass	84	0.000	0.036	+0.036
	Context mass	84	0.143	0.333	+0.190
Vicuna-33B	Attn. mean	75	0.507	0.013	-0.493
	Max-doc mass	75	0.000	0.027	+0.027
	Context mass	75	0.000	0.160	+0.160

Table S68: Five held-out threshold ranges for every predeclared metric and model. Training uses matched clean-correct rows outside each fold only.

Model	Train N/fold	Attn.-mean τ range	Max-doc τ range	Context-mass τ range
Llama 3-8B	40–44	0.00626–0.00713	0.710–0.749	0.886–0.904
Llama 2-7B	39–47	0.00141–0.00190	0.161–0.171	0.406–0.419
Vicuna-7B	38–44	0.00673–0.00913	0.740–0.777	0.948–0.972
Vicuna-13B	43–48	0.00593–0.00611	0.643–0.704	0.903–0.909
Vicuna-33B	47–50	0.00496–0.00674	0.636–0.678	0.943–0.956

J.4 Cross-Dataset Threshold Transfer

For each model and direction, the length-conditioned gate's residualization, robust scales, and 75th-percentile threshold are fit on source matched-clean examples only and applied unchanged to the target dataset. Every clean and poisoned condition has 100 target examples; the table's attacked N is the target condition's baseline-attacked denominator. Native target calibration has clean cost 0.24 throughout. Tables S69 and S70 report the 20-cell cross-dataset transfer analysis, which is distinct from the 48-cell last-token detector transfer.

Table S69: HotpotQA→MS MARCO unchanged-threshold transfer. Target-native clean cost is 0.24; transferred clean cost ranges from 0.12 to 0.41 and neutralization is direction/model specific.

Model	Condition	Att. N	Transfer cost	Transfer neut.	Native neut.	AUROC
Llama 3-8B	Matched	85	0.16	0.047	0.094	0.428
Llama 3-8B	Diluted	82	0.16	0.024	0.049	0.356
Llama 2-7B	Matched	91	0.41	0.143	0.077	0.370
Llama 2-7B	Diluted	87	0.41	0.540	0.287	0.609
Vicuna-7B	Matched	84	0.19	0.083	0.274	0.598
Vicuna-7B	Diluted	86	0.19	0.035	0.070	0.422
Vicuna-13B	Matched	86	0.12	0.035	0.198	0.475
Vicuna-13B	Diluted	82	0.12	0.000	0.049	0.372
Vicuna-33B	Matched	84	0.23	0.214	0.310	0.533
Vicuna-33B	Diluted	78	0.23	0.167	0.154	0.467

Table S70: MS MARCO→HotpotQA unchanged-threshold transfer. The directional asymmetry shows that mass-aware aggregation and target-local clean calibration are separate requirements.

Model	Condition	Att. N	Transfer cost	Transfer neut.	Native neut.	AUROC
Llama 3-8B	Matched	82	0.31	0.037	0.000	0.362
Llama 3-8B	Diluted	85	0.31	0.000	0.000	0.200
Llama 2-7B	Matched	87	0.16	0.023	0.046	0.431
Llama 2-7B	Diluted	84	0.16	0.238	0.417	0.650
Vicuna-7B	Matched	83	0.28	0.157	0.012	0.542
Vicuna-7B	Diluted	80	0.28	0.013	0.000	0.370
Vicuna-13B	Matched	76	0.38	0.118	0.013	0.519
Vicuna-13B	Diluted	77	0.38	0.000	0.000	0.391
Vicuna-33B	Matched	82	0.26	0.085	0.073	0.584
Vicuna-33B	Diluted	72	0.26	0.111	0.139	0.561