Semantic Distractor Robustness in LLMs: Exploring the Underlying Mechanisms

Anonymous ACL submission

Abstract

Large language models drop in performance when their context contains *semantic distractors*: context topically related to the question but off the reasoning chain. Prior mechanistic work tracks a single injected distractor token, missing more diverse failures such as the model predicting other context tokens or out-of-context but topically related ones. To capture all failure modes, we instead track the model's actual top non-gold prediction and ask which components determine its ranking against gold. Failures localize to a small set of mid-to-late-layer attention heads in mid-to-late layers which we call *answer-selection heads* (AS-Head). Patching their outputs from a clean forward into a distracting one recovers 60–86% of failed predictions across four models from two families (random heads recover $\leq$ 8%), and the same heads emerge across MuSiQue, HotpotQA, and 2WikiMultiHopQA. These heads preferentially attend to the gold answer; redirecting their attention onto an entity span causes the model to transcribe that span at 22–83% (vs. $\leq$ 7.4% for random heads), showing that intervening on their attention redirects the prediction. At inference, using their attention to filter context lifts F1 by up to +35 points on samples that originally fail under distraction, competitive with retrieval-head methods.

1 Introduction

Distracting context causing performance degradation in LLMs has attracted growing attention (Niu et al., 2025; Yang et al., 2025; Lee et al., 2026; Wang et al., 2026). This problem is especially relevant in current agentic use cases, where LLMs must retrieve information from the open web or search user file systems to complete multi-step tasks, and may surface content that is topically related but irrelevant to the task. In our work, we refer to such content as *semantic distractors*: passages that share entities, relations, or surface form with the question but contribute no fact required for the correct reasoning chain. We found that semantic distractors cause more performance degradation than semantically-unrelated ones (Table 1). Understanding their underlying mechanisms is therefore essential for diagnosing this failure and designing principled interventions.

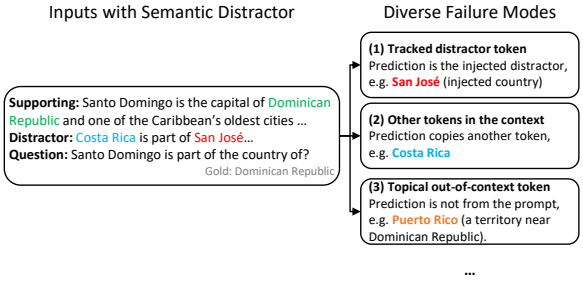

Figure 1: **Semantic distractors induce diverse failure modes.** Even with a deliberately injected distractor (left), the incorrect prediction may be (1) that tracked token, (2) a different token copied from elsewhere in the context, or (3) a topically related token absent from the prompt. This motivates a mechanistic analysis and intervention that captures all failure modes.

While distractor-induced performance drops are widely documented, prior analysis has largely been behavioral, treating the model as a black box and characterizing how accuracy varies with factors such as distractor quantity, training data, and prompting strategy (Shi et al., 2023; Yoran et al., 2023; Yang et al., 2025; Lee et al., 2026). In contrast, mechanistic study of the internal computations that produce these failures remain scarce. The closest such study, Niu et al. (2025), identifies contextual entrainment attention heads using two design choices: the probability of a deliberately injected distractor token as the localization metric for failure analysis and a differentiable mask trained to optimize this metric separately for each relation (e.g. country–capital city). However, failures under distraction are diverse: the incorrect prediction

Model	Type	Clean		Unrelated		Distracting		Distracting/Unrelated Δ Ratio	
		F1	EM	F1 (Δ)	EM (Δ)	F1 (Δ)	EM (Δ)	F1	EM
Qwen2.5-72B	Instruct	69.5	57.1	61.6 (−7.9)	49.9 (−7.2)	51.7 (−17.7)	40.2 (−17.0)	**2.26**×	**2.34**×
	Base	70.2	58.0	58.0 (−12.2)	45.1 (−12.8)	41.6 (−28.7)	30.3 (−27.6)	**2.34**×	**2.15**×
Llama-3.1-70B	Instruct	69.7	58.0	63.6 (−6.1)	51.4 (−6.7)	54.9 (−14.8)	43.9 (−14.1)	**2.42**×	**2.12**×
	Base	56.7	43.4	54.4 (−2.3)	41.2 (−2.2)	37.9 (−18.8)	25.9 (−17.5)	**8.15**×	**8.12**×
Llama-3.1-8B	Instruct	62.8	50.0	53.6 (−9.1)	41.0 (−9.0)	40.5 (−22.2)	29.0 (−21.1)	**2.43**×	**2.35**×
	Base	53.1	40.1	40.9 (−12.2)	29.3 (−10.9)	31.1 (−22.0)	20.1 (−20.0)	**1.81**×	**1.84**×
Qwen2.5-7B	Instruct	62.3	49.5	52.4 (−9.8)	41.0 (−8.5)	37.0 (−25.3)	26.0 (−23.5)	**2.57**×	**2.76**×
	Base	60.9	48.6	47.6 (−13.3)	35.5 (−13.1)	32.0 (−28.9)	21.4 (−27.2)	**2.17**×	**2.07**×

Table 1: MuSiQue validation results under three contexts: **Clean** (supporting passages only), **Unrelated** (length-matched topically-distant paragraphs), and **Distracting** (original semantic distractors). See Appendix B for details.

is not limited to the injected distractor token that they track, but also can be other token from the context, or even a topically related token outside the context. Our preliminary experiments verified that only about 20% of failures in their dataset land on the specific token they track. Two limitations therefore motivate our work. First, their metric captures only the tracked-token failure mode, leaving the others unaccounted for. Second, the heads they identify are relation-specific and must be re-identified for each new task, limiting their use for practical intervention.

To tackle the limitations, we propose a new localization metric to capture all failure modes. Rather than tracking an injected distractor token, we track the model's actual top non-gold prediction under a distracting prompt and ask which components determine its ranking against the gold token at the final position. To avoid relation-specific heads, we use a set of diagnostics that cross-validate localization: attribution patching for scalable localization, activation patching for verification, and ablation with attention analysis for characterization. We apply these methods on the diverse failures, so the identified heads are not relation-specific. We then test whether these heads serve as a general intervention target by using their attention to filter context at inference time, keeping the top-k paragraphs and re-answering on the filtered prompt.

We conducted a range of experiments and several interesting findings emerge from our experiments. We find that distractor-induced failures concentrate at a small set of attention heads in mid-to-late layers, typically 19–35 heads per model, that we call *answer-selection heads* (AS-Head). These heads preferentially attend to the gold answer over other tokens. Patching their outputs from a clean forward pass into a distracting one restores the correct answer on 60–86% of failed samples across four models, whereas matched-size random heads recover at most 8%. Redirecting their attention at the final query position onto an arbitrary entity causes the model to transcribe that span at 22–83% (versus $\leq 7.4\%$ for matched random head controls on final-answer and distractor-entity targets), demonstrating that intervening on these heads' attention can redirect what the model predicts. We find substantial overlap in the head sets identified when the procedure is repeated on HotpotQA and 2WikiMultiHopQA. Our method therefore identifies an intervention point that is generalizable across datasets.

Using the answer-selection heads' attention to filter context at inference time lifts F1 by up to $+35$ points on samples where the model fails under distraction, competitive with retrieval-head methods from the broader long-context retrieval literature (Wu et al., 2024; Zhang et al., 2025) on this subset across all four models evaluated. This provides convergent evidence that the answer-selection heads implement genuine selectivity between supporting and distracting context.

Our main contributions are as follows:

1. We propose a new localization metric and identify a small set of attention heads in mid-to-late layers at which semantic-distractor failures can be recovered by intervention.

2. We characterize these heads as implementing answer-selection under normal conditions, and show their attention pattern is causally responsible for the model's prediction.

3. We demonstrate that the answer-selection heads serve as an effective context-filtering signal at inference time, competitive with retrieval-head methods.

2 Related Work

2.1 Robustness of LLMs to distracting context

As LLMs are increasingly deployed in settings that consume long and heterogeneous context (RAG, multi-document QA, agentic tool-use), their ability to ignore irrelevant material has emerged as a central reliability concern. Most existing analyses are behavioral rather than mechanistic, characterizing how accuracy varies with factors such as the quantity of irrelevant passages, training data, or prompting strategy (Shi et al., 2023; Yoran et al., 2023; Yang et al., 2025; Lee et al., 2026). Within this line, *distracting context* and *irrelevant context* are used loosely and largely interchangeably, with some work noting that "irrelevant" is itself ambiguous (Niu et al., 2025). We keep *distracting context* as the umbrella phenomenon but make the operative axis precise: the construct we study is the *semantic distractor*. The closest prior mechanistic study to ours is Niu et al. (2025), who identify *context entrainment heads* that elevate the probability of injected distractor token. This distractor token, however, captures only a fraction of failure cases, and the heads identified by tracking it are relation-specific. We pursue a complementary approach: we track the model's actual top non-gold prediction rather than a pre-defined token, which is also more applicable to more realistic tasks where it is hard to define a specific distractor token to track.

2.2 Mechanistic analysis of specialized attention heads

A growing body of mechanistic analysis on language models has identified small sets of attention heads with specific roles in how the model processes context. One group of heads implements token- or entity-level copying: *induction heads* attend to and copy tokens that have appeared earlier in the sequence, underlying much of in-context learning (Olsson et al., 2022); in the indirect-object identification circuit, *name-mover* and *S-inhibition* heads coordinate to select the correct entity over a competing one and write it to the output (Wang et al., 2022), establishing the patching and causal-intervention methodology used in much subsequent work, including ours; and *copy-suppression heads* act in the opposite direction, reducing model confidence in over-copied tokens to improve calibration (McDougall et al., 2024). As context windows have grown, attention has shifted to heads that locate relevant content within long inputs. Wu et al. (2024) identify *retrieval heads*, a sparse subset that surfaces the answer span in needle-in-a-haystack–style tasks, and Zhang et al. (2025) extend this to query-conditioned settings with *QR-heads* selected by their attention to query-relevant paragraphs. A third line studies heads that route between parametric memory and external context: Jin et al. (2024) identify *memory* and *context* heads via path patching and prune conflicting ones to control which source the model relies on; Li et al. (2025) show these heads often superpose both roles and propose a test-time intervention to mitigate this; and Sun et al. (2025) decouple Copying Heads from Knowledge FFNs to detect RAG hallucinations from over-reliance on parametric knowledge. Our work complements these analyses by identifying *answer-selection heads* through a different localization metric. Rather than first specifying a target function (copying, retrieval) or a source dichotomy (memory vs. context) and then locating heads that implement it, we localize heads by their causal effect on distractor-induced failures, with the goal of finding stable, model-level intervention points applicable across datasets.

3 Task Formalization

We study multi-hop question answering under distracting context. Under this setting, an instance consists of a question, supporting passages containing the facts required to answer it, distractor passages with none of those facts, and a ground-truth (gold) answer token $y^\star$.

We compare two prompts: a **clean** prompt x_{clean} with only the question and supporting passages, and a **distracted** prompt x_{dist} that additionally inserts the semantic distractors. We write $\mathrm{logit}(y \mid x)$ for the pre-softmax logit on the first answer token y given x, and $\mathrm{rank}(y \mid x)$ for the rank of y under that ordering (1 = top).

For our localization metric, we run one unintervened forward pass on the distracting prompt and define y^- as the model's top non-gold prediction for the first answer token:

$$y^- = \arg\max_{y \neq y^\star} \mathrm{logit}(y \mid x_{\text{dist}}). \quad (1)$$

By construction, y^- is the token the model predicts whenever it fails on that sample. We identify y^- once and hold it fixed throughout all analyses, which investigate the components that cause the model to predict y^- rather than $y^\star$. We then categorize the samples into two subsets. The **disrupted**

subset, where distraction breaks a previously-correct prediction, has $\mathrm{rank}(y^\star \mid x_{\text{clean}}) = 1$ and $\mathrm{rank}(y^\star \mid x_{\text{dist}}) \neq 1$. The **resistant subset**, where the model is correct on both prompts, has $\mathrm{rank}(y^\star \mid x_{\text{clean}}) = \mathrm{rank}(y^\star \mid x_{\text{dist}}) = 1$.

4 Methods

Our analysis proceeds in three stages: *localize*, *characterize*, and *intervene*.

4.1 Localizing the Heads

To localize the heads causally responsible for distractor-induced failure, we need a metric that detects when the failure has occurred. We therefore target the gold–distractor logit gap on the distracting prompt,

$$L = \mathrm{logit}(y^\star \mid x_{\text{dist}}) - \mathrm{logit}(y^- \mid x_{\text{dist}}), \quad (2)$$

which is positive when the model favors gold and shrinks (typically flipping sign) under successful distraction. Using L as the metric, we run the search below.

Multi-step search. We use attribution patching (AtP; Nanda, 2023; Syed et al., 2024; Kramár et al., 2024), a fast, gradient-based approximation of activation patching (Vig et al., 2020; Meng et al., 2022), to rank candidate heads by their causal contribution to L. We apply AtP at the final token position to every individual attention head. We then re-test the top AtP candidates with full activation patching, and a head enters the answer-selection set only if both procedures agree in sign under both the forward (clean→distracting) and reverse (distracting→clean) patches. We also verify the head discovery and the patching results using 2-fold cross-validation (Appendix B).

Causal test. We test the identified heads by output patching: at the final token position, we patch their per-head outputs (taken before the output projection) from a clean run into a distracting run, and read off $\mathrm{rank}(y^\star \mid x_{\text{dist}})$ under the patched model.

4.2 Characterizing the Function of the Identified Heads

Having identified the candidate head set in §4.1, we next probe what these heads do under normal conditions. The diagnostics below cover four complementary aspects of their role: functional necessity, what spans they attend to, what their attention causally controls, and where in the layer trajectory their effect appears.

Zero and mean ablation. We zero out each answer-selection head's output contribution to the residual stream at the final token position (zero ablation), or replace it with the head's mean output over the samples (mean ablation), and measure how much $\mathrm{rank}(y^\star \mid x)$ degrades on each prompt x.

Mean per-token attention. For each head, we measure how much attention the final-token query sends to four token classes: (i) the final-answer span, (ii) the bridging-hop entity, (iii) a generic NER-tagged supporting entity, and (iv) the distractor entity (proxied by the distractor paragraph title; details in Appendix B). For each class, we sum the attention over the tokens in the span and divide by the span length to obtain a per-token rate, then report the ratio against the final-answer span. We compute these values on both clean and distracting prompts: the clean-prompt values reflect the heads' intrinsic behaviour, and the distracting-prompt values show how it changes under distraction.[1]

α-redirection. On the distracting prompt, we replace each answer-selection head's final-token attention α with uniform attention over a chosen target span, and measure the rate at which the model's top-1 prediction is a token from that span. A high rate indicates that these heads' attention pattern alone can determine what the model predicts.

Logit-lens trajectory. We apply the logit lens (nostalgebraist, 2020; Belrose et al., 2023) at the final query position across all layers, tracking the per-layer ranks of $y^\star$ and y^-. We then ablate the answer-selection heads and ask whether the layer at which gold overtakes the distractor in rank can be attributed to these heads.

4.3 Intervention via Context Filtering with the Identified Heads

Having shown that the identified heads pick out the answer-relevant span under both clean and distracting conditions (§4.2), we test whether their attention can be *read out* at inference time to mitigate distractor-induced failure. On a single forward pass over the full distracting prompt, we take each answer-selection head's post-softmax attention at the final token position. Let $\alpha_h(p)$ denote the total

[1] We normalize by span length because span lengths differ by orders of magnitude across the four models (gold-answer spans average ~ 5 tokens vs. $\sim 2{,}000$ tokens of distractor content per prompt); without per-token normalization, the comparison would conflate a head's per-token preference with raw class size. Unnormalized totals are in Appendix Table 12.

attention head h assigns to paragraph p (summed over the paragraph's tokens). Each paragraph p then receives a weight

$$w(p) \;=\; \sum_{h \in \mathcal{A}} \mathrm{ACP}_{\mathrm{fwd}}[h] \cdot \alpha_h(p), \qquad (3)$$

where $\mathcal{A}$ is the answer-selection head set and $\mathrm{ACP}_{\mathrm{fwd}}[h]$ is head h's forward activation-patching effect on L (from §4.1), so heads with larger measured causal contribution to L contribute more. We keep the top-k paragraphs by $w(p)$, re-run the model on a reduced prompt containing only those paragraphs (in original order) plus the question, and read its greedy answer. We report two settings for k. Our headline numbers use *adaptive* k, where k is set per sample to the MuSiQue-annotated number of supporting paragraphs (the question's hop count, $k \in \{2, 3, 4\}$), matching each question's natural complexity. We also evaluate a *fixed-*k variant in which a single $k \in \{1, 2, 3\}$ is used for every sample, as a check that our conclusions do not depend on the per-sample hop-count annotation (Appendix D.2). We also report an unweighted ablation: each head casts a single vote for the paragraph it attended to most, and paragraphs are ranked by vote count, matching the aggregation used for the baselines below. The intervention is training-free.

Baselines. We compare against three baseline head sets in the main results: (i) *random heads*, equally many heads randomly drawn from the answer-selection heads' own layers but disjoint from that set; (ii) *retrieval heads*, the same number of top heads by the retrieval score of Wu et al. (2024), detected on MuSiQue; and (iii) *QR-heads*, the same number of top heads by the query→paragraph attention score of Zhang et al. (2025), re-detected on a fresh 128-sample MuSiQue subset using their published procedure, calibrated against a null-query baseline. For completeness we additionally run an *all-heads* control and the per-sample *entrainment heads* of Niu et al. (2025) as sanity checks. As expected, both rank at the bottom on the disrupted subset, and we therefore do not extend them to the full MuSiQue validation (Appendix D.1).

Evaluation. We report EM/F1 on the validation set in the main results.

5 Experimental Setup

Models. We study four open-weight base language models spanning two families: `Llama-3.1-8B`, `Llama-3.1-70B`, `Qwen2.5-7B`, and `Qwen2.5-72B`.

	Llama-8B (N=276)	Qwen-7B (N=372)	Llama-70B (N=179)	Qwen-72B (N=429)
Supporting	14.1%	10.8%	19.0%	17.9%
Bridge	55.4%	32.3%	31.8%	27.7%
Distractor	30.4%	57.0%	49.2%	54.3%

Table 2: Source of the model's top-1 prediction on samples it fails under distraction (MuSiQue validation, base models). N is the count of interpretable failed samples. *Bridge* is the answer of a non-final-hop question, *Supporting* is a token from a supporting paragraph that does not match the bridge, and *Distractor* is a token from a distractor paragraph.

	Llama-3.1-8B	Qwen-2.5-7B
Correct under clean	2,761	2,551
Incorrect under distractor	251 (9.1%)	216 (8.5%)
↪ injected distractor	54 (**21.5%**)	41 (**19.0%**)
↪ other token	197 (78.5%)	175 (81.0%)

Table 3: Only about 20% incorrect predictions is the injected distractor; the rest are other topically related tokens. Examples: gold "Dominican Republic" / distractor "San José" → model says "Costa Rica" (another Latin America country); gold "China" / distractor "United States" → "Hong Kong" (related to China)

Data. Our main analysis uses the validation split of MuSiQue (Trivedi et al., 2022). We additionally evaluate generalization on HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020). We do the analysis (§6.1, §6.2, §6.3) on the disrupted subset of the validation set. For the context-filtering intervention (§6.4), we identify the heads on the training set and evaluated on the validation set, with disrupted and resistant subsets reported separately.

6 Results

6.1 Distractor-induced failures are diverse

In the MuSiQue dataset, among samples the model answers correctly on the clean prompt but incorrectly once semantic distractors are introduced, the wrong predictions come from multiple sources beyond the distractor paragraph. Table 2 shows that bridging-hop and other supporting-paragraph tokens together account for 43–70% of wrong predictions across models. In the LRE dataset of Niu et al. (2025), Table 3 confirms the same pattern: only about 20% of incorrect predictions are the injected distractor token, and the remainder are

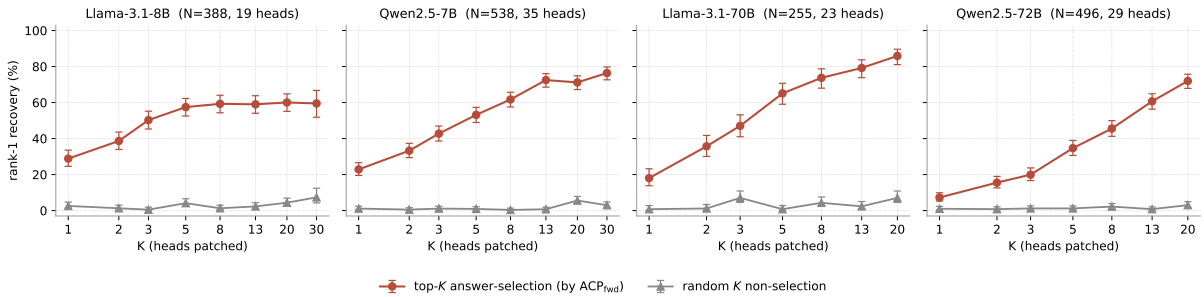

Figure 2: Cumulative head-output patching on the disrupted subset. Each point is the post-intervention rank-1 rate when the indicated head set's pre-`o_proj` outputs at the final query position are replaced with their clean-run values.

		entity targets			*non-entity* targets	
Source set	K	`final`	`bridging`	`distr_entity`	`random_supp`	`random_distr`
Llama-3.1-8B						
AS-Head	19	**62.5**%	58.7%	**60.4**%	4.5%	1.6%
Random head	19	2.9%	7.7%	2.6%	1.1%	0.5%
Qwen2.5-7B						
AS-Head	35	**22.0**%	24.1%	**42.6**%	3.8%	3.2%
Random head	35	0.4%	3.3%	1.7%	1.7%	0.9%
Llama-3.1-70B						
AS-Head	23	**83.1**%	64.2%	**78.0**%	5.5%	2.8%
Random head	23	0.0%	2.7%	0.4%	0.0%	0.0%
Qwen2.5-72B						
AS-Head	29	**68.4**%	49.3%	**38.1**%	2.5%	2.0%
Random head	29	7.4%	21.6%	3.5%	0.6%	0.4%

Table 4: α-redirection rank-1 transcription rates on the disrupted subset. Targets `final`, `bridging`, and `distr_entity` are entity-class spans (gold answer, intermediate-hop entity, and distractor entity), while `random_supp` and `random_distr` are random sub-word fragments from a supporting or distractor paragraph.

other topically related tokens. Distractor context disrupts prediction in several ways, and the model prediction does not always land on the injected distractor. A localization metric that tracks only this token therefore captures one failure mode while missing the rest. To mechanistically characterize how distractors derail prediction and to build an intervention effective across all failures, the metric should cover every failure the model exhibits.

6.2 We localize a small set of attention heads in mid-to-late layers important for distractor robustness

Using the localization procedure in §4.1 with 2-fold cross-validation, we identify a small answer-selection head set per model (19–35 heads; identical across folds, Jaccard 1.000). Cumulative patching of these heads' pre-`o_proj` outputs from a clean run into a distracting run restores $\mathrm{rank}(y^\star \mid x_{\mathrm{dist}})=1$ on 60.1–85.9% of disrupted samples across the four models, while matched-K random head subsets recover $\leq$ 8% at every K on every model (Figure 2). The answer-selection heads therefore identify a small, model-specific intervention point at which most failures can be recovered.

6.3 These heads' attention is causally important for what the model predicts

Redirecting the heads' attention causally drives the prediction. On unintervened passes, the heads' argmax-attended source token already coincides with the model's first-token prediction at a much higher rate than for random heads. To test whether this attention pattern causally drives the prediction, we replace the heads' attention at the final query position with uniform attention over a chosen entity. As Table 4 reports, the model then transcribes the chosen span at 22–83% across the four models, versus $\leq$ 7.4% for matched random head controls on final-answer and distractor-entity targets, a causal effect at least an order of magnitude larger on every model. As a further check, on the resistant subset the logit-lens gold-over-distractor rank crossover is destroyed only when

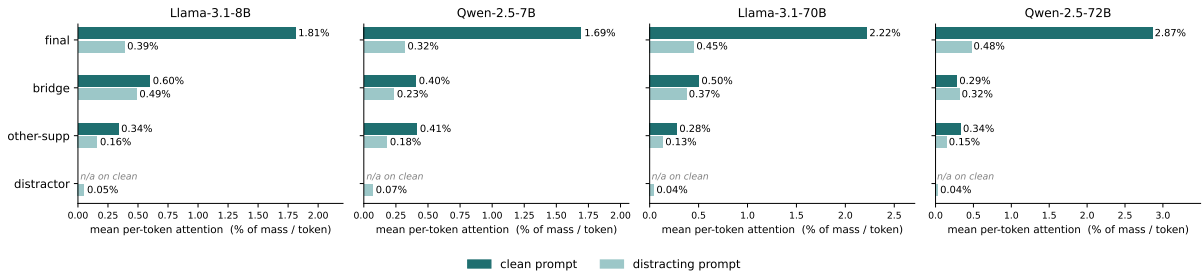

Figure 3: Mean per-token attention from the AS-Heads to each token class, on clean and distracting prompts.

Method	Disrupted EM	Disrupted F1	Resistant EM	Resistant F1	Full val EM	Full val F1
Llama-3.1-8B (N_f=2417, N_d=388, N_r=324)						
No intervention	3.6	15.6	69.1	82.1	20.1	31.1
Oracle (clean)	83.5	93.1	75.3	87.0	39.5	52.4
Random head	24.5	34.7	39.2	52.8	19.6	30.2
Retrieval head	31.4	40.7	50.9	64.3	23.0	33.5
QR-head	33.0	41.8	51.9	64.4	24.5	34.7
AS-Head (ours)	32.7	41.7	49.1	62.8	23.9	34.7
Llama-3.1-70B (N_f=2417, N_d=255, N_r=347)						
No intervention	1.6	12.9	74.9	84.4	26.1	38.2
Oracle (clean)	90.6	96.4	81.3	88.4	34.6	44.1
Random head	19.6	28.0	48.1	57.1	21.7	32.2
Retrieval head	26.7	34.0	64.3	73.2	25.9	37.4
QR-head	35.3	43.5	66.3	74.6	29.1	40.2
AS-Head (ours)	34.1	41.4	64.8	73.5	27.8	38.8

(a) Llama family

Method	Disrupted EM	Disrupted F1	Resistant EM	Resistant F1	Full val EM	Full val F1
Qwen-2.5-7B (N_f=2417, N_d=538, N_r=329)						
No intervention	3.0	11.2	68.8	79.6	21.1	31.4
Oracle (clean)	88.3	95.3	75.2	86.2	48.2	60.5
Random head	19.8	27.1	43.0	53.8	19.9	28.7
Retrieval head	25.6	33.2	51.5	62.7	24.7	34.7
QR-head	26.3	34.4	51.2	63.1	25.4	35.6
AS-Head (ours)	28.0	35.2	54.2	65.1	26.0	35.7
Qwen-2.5-72B (N_f=2417, N_d=496, N_r=605)						
No intervention	1.4	9.3	57.9	79.6	18.1	29.4
Oracle (clean)	72.0	88.4	63.6	83.9	36.0	52.5
Random head	22.2	30.5	58.3	69.0	32.4	42.3
Retrieval head	24.2	33.1	65.5	76.3	34.1	44.4
QR-head	26.8	35.9	64.3	75.2	34.9	45.5
AS-Head (ours)	24.6	32.8	64.3	74.5	35.1	45.2

(b) Qwen family

Table 5: Context filtering with head-attention voting on the full MuSiQue validation set, adaptive k = hop count. All rows use *equal-weight votes*: each head casts a single vote for the paragraph it attended to most, and paragraphs are ranked by vote count. *No intervention* runs the model on the full distracting prompt with no filtering (lower reference); *Oracle (clean)* runs the model on the supporting-paragraphs-only prompt with all distractors removed (upper reference).

these heads are ablated, and the same head-set ablation also collapses clean-prompt accuracy, so the heads are necessary for normal answer extraction rather than merely the route by which distractor content enters the prediction. Together, these results identify the heads' attention pattern as both sufficient (redirecting it changes the prediction) and necessary (ablating the heads collapses prediction) for the model's answer selection, establishing these heads as causally important for distractor robustness. See Appendix C for the detailed analysis results mentioned in this paragraph.

The heads' attention strongly prefers the gold answer over other entities. Figure 3 shows the per-token attention pattern: on clean prompts the heads' attention concentrates on the gold-answer span, several-fold above other token classes. Under distraction this concentration weakens but does not redirect to the distractor, and the per-token preference for gold over distractor is preserved at $\sim 8\times$. The heads' attention thus continues to favor the gold span even under distraction, a feature that we can use for intervention.

Summary. The diagnostics agree on the heads' function: they selectively attend to the gold-answer span over other entities, and their attention pattern is causally important for what the model predicts (α-redirection raises transcription to 22–83% versus $\leq 7.4\%$ for random heads, and ablating them cause wrong prediction). We therefore call these *answer-selection heads*.

6.4 Using the answer-selection heads as intervention recovers the failures

From § 6.3, the answer-selection heads concentrate attention on the gold-answer tokens, and the same selectivity carries over to the paragraph level even on samples the model gets wrong (Appendix C).We

exploit this signal by ranking paragraphs by per-head equal-vote count and re-prompting the model on the top-k paragraphs. On the full MuSiQue validation this lifts EM from 20.1% to 23.9% on Llama-3.1-8B and from 21.1% to 26.0% on Qwen-2.5-7B, with the lift concentrating on the disrupted subset where the intervention is needed (3.6% → 32.7% on 8B, 3.0% → 28.0% on Qwen-2.5-7B).

Comparison to baselines. From Table 5, AS-Head consistently outperforms the random head baseline by +2.7 to +6.1 EM on full validation across all four models and slightly outperforms the Retrieval head baseline by +0.9 to +1.9 EM. AS-Head is competitive with QR-Head, the heads which are identified by directly ranking heads by their query-to-context attention mass. Our AS-Heads are instead identified by their causal effect on the model's prediction and their importance for distractor robustness. AS-Head share only Jaccard 0.05–0.21 with QR-Head (Appendix D.4). That two largely-disjoint head sets selected by very different criteria reach comparable performance provides convergent evidence that the answer-selection heads we identify have a strong ability to distinguish supporting context from semantic distractors and suitable to be used as a general intervention point for distractor robustness.

Native weighting separates the methods. Each baseline ships with a per-head score (Wu retrieval score (Wu et al., 2024), QR query-focused score (Zhang et al., 2025)); we use $\mathrm{ACP}_{\mathrm{fwd}}$ in the analogous role. Re-weighting each method's vote by its own score consistently strengthens the result (Table 6). Under their native weightings, AS-Head outperforms both weighted baselines on every metric on Llama-3.1-8B and Qwen-2.5-7B.

Robustness to the per-sample hop-count annotation. With $k \in \{1, 2, 3\}$ chosen globally per method on full-validation EM (in place of the per-sample hop count), AS-Head and QR-head remain competitive on every model under equal-weight voting (Appendix D.2), so the result in Table 5 does not depend on the per-sample k annotation.

6.5 The answer-selection heads and the mechanism generalize across datasets

Repeating the localization on HotpotQA and 2WikiMultiHopQA yields head sets in the same mid-to-late layers with substantial overlap, and patching the MuSiQue-identified heads also recovers performance on these two datasets (Table 7).

	Disrupted		Resistant		Full val	
Method	EM	F1	EM	F1	EM	F1
Llama-3.1-8B						
Retrieval head	38.1	48.5	57.7	70.4	26.8	37.6
QR-head	38.1	48.3	57.1	68.6	26.8	37.4
AS-Head (ours)	40.5	50.4	58.6	70.8	27.8	38.4
Qwen-2.5-7B						
Retrieval head	29.1	36.1	53.6	65.7	25.7	35.7
QR-head	28.1	35.8	54.8	66.8	26.1	36.3
AS-Head (ours)	30.4	37.7	54.8	66.8	26.6	36.8

Table 6: Each method under its native per-head weighting on the full MuSiQue validation set, adaptive k = hop count. Each method's per-head score re-weights its vote when ranking paragraphs.

	Llama-3.1-8B	Qwen2.5-7B
MuSiQue (in-distribution)	60.1%	75.1%
HotpotQA *(transfer)*	74.7%	76.7%
2WikiMultihopQA *(transfer)*	79.1%	70.8%

Table 7: Cumulative-patching recovery rate (fraction of failed predictions restored to rank-1) of the MuSiQue-identified answer-selection heads, transferred to the disrupted subsets of HotpotQA and 2WikiMultiHopQA. The MuSiQue row matches Figure 2.

7 Conclusion

By tracking the model's actual incorrect prediction rather than a pre-defined distractor token, we localize semantic-distractor failures in multi-hop QA to a small set of attention heads in mid-to-late layers: patching their outputs from a clean run recovers 60–86% of failed predictions across MuSiQue, HotpotQA, and 2WikiMultihopQA. We characterize these as *answer-selection heads*: their attention preferentially concentrates on the gold-answer span over other entities and drives the model's prediction (redirecting it transcribes any chosen span; ablating them collapses answer extraction even without a distractor). Using their attention to filter context at inference time (without fine-tuning) lifts F1 by up to +35 points on the disrupted subset and +5–7 points on the full MuSiQue validation, matching retrieval-head methods despite a largely-disjoint head set. The answer-selection heads are thus stable, dataset-general heads that are important for semantic distractor robustness; we hope they anchor future work tracing the circuit and mechanism that configures their attention.

Limitations

Scope of analysis. Our analysis is restricted to MuSiQue-style open-form multi-hop question answering with paired clean and distractor-bearing prompts. Although we replicate the answer-selection head set on HotpotQA and 2WikiMultihopQA, we do not claim the same mechanism will appear unchanged in other reasoning tasks (e.g., mathematical or commonsense reasoning) or other distractor constructions (e.g., adversarial paraphrases)..

Model coverage. We study four open-weight base models from two families (Llama-3.1 and Qwen2.5) at two scales (7–8B and 70–72B). The answer-selection set varies in how concentrated or distributed it is across these models, so the precise circuit organization may differ in other architectures, in instruction-tuned or reasoning-tuned variants, or at substantially different scales.

Causal pathway is partial. Our localization identifies the answer-selection heads as a causally consequential intervention point for distractor robustness, recovering substantial failed predictions across four models, but we do not characterize the full causal pathway that produces the failure. We do not identify where the contaminating signal originates: upstream components could write a corrupted signal into the residual stream that the answer-selection heads then read, the heads' own Q/K computation could itself misselect under distraction, or both, and our intervention point is downstream of either possibility. Components outside the answer-selection set may also contribute to distractor-induced failures in ways our localization does not capture.

We see this work as an initial mechanistic step: reframing distractor-failure localization around the model's actual top non-gold prediction and identifying the answer-selection heads as a stable, intervention-capable locus across models and datasets. Tracing the upstream sources of the contamination, disentangling the role of the heads' own Q/K computation, and extending the framework to other reasoning settings are natural follow-ups we leave to future work.

Potential Risks

Our work is a mechanistic-interpretability contribution: we study how language models fail under semantic-distractor context and propose an inference-time context-filtering intervention. We do not introduce new model capabilities or modify model weights. The most plausible downstream risk is misuse of the context-filtering intervention as a general-purpose content filter outside the multi-hop QA setting we evaluate. The intervention may extend to other settings, but we have not validated it there; users should test it before deploying it. Our findings on open-weight base models in a research context do not transfer to safety-critical deployments without further validation.

References

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

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pages 10504–10526.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

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

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. Atp*: An efficient and scalable method for localizing llm behaviour to components. *arXiv preprint arXiv:2403.00745*.

Seongyun Lee, Yongrae Jo, Minju Seo, Moontae Lee, and Minjoon Seo. 2026. Lost in the noise: How reasoning models fail with contextual distractors. *arXiv preprint arXiv:2601.07226*.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. Taming knowledge conflicts in language models. *arXiv preprint arXiv:2503.10996*.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2024. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. *URL: https://www.neelnanda.io/mechanistic-interpretability/attribution-patching*, 15:19.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H Abdi. 2025. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16218–16239.

nostalgebraist. 2020. interpreting GPT: the logit lens. LessWrong. Accessed: 2026-05-17.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

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

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, volume 2025, pages 50250–50279.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Yanbo Wang, Zixiang Xu, Yue Huang, Gao Chujie, Siyuan Wu, Jiayi Ye, Pin-Yu Chen, Xiuying Chen, and Xiangliang Zhang. 2026. Adaptive distraction: Probing llm contextual robustness with automated tree search. *Advances in Neural Information Processing Systems*, 38:165093–165131.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. 2025. How is llm reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13340–13358.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23802–23816.

A Distractor token statistics

This appendix reports the statistics that support the distractor-definition reframing of §6.1: across both LRE and MuSiQue, the injected distractor token of Niu et al. (2025) accounts for only a subset of the model's incorrect predictions, motivating our model-behavior-based definition.

A.1 LRE: Niu et al. (2025)'s definition vs. ours

Table 8 compares two distractor-token definitions on the LRE dataset (Hernandez et al., 2024): the in-context token the distractor context is constructed to promote (Niu et al., 2025), and the model's actual top non-gold prediction (ours). Fewer than 20% of model flips land on the prior distractor, whereas our behavioral definition captures every flip by construction.

Distractor definition	Llama-3.1-8B prior	Llama-3.1-8B ours	Qwen-2.5-7B prior	Qwen-2.5-7B ours
Rank on clean (median)	292	2	256	4
In top-10 on clean (%)	4.0	87.1	2.9	75.2
Δrank when distr. added	$+261$	$+1$	$+143$	$+2$
$\Delta \log p$ when distr. added	$+3.55$	$+0.24$	$+1.05$	$+0.45$
% of failures on this token	21.5	100.0	19.0	100.0

Table 8: Two distractor-token definitions on the same LRE prompts (n=2,761 Llama, 2,551 Qwen). The **prior** definition (Niu et al., 2025) is the object the distractor context is constructed to promote, while **ours** is the model's top non-gold prediction on the distracting prompt. Δ values are medians and $\log$ p in nat.

	Llama-3.1-8B	Qwen-2.5-7B
Clean-correct subset (aggregate, no segregation)		
Distractor rank, clean (median)	4	4
Distractor in top-10, clean (%)	75.4	70.2
Δrank when distr. added	$+2$	$+2$
Top-1 flipped (%)	54.5	62.1
n (clean-correct)	712	867

Table 9: Distractor-token statistics on MuSiQue under our behavioural definition.

A.2 MuSiQue: distractor rank statistics under our definition

Table 9 reports the analogous statistics on MuSiQue. The model's top non-gold prediction y^- is in the top-10 candidates on 70–87% of clean prompts across the four models, consistent with the heterogeneity in Table 2: bridge and supporting-token failures re-rank already-strong candidates, whereas distractor-paragraph failures promote deeply-buried tokens.

B Additional Experimental Details

Table 1 Implementation Details. Table 1 compares three conditions on MuSiQue: *Clean* (only the supporting paragraphs needed to answer the question), *Distracting* (the dataset's semantic distractors added, i.e., the standard MuSiQue distracting prompt), and *Unrelated* (a length-matched control in which each distractor paragraph is replaced by a coherent, in-distribution paragraph drawn from the most topically-distant MuSiQue samples, ranked by question-embedding cosine similarity using `all-MiniLM-L6-v2`, and length-matched per paragraph within $\pm 10\%$ characters). The unrelated control holds length, English coherence, and entity density approximately constant while removing topical and entity overlap with the question.

Prompt template The prompt template is the same for all models and both conditions (clean / distracting):

```
System prompt: You're a helpful
assistant.
User prompt:
Answer the question based on the context
below. Respond with ONLY the answer,
nothing else.
Context:{context}
Question: {question}
Answer:
```

Disrupted subset filters. Beyond the clean-rank-1/distracting-not-rank-1 criterion in the main text, we restrict to samples with `content_position` $= 0$ on both conditions: the gold answer's first sub-token is the first token the model generates, with no leading whitespace sub-token in between. This keeps the analysis focused on the gold-vs-distractor decision at the first content token.

Localizing the answer-selection heads. Head identification proceeds in four stages per model on the disrupted subset. A head qualifies as an *answer-selection head* iff its causal contribution is sign-consistent under both attribution patching (the gradient-based linearization) and activation patching (the real causal swap), in both directions.

(i) Forward attribution patching, clean→distracting. For every attention head (L, H) we compute the subset-mean attribution score

$$\mathrm{AP}_{\mathrm{fwd}}(L,H) \;=\; (X_{L,H}^{\mathrm{clean}} - X_{L,H}^{\mathrm{dist}}) \cdot \left.\frac{\partial L}{\partial X}\right|_{\mathrm{dist}}, \tag{4}$$

the gradient-based linearization of activation patching at the final-token pre-W_O slice of the head's output. This is cheap enough to compute for every head.

(ii) Reverse attribution patching, distracting→clean. The same procedure with the clean↔distracting roles swapped yields $\mathrm{AP}_{\mathrm{rev}}(L,H)$.

(iii) Forward activation patching, clean→distracting. On the top-30 union of heads by $|\mathrm{AP}|$ across both directions, we then run *real* activation patching: the head's clean activation is spliced into an otherwise-distracting forward pass and the resulting shift in L is measured directly, yielding $\mathrm{ACP}_{\mathrm{fwd}}(L,H)$. This verifies the linearization on the strongest candidates rather than trusting it globally.

(iv) Reverse activation patching, distracting→clean. Same as (iii) with reversed roles, yielding $\mathrm{ACP}_{\mathrm{rev}}(L, H)$.

Membership. A head enters the answer-selection set iff

$$\mathrm{AP}_{\mathrm{fwd}} > 0 \;\wedge\; \mathrm{ACP}_{\mathrm{fwd}} > 0 \;\wedge \\ \mathrm{AP}_{\mathrm{rev}} < 0 \;\wedge\; \mathrm{ACP}_{\mathrm{rev}} < 0. \tag{5}$$

Forward scores are positive on a head whose clean-to-distracting swap recovers L; reverse scores flip sign on the same head when the contrast is swapped. Requiring the gradient-based and real-patching procedures to agree on both signs filters out heads where the linearization is noisy and heads whose causal effect is confound-driven rather than distraction-driven. The resulting answer-selection set has 19 / 35 / 23 / 29 heads on Llama-3.1-8B / Qwen-2.5-7B / Llama-3.1-70B / Qwen-2.5-72B respectively. On the top-30 union the attribution-patching linearization agrees with real activation patching at Pearson $r \geq +0.976$ with 100% sign agreement (calibration slope ≈ 1.2–1.4; attribution patching slightly over-predicts magnitudes).

Paragraph ordering. The clean prompt concatenates the supporting paragraphs in ascending `idx` order (the original MuSiQue field), separated by `"\n\n"`. The distracting prompt concatenates *all* paragraphs (supporting and distractor) in the same ascending `idx` order, so distractors are interleaved at their original positions rather than appended at the end. Each paragraph is rendered as `{title}: {paragraph_text}`. The ordering is fully deterministic and identical across all three models.

Distractor Entity Proxy In §4.2 we measure attention to "the distractor entity" as one of four comparison classes. We proxy this by the distractor paragraph's title: MuSiQue distractor paragraphs are drawn from Wikipedia, where the title names the canonical entity the paragraph describes. We compute attention to each title separately and average across all distractor titles in the prompt; multi-token titles use the same per-token-mean treatment as the other comparison spans.

C Characterizing the identified heads

C.1 Naive ablation harms performance in both clean and distracting runs

We cross-check with mean ablation (Table 11): each answer-selection head's pre-`o_proj` output is replaced with its per-condition mean over a 64-sample calibration set, instead of zero. Mean-ablated metrics are 2–5 pp closer to baseline than zero-ablation (Table 10) but every direction matches, so the qualitative finding (Finding 6.3) is not an artefact of the zero substitute.

C.2 Attention analysis

Total-mass view of where answer-selection-head attention goes. Table 12 reports total per-category attention mass at the final query position, complementing the per-token view of Figure 3. *(i)* By total mass, distractor paragraphs receive the largest content-category share (33–48%), but this is a length effect: distractor paragraphs total $\sim$2,000 tokens versus $\sim$5 for the gold span, so a uniform-attention head would already allocate $\sim 400\times$ more mass to distractors. The per-token preference for gold over distractor is preserved (Figure 3). *(ii)* Answer-selection heads attend to every content category more than the random head, so they are content-aware mass-wise and gold-selective per-token.

Mean vs. peak per-token attention across subsets. A mean per-token statistic could dilute a small number of high-attention tokens in a long span. Table 13 therefore also reports the per-token peak (maximum on any single token of class c) on both subsets. *(i)* On the resistant subset, the per-token peak on gold exceeds the peak on the distractor entity by 2.4–$2.6\times$ on both models, so the heads' preference for gold over distractor holds at the peak token, not just on average. *(ii)* On the disrupted subset the peak shifts to a distractor token, but this is a sign of failure rather than evidence against the gold preference: distractor mean and peak are essentially identical across subsets, while gold mean and peak both roughly double on the resistant subset.

C.3 Selector uniqueness within the head population

To check that the attention preference for the gold-answer span is specific to the answer-selection heads rather than a generic property of heads in this depth band, we rank all $L \cdot H$ heads in each model by the per-head selectivity score

$$s(L, H) \;=\; \overline{\alpha}_{\text{final}}(L, H) \;-\; \max_{c \in \mathcal{C}_{\text{supp}}} \overline{\alpha}_c(L, H) \tag{6}$$

on clean prompts, where $\overline{\alpha}_c(L, H)$ is the mean per-token attention from the final query position to

Model	Subset (N)	Pass	rank-1 %	EM %	F1 %	logit-gap
Llama-3.1-8B	full (2,417)	clean	$29.6 \to 7.7$	$39.5 \to 19.2$	$52.3 \to 30.1$	$+0.30 \to -1.02$
		distracting	$15.0 \to 6.7$	$20.4 \to 12.5$	$31.3 \to 23.3$	$-2.07 \to -1.78$
	disrupted (388)	clean	$98.7 \to 9.8$	$82.7 \to 11.6$	$92.4 \to 19.2$	$+2.87 \to +0.07$
		distracting	$0.8 \to 4.6$	$4.1 \to 4.4$	$16.2 \to 14.9$	$-2.14 \to -1.75$
Qwen2.5-7B	full (2,417)	clean	$38.6 \to 4.8$	$48.7 \to 15.9$	$60.8 \to 29.0$	$+1.89 \to -0.16$
		distracting	$15.5 \to 2.9$	$20.8 \to 6.2$	$30.9 \to 14.4$	$-1.88 \to -1.40$
	disrupted (538)	clean	$99.6 \to 5.4$	$87.9 \to 10.8$	$95.3 \to 26.3$	$+5.29 \to +0.59$
		distracting	$0.9 \to 1.1$	$3.4 \to 1.5$	$11.7 \to 8.1$	$-2.96 \to -1.95$

Table 10: Zero-ablating the answer-selection set at the final-query position and re-running the full prompt. Each cell reports baseline $\to$ ablated. Baseline clean rank-1 deviates from 100% because the subset is defined from the vLLM eval whereas the ablation re-runs the model in HuggingFace eager mode. Ablation collapses clean rank-1 accuracy by ≥ 85 pp on every model while leaving distracting-prompt accuracy unchanged or worsened, ruling out a distractor-only role for these heads.

Model	Subset (N)	Pass	rank-1 %	EM %	F1 %	logit-gap
Llama-3.1-8B	full (2,417)	clean	$29.6 \to 9.8$	$39.5 \to 9.6$	$52.3 \to 18.8$	$+0.30 \to -1.02$
		distracting	$15.0 \to 6.5$	$20.4 \to 4.5$	$31.3 \to 11.5$	$-2.07 \to -1.72$
	disrupted (388)	clean	$98.7 \to 13.9$	$82.7 \to 10.3$	$92.4 \to 18.8$	$+2.87 \to -0.05$
		distracting	$0.8 \to 5.7$	$4.1 \to 3.4$	$16.2 \to 10.8$	$-2.14 \to -1.67$
Qwen2.5-7B	full (2,417)	clean	$38.6 \to 10.1$	$48.7 \to 17.6$	$60.8 \to 28.8$	$+1.89 \to +0.27$
		distracting	$15.5 \to 5.0$	$20.8 \to 5.9$	$30.9 \to 13.1$	$-1.88 \to -1.26$
	disrupted (538)	clean	$99.6 \to 14.7$	$87.9 \to 12.3$	$95.3 \to 25.2$	$+5.29 \to +1.30$
		distracting	$0.9 \to 3.7$	$3.3 \to 2.4$	$11.7 \to 8.0$	$-2.96 \to -1.74$

Table 11: Mean-ablation variant of Table 10: each answer-selection head's `o_proj` input is replaced with its per-condition corpus mean (over a 64-sample calibration set) instead of zero. Each cell reports baseline $\to$ ablated.

tokens of class c (averaged across the subset) and $\mathcal{C}_{\text{supp}}$ is the set of supporting NER-entity classes (bridging-hop entity, supporting-paragraph title, supporting-paragraph NER-tagged body entity).

The top causal answer-selection heads are population outliers in attention selectivity. The top-5 heads by ACP_{fwd} have median selectivity rank in the top 0.1–2.0% of the head population on every model (Table 14). Out of the hundreds of candidate heads per model, the ones the causal analysis flags as the top answer-selection heads are precisely the heads that the attention-selectivity analysis flags as the top selectors of the gold-answer span. The gold preference is therefore a distinguishing property of these heads, not a generic property of the depth band.

C.4 Per-layer logit-lens analysis

The answer-selection heads are important for the gold-over-distractor crossover on resistant prompts. A per-layer logit-lens at the final query position gives the same picture layer-resolved (Figure 4). On the disrupted subset the distractor outranks gold and never recovers. On the resistant subset gold overtakes the distractor at the layer of the top answer-selection head (`L17` on 8B, `L23` on Qwen-7B) and holds the lead to the output, but only if the answer-selection heads are intact: zero-ablating them on the resistant subset destroys the crossover, while random head ablation preserves it.

D Intervention: context filtering results

D.1 Baselines excluded from the full-validation experiment

The full-validation context-filtering tables (Table 5, Appendix Table 15) excludes two potential baselines: an "all-heads" control and the entrainment-head set of Niu et al. (2025). We evaluated them on the smaller disrupted subset under a fixed-k sweep. Both underperform on the disrupted subset, so we did not scale them to the full validation.

All-heads control. Pooling votes across every (L, H) in the model produces diffuse paragraph scores dominated by attention-sink tokens (BOS, the answer-prefix) rather than answer-relevant content. On the disrupted subset, EM at k=2 is 22.2 on Llama-3.1-8B, 16.9 on Qwen-2.5-7B, 2.8 on Llama-3.1-70B, and 8.7 on Qwen-2.5-72B, with

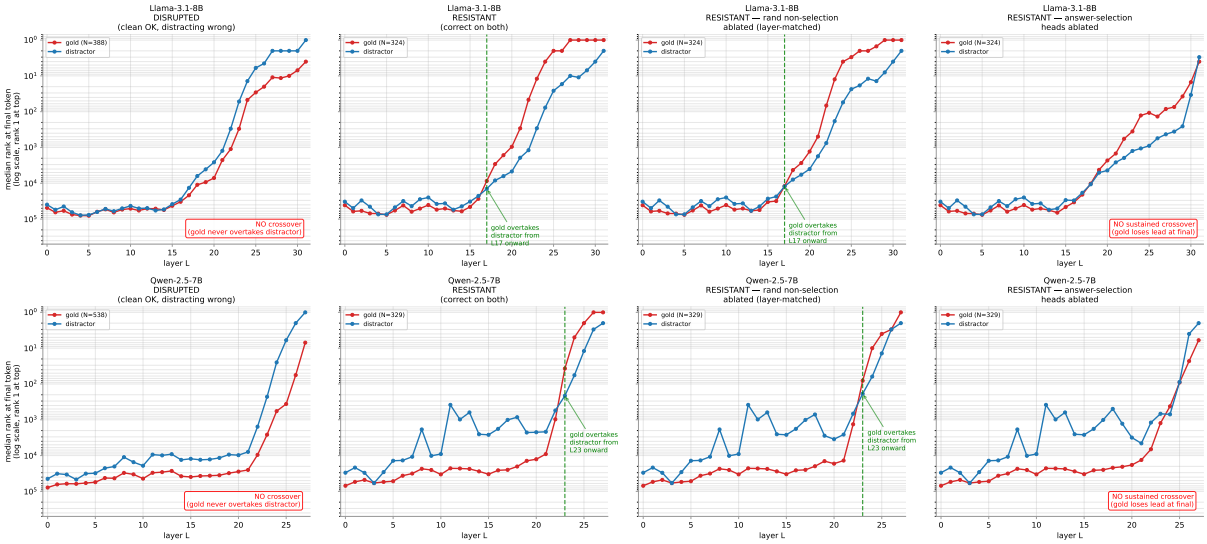

Figure 4: Logit-lens median rank of gold and distractor tokens at the final query position on the distracting pass, per layer. Rows: Llama-3.1-8B, Qwen-2.5-7B. Columns: disrupted subset, resistant subset, resistant subset with random heads ablated, resistant subset with the answer-selection heads ablated.

the gap widening on the larger models.

Entrainment heads (Niu et al., 2025). The differentiable-mask procedure learns a per-head mask that maximizes the probability of the injected distractor token. Because entrainment heads are relation-specific (Niu et al., 2025), the procedure does not converge to a consistent head set when fit jointly across MuSiQue's diverse samples, and fitting per-relation would require a relation taxonomy that MuSiQue does not provide. We therefore fit per-sample. By construction the resulting heads are those that *attend to the distractor*, not heads that attend to the answer: their top vote lands on a distractor paragraph 68–77% of the time. EM at k=2 on the disrupted subset is 19.3 on Llama-3.1-8B and 11.0 on Qwen-2.5-7B; the 70B/72B cells were not run because the per-sample mask search are resource-heavy on these larger models. The entrainment-head set is well-suited to its original use (ablation against distractor influence) but is the wrong head set for context-filtering read-out.

D.2 Fixed-k context filtering on the full validation

The headline intervention (Table 5) sets k per sample to MuSiQue's annotated hop count. To check that the comparison does not depend on this per-sample annotation, we also evaluate a fixed-k variant in which each method picks a single $k \in \{1, 2, 3\}$ and uses it for every sample (Table 15). The answer-selection heads remain competitive under this setting.

D.3 Paragraph category breakdown

The answer-selection, retrieval, and QR-head methods recover an order of magnitude more supporting-paragraph mass than All-heads / Entrainment, with the random head in between (Table 16). The same ordering holds at 70B/72B (Table 17).

D.4 Overlap between the three head-selection criteria

Table 18 reports pairwise overlap between the answer-selection (AS), QR-head, and retrieval-head sets per model. AS-Head shares at most 0.21 Jaccard with either baseline; even the QR-head / retrieval-head pair, which uses closely related selection criteria, shares at most 0.37. The three head sets are therefore largely disjoint, so the comparable EM/F1 under equal-weight voting in Table 5 reflects three distinct viable subsets of distractor-filtering heads rather than three relabelings of one underlying set.

E Cross-dataset Generalization

We re-run the full localization procedure (§4.1) independently on the disrupted subsets of MuSiQue, HotpotQA, and 2WikiMultiHopQA. Figure 5 visualizes the resulting head sets per model. The substantial overlap across datasets, including identity of the rank-1 (top AtP-scored) head on every model, indicates that the answer-selection heads are not dataset-specific artifacts of MuSiQue but a stable property of how each model handles multi-

hop reasoning under distraction.

Patching the MuSiQue-identified answer-selection heads also recovers 71–79% of failures on HotpotQA and 2WikiMultiHopQA (Table 7), confirming that the localized intervention point itself is dataset-stable. 70B / 72B transfer was not run for cost reasons.

F Datasets, Models, and Artifact Use

Datasets. We use four publicly released datasets: MuSiQue (Trivedi et al., 2022) (CC-BY-4.0, validation split), HotpotQA (Yang et al., 2018) (CC-BY-SA-4.0, validation split), 2WikiMultiHopQA (Ho et al., 2020) (Apache-2.0, validation split), and LRE (Hernandez et al., 2024) (MIT, used only in Appendix A.1). MuSiQue, HotpotQA, and 2WikiMultiHopQA are English-language multi-hop question-answer datasets constructed over Wikipedia paragraphs. LRE consists of English-language subject-relation-object prompts (e.g., *"The capital of France is"*); its factual relations are drawn from Wikidata and from the COUNTERFACT dataset (Meng et al., 2022), while its linguistic, bias, and commonsense relations were curated by the LRE authors. Our use of each dataset is non-commercial research, consistent with its intended use. The source material references public figures and historical or geographic entities; we did not collect or annotate any new personal data, and did not encounter content requiring redaction.

Models. We use four open-weight base language models: Llama-3.1-8B and Llama-3.1-70B (Llama 3.1 Community License), Qwen-2.5-7B and Qwen-2.5-72B (Apache-2.0). Our use is non-commercial mechanistic-interpretability research, consistent with each model's intended use under the respective license.

Software and Code. We use the standard ML stack: PyTorch for tensor compute and forward/backward passes, Hugging Face Transformers for model loading and hooked inference, Hugging Face Datasets for loading the four datasets, and vLLM for the EM/F1 evaluation passes. NumPy and SciPy are used for aggregation and statistical tests. Analysis code with the exact dependency details will be released upon publication.

Per-subset statistics. The MuSiQue validation set has 2,417 samples. Per-model disrupted and resistant subset sizes are reported in Table 5. HotpotQA and 2WikiMultiHopQA are used only for cross-dataset generalization (Table 7, validation splits); the disrupted-subset sizes are 779 / 995 on HotpotQA and 2,253 / 2,570 on 2WikiMultiHopQA (Llama-3.1-8B / Qwen-2.5-7B respectively). LRE prompts (Appendix A.1) are drawn from the original release across 15 relations. We restrict the LRE analysis to samples on which the model answers correctly without a distractor, yielding $n = 2{,}761$ on Llama / $n = 2{,}551$ on Qwen.

Compute resources. Model parameter counts are approximately 8.0B (Llama-3.1-8B), 70B (Llama-3.1-70B), 7.6B (Qwen-2.5-7B), and 72B (Qwen-2.5-72B). All experiments were run on a server with 8× NVIDIA A100 80GB GPUs. All compute was used for analysis (forward and backward passes through frozen pretrained models); no training or fine-tuning was performed. The most expensive components are the four-stage localization pipeline (attribution + activation patching) on the 70B/72B models and the cumulative head-output patching evaluation, which together dominate the total compute.

Decoding and inference. We use greedy decoding throughout (`temperature` $= 0$, no sampling). Hugging Face Transformers forward passes use eager attention when attention weights or gradients are captured and SDPA otherwise; models are loaded in bf16. vLLM evaluations use the default greedy configuration with the generation length capped at 64 tokens (sufficient for MuSiQue answer spans). We do not perform hyperparameter search; the only user-facing hyperparameter is the context-filtering k, set per sample to the question's hop count (Table 5) or globally per method on full-validation EM (Appendix Table 15).

Statistical reporting. Figure 2 reports Wilson 95% confidence intervals over per-sample rank-1 outcomes. Other reported numbers (EM, F1, rank statistics, attention densities) are point estimates over the relevant subset. Because all experiments use deterministic greedy decoding (`temperature` $= 0$) on frozen pretrained models without data shuffling, the reported metrics are reproducible to within reporting precision across runs on the same hardware; we therefore do not report run-to-run variance.

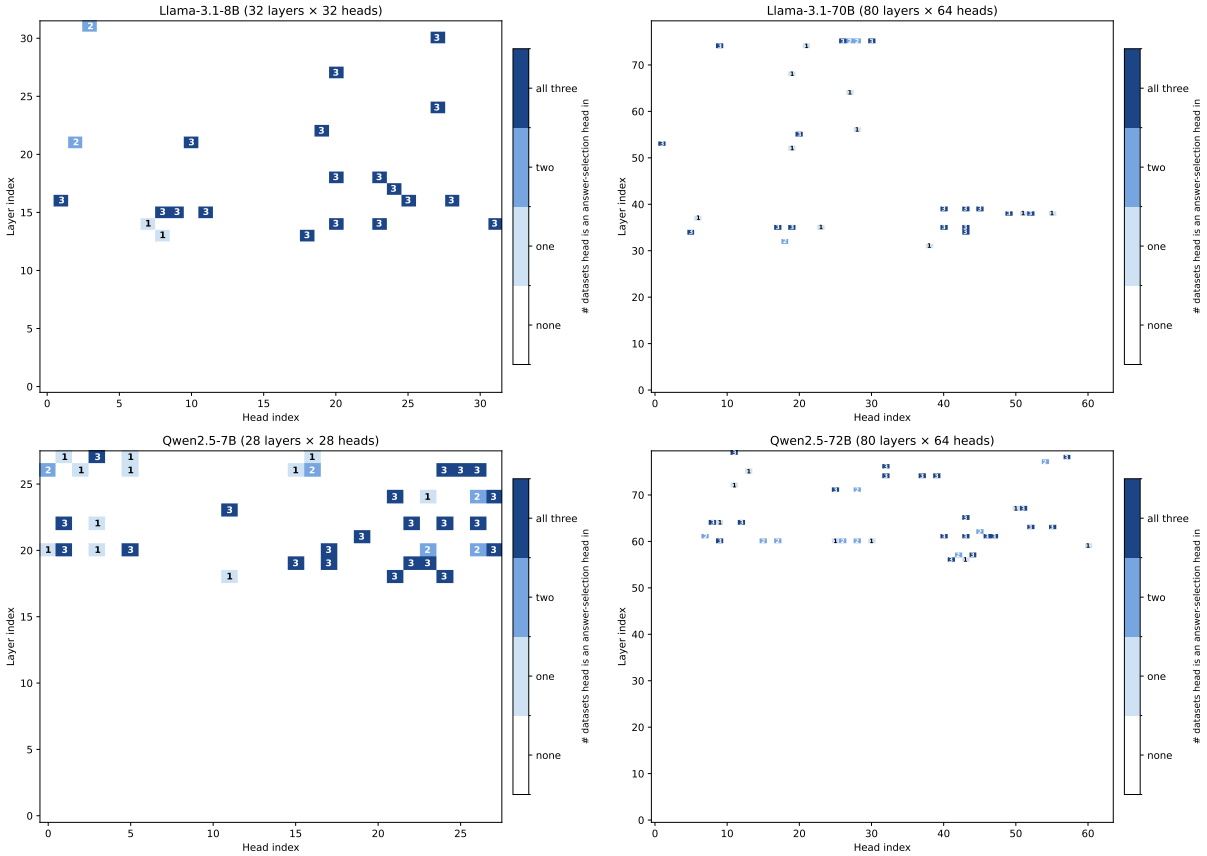

Figure 5: Answer-selection heads identified independently on three multi-hop QA datasets (MuSiQue, HotpotQA, 2WikiMultiHopQA), per model. Each colour marks the head set identified on one dataset, and overlap (darkened regions) indicates heads identified by multiple datasets.

		Final	Bridge	Other-sup	Distr.	Others
Llama-3.1-8B (N=388, $\lvert\mathcal{A}\rvert$=19)						
clean	AS-Head (weighted)	9.3	3.1	42.6	–	44.9
	AS-Head (unweighted)	6.1	2.6	35.2	–	56.1
	Random head (matched)	1.8	1.6	18.1	–	78.5
	All non-AS Heads	1.0	0.8	14.8	–	83.4
distracting	AS-Head (weighted)	2.1	3.1	20.5	33.2	41.2
	AS-Head (unweighted)	1.4	2.2	15.4	29.0	52.0
	Random head (matched)	0.4	1.0	6.3	18.3	74.1
	All non-AS Heads	0.2	0.4	4.1	16.4	78.8
Qwen2.5-7B (N=538, $\lvert\mathcal{A}\rvert$=35)						
clean	AS-Head (weighted)	9.2	2.2	49.8	–	38.9
	AS-Head (unweighted)	6.3	2.0	40.7	–	51.0
	Random head (matched)	1.8	1.1	23.3	–	73.8
	All non-AS Heads	0.7	0.5	17.2	–	81.5
distracting	AS-Head (weighted)	1.6	1.6	18.7	47.8	30.4
	AS-Head (unweighted)	1.2	1.2	14.7	43.4	39.4
	Random head (matched)	0.4	0.5	7.6	29.1	62.4
	All non-AS Heads	0.1	0.2	4.3	22.0	73.3
Llama-3.1-70B (N=255, $\lvert\mathcal{A}\rvert$=23)						
clean	AS-Head (weighted)	11.1	3.1	43.8	–	42.0
	AS-Head (unweighted)	7.1	2.4	39.1	–	51.4
	Random head (matched)	1.0	0.8	13.0	–	85.3
	All non-AS Heads	0.4	0.4	7.9	–	91.3
distracting	AS-Head (weighted)	2.1	2.6	18.8	40.9	35.6
	AS-Head (unweighted)	1.4	1.9	15.4	39.3	42.0
	Random head (matched)	0.3	0.5	4.8	17.6	76.8
	All non-AS Heads	0.1	0.2	2.3	11.5	86.0
Qwen2.5-72B (N=496, $\lvert\mathcal{A}\rvert$=29)						
clean	AS-Head (weighted)	10.8	1.4	37.8	–	50.0
	AS-Head (unweighted)	9.1	1.4	32.5	–	56.9
	Random head (matched)	2.2	1.1	21.7	–	74.9
	All non-AS Heads	0.7	0.4	11.4	–	87.5
distracting	AS-Head (weighted)	1.7	1.4	14.8	34.8	47.3
	AS-Head (unweighted)	1.5	1.3	12.4	31.7	53.2
	Random head (matched)	0.4	0.7	7.1	25.6	66.3
	All non-AS Heads	0.1	0.2	3.0	14.8	81.8

Table 12: Total per-category attention mass at the final query position, disrupted subset. Each row is a probability distribution over destination-token categories (Distractor on clean is "–" since clean prompts contain no distractor paragraphs). Head-set aggregations: *AS-Head (weighted)* averages answer-selection heads with $\mathrm{ACP}_{\mathrm{fwd}}$ weights; *AS-Head (unweighted)* averages them uniformly. The two controls are *Random head (matched)*, the mean over $1{,}000$ Monte-Carlo draws of a layer-matched random subset of heads outside $\mathcal{A}$ (each draw picks, in every layer that contains AS-Heads, the same number of non-AS heads as $\mathcal{A}$ has in that layer, uniformly without replacement), and *All non-AS Heads*, the uniform mean over every head outside $\mathcal{A}$. Answer-selection heads attend ~ 1.5–$5\times$ more to supporting and distractor content than even the matched random head control on every category and condition.

	Clean				Distracting			
	Distracted		Resistant		Distracted		Resistant	
Class	mean	peak	mean	peak	mean	peak	mean	peak
Llama-3.1-8B (N_{dist}=388, N_{res}=324, $\lvert\mathcal{A}\rvert$=19)								
final (gold)	2.81	5.89	2.81	6.21	**0.58**	**1.29**	**1.14**	**2.78**
bridging	0.72	1.80	0.55	1.50	0.70	2.06	0.51	1.68
supp_title_excl	0.22	0.95	0.21	1.00	0.14	0.67	0.15	0.75
supp_body_excl	0.20	9.11	0.19	9.01	0.10	5.36	0.11	6.51
distr_title	–	–	–	–	0.05	**1.22**	0.05	**1.09**
distr_body	–	–	–	–	0.02	3.39	0.01	2.91
Qwen-2.5-7B (N_{dist}=538, N_{res}=329, $\lvert\mathcal{A}\rvert$=35)								
final (gold)	2.48	6.31	2.51	6.20	**0.45**	**1.12**	**1.29**	**3.25**
bridging	0.44	1.39	0.41	1.39	0.29	1.13	0.26	0.92
supp_title_excl	0.25	1.33	0.26	1.51	0.12	0.65	0.11	0.78
supp_body_excl	0.21	17.18	0.18	17.48	0.08	6.98	0.10	10.30
distr_title	–	–	–	–	0.07	**1.96**	0.05	**1.37**
distr_body	–	–	–	–	0.02	8.64	0.02	6.17

Table 13: Mean vs. peak per-token attention at the answer-selection heads on the disrupted and resistant subsets. Values are percentages of the head's final-query-position attention, AP-weighted across the answer-selection-head set (same aggregation as *AS-Head (weighted)* in Table 12). *mean* averages attention over all tokens of class c, while *peak* reports the maximum on any single token of class c.

top-k by ACP_{fwd}	Llama-3.1-8B ($L{\cdot}H$=1024)	Qwen-2.5-7B ($L{\cdot}H$=784)	Llama-3.1-70B ($L{\cdot}H$=5120)	Qwen-2.5-72B ($L{\cdot}H$=5120)
1	L17H24 / **1**	L23H11 / 16	L35H19 / **2**	L64H12 / 60
2	L16H1 / 17	L24H21 / 5	L35H43 / 8	L64H8 / 54
3	L15H11 / 906	L22H24 / **1**	L35H17 / 7	L60H9 / **1**
4	L18H20 / 6	L22H1 / 46	L39H40 / 5	L56H41 / 3247
5	L18H23 / 35	L20H5 / 171	L39H43 / 3	L63H52 / 2148
median rank	**17**	**16**	**5**	**60**
median percentile	**top 1.7%**	**top 2.0%**	**top 0.1%**	**top 1.2%**

Table 14: Top-5 answer-selection heads by ACP_{fwd} and their rank within the full $L{\cdot}H$ population by attention-selectivity $s = \overline{\alpha}_{\text{final}} - \max_c \overline{\alpha}_c$ on clean prompts (max over the supporting NER-entity classes). Each cell is head / *selectivity rank*, with bold ranks marking population top-1/top-2.

	Disrupted		Resistant		Full val	
Method	EM	F1	EM	F1	EM	F1
Llama-3.1-8B						
No intervention	3.6	15.6	69.1	82.1	17.3	28.5
Oracle (clean)	83.5	93.1	75.3	87.0	38.8	50.9
QR-head	25.3	33.8	43.2	56.8	18.8	29.0
AS-Head (ours)	22.7	31.6	46.3	60.3	18.3	28.6
Qwen-2.5-7B						
No intervention	3.0	11.2	68.7	79.4	16.3	27.1
Oracle (clean)	88.5	95.5	75.4	86.2	47.0	59.1
QR-head	19.3	27.4	45.9	57.2	18.9	29.5
AS-Head (ours)	19.7	27.5	47.7	58.9	19.1	29.3
Llama-3.1-70B						
No intervention	1.6	12.9	74.9	84.4	19.4	31.8
Oracle (clean)	90.6	96.4	81.3	88.4	35.6	44.0
QR-head	17.3	21.8	46.7	53.5	15.7	22.7
AS-Head (ours)	18.0	21.8	48.1	54.9	16.1	23.1
Qwen-2.5-72B						
No intervention	1.4	9.3	57.9	79.6	22.8	36.3
Oracle (clean)	72.0	88.4	63.6	83.9	46.4	63.5
QR-head	16.7	26.1	55.7	66.2	27.2	38.2
AS-Head (ours)	18.8	27.9	56.5	66.6	28.3	39.0

(a) $k = 1$

	Disrupted		Resistant		Full val	
Method	EM	F1	EM	F1	EM	F1
Llama-3.1-8B						
No intervention	3.6	15.6	69.1	82.1	17.3	28.5
Oracle (clean)	83.5	93.1	75.3	87.0	38.8	50.9
QR-head	31.7	40.7	49.7	61.5	21.1	31.0
AS-Head (ours)	31.2	40.7	49.1	62.3	21.2	31.7
Qwen-2.5-7B						
No intervention	3.0	11.2	68.7	79.4	16.3	27.1
Oracle (clean)	88.5	95.5	75.4	86.2	47.0	59.1
QR-head	26.6	34.8	51.1	62.3	22.6	33.1
AS-Head (ours)	27.7	35.4	52.3	64.1	23.0	33.2
Llama-3.1-70B						
No intervention	1.6	12.9	74.9	84.4	19.4	31.8
Oracle (clean)	90.6	96.4	81.3	88.4	35.6	44.0
QR-head	25.5	30.5	54.2	60.1	18.9	25.4
AS-Head (ours)	25.1	30.5	52.4	59.5	18.1	25.0
Qwen-2.5-72B						
No intervention	1.4	9.3	57.9	79.6	22.8	36.3
Oracle (clean)	72.0	88.4	63.6	83.9	46.4	63.5
QR-head	27.2	35.7	63.0	73.2	33.1	43.7
AS-Head (ours)	25.0	32.7	63.1	72.4	32.6	42.7

(b) $k = 2$

	Disrupted		Resistant		Full val	
Method	EM	F1	EM	F1	EM	F1
Llama-3.1-8B						
No intervention	3.6	15.6	69.1	82.1	17.3	28.5
Oracle (clean)	83.5	93.1	75.3	87.0	38.8	50.9
QR-head	29.6	39.2	53.1	65.2	21.6	31.9
AS-Head (ours)	34.0	43.8	52.8	66.4	22.8	33.6
Qwen-2.5-7B						
No intervention	3.0	11.2	68.7	79.4	16.3	27.1
Oracle (clean)	88.5	95.5	75.4	86.2	47.0	59.1
QR-head	26.2	34.0	52.9	63.9	22.1	32.3
AS-Head (ours)	26.2	34.2	55.6	65.4	22.9	32.7
Llama-3.1-70B						
No intervention	1.6	12.9	74.9	84.4	19.4	31.8
Oracle (clean)	90.6	96.4	81.3	88.4	35.6	44.0
QR-head	25.9	32.3	57.6	64.4	21.3	29.0
AS-Head (ours)	27.5	32.2	56.8	64.2	21.0	29.0
Qwen-2.5-72B						
No intervention	1.4	9.3	57.9	79.6	22.8	36.3
Oracle (clean)	72.0	88.4	63.6	83.9	46.4	63.5
QR-head	26.2	36.1	65.3	75.9	33.5	44.9
AS-Head (ours)	24.6	32.8	66.0	75.6	33.4	44.2

(c) $k = 3$

Table 15: Fixed-k context-filtering on the **full MuSiQue validation set** for $k \in \{1, 2, 3\}$ (paragraphs kept), one subtable per k. All rows use equal-weight votes, matching the aggregation in Table 5. *No intervention* is the model on the full distracting prompt; *Oracle (clean)* is the supporting-paragraphs-only oracle.

Head set	k=1 1st	Mid	Final	Distr.	k=2 1st	Mid	Final	Distr.	k=3 1st	Mid	Final	Distr.
Llama-3.1-8B (N=388)												
All heads	7.7	1.8	10.8	79.6	14.7	2.4	16.6	66.2	14.9	2.1	15.0	68.0
Random head	27.3	4.1	23.5	45.1	23.2	3.6	18.3	54.8	19.3	3.4	17.6	59.7
Entrainment	10.7	1.6	13.4	74.3	14.1	1.9	15.9	68.1	14.3	2.0	13.8	69.9
Retrieval	35.1	4.1	**30.7**	**30.2**	23.5	3.9	22.4	50.3	17.7	4.1	17.9	60.3
QRHead	33.8	4.6	28.9	32.7	24.2	4.7	**22.6**	**48.6**	19.2	3.9	18.6	58.3
AS-Head (ours)	32.0	4.4	29.6	34.0	23.6	3.2	21.4	51.8	20.4	3.4	**19.8**	**56.4**
weighted	39.7	4.1	28.9	27.3	28.6	4.6	29.5	37.2	22.9	4.5	24.7	47.9
Qwen2.5-7B (N=538)												
All heads	5.2	1.9	5.4	87.5	9.5	2.0	10.9	77.7	10.1	2.4	11.1	76.5
Random head	13.4	1.9	17.5	67.2	12.8	2.0	14.5	70.8	12.8	2.2	13.1	71.9
Entrainment	5.4	1.9	5.4	87.3	9.1	1.9	11.8	77.2	9.3	2.1	11.9	76.7
Retrieval	22.9	2.8	23.3	51.0	18.1	2.9	18.9	60.1	15.2	2.8	15.5	66.5
QRHead	23.4	3.0	**23.8**	**49.8**	18.7	3.0	19.5	**58.8**	15.1	2.7	16.2	66.0
AS-Head (ours)	20.8	2.2	23.4	53.5	17.5	2.6	**20.9**	59.0	14.9	2.5	**16.9**	**65.7**
weighted	21.4	3.3	23.4	51.9	17.8	3.2	23.4	55.7	15.4	3.1	19.8	61.7

Table 16: Hop-role distribution of the paragraphs voted in by each head set on the disrupted subset (MuSiQue validation, fixed $k \in \{1, 2, 3\}$). Each retrieved paragraph is labelled against the question's decomposition as *1st-hop*, *Mid-hop*, *Final-hop* (the gold-answer paragraph), or *Distractor*. Cells are paragraph-share percentages and each k-block row sums to 100. Primary rows use unweighted votes; the indented *weighted* row shows AS-Head with $\mathrm{ACP}_{\mathrm{fwd}}$-weighted votes. Bold marks the highest *Final-hop* share and the lowest *Distractor* share per (k, model) among the equal-weight primary rows.

Head set	k=1 1st	Mid	Final	Distr.	k=2 1st	Mid	Final	Distr.	k=3 1st	Mid	Final	Distr.
Llama-3.1-70B (N=176)												
All heads	6.2	1.1	6.8	85.8	8.0	1.4	6.0	84.7	9.8	1.5	6.1	82.6
Random head	8.0	0.6	10.8	80.7	12.8	1.7	10.8	74.7	13.7	2.7	11.8	71.8
Retrieval	30.7	2.3	19.9	47.2	22.2	2.6	15.9	59.4	17.1	2.7	14.2	66.0
QRHead	30.7	3.4	25.0	**40.9**	22.7	4.0	21.0	**52.3**	18.6	3.9	**18.8**	**58.7**
AS-Head (ours)	25.6	3.4	**27.8**	43.2	21.9	3.1	**22.2**	52.8	18.4	3.6	17.6	60.3
weighted	30.7	2.8	24.4	42.0	23.6	3.7	24.1	48.6	18.4	4.4	20.3	57.0
Qwen2.5-72B (N=275)												
All heads	7.6	2.5	7.6	82.2	8.4	2.2	7.3	82.2	9.0	2.1	8.2	80.7
Random head	27.6	1.5	22.9	48.0	20.9	2.5	17.5	59.1	17.0	1.8	14.8	66.4
Retrieval	26.9	1.8	22.2	49.1	22.5	1.6	19.1	56.7	18.2	1.9	16.5	63.3
QRHead	35.6	1.8	18.9	**43.6**	25.1	2.6	**19.5**	**52.8**	20.7	2.3	**17.1**	**59.9**
AS-Head (ours)	25.1	2.2	**23.6**	49.1	19.6	1.8	18.7	59.8	15.6	2.1	16.3	66.1
weighted	21.8	1.8	23.3	53.1	20.4	2.0	21.1	56.5	17.5	1.8	18.8	61.9

Table 17: Companion to Table 16 on Llama-3.1-70B and Qwen2.5-72B, with the same procedure, definitions, and bolding scheme.

Model	$\lvert\mathcal{A}\rvert$	AS∩QR	AS∩Ret	QR∩Ret
Llama-3.1-8B	19	3 (0.09)	2 (0.06)	4 (0.12)
Qwen-2.5-7B	35	9 (0.15)	12 (0.21)	19 (0.37)
Llama-3.1-70B	23	5 (0.12)	2 (0.05)	5 (0.12)
Qwen-2.5-72B	29	3 (0.05)	2 (0.04)	7 (0.14)

Table 18: Pairwise overlap between the three answer-selection-style head sets per model: the answer-selection set $\mathcal{A}$ (AS), top-$|\mathcal{A}|$ QR-heads by the query-focused score of Zhang et al. (2025), and top-$|\mathcal{A}|$ retrieval heads by the copy-during-decoding score of Wu et al. (2024). Cells report intersection size and Jaccard index. The three sets are largely disjoint: AS-Head shares at most 0.21 Jaccard with either baseline.