Failure Modes in Multi-Hop QA: The Weakest Link Effect and the Recognition Bottleneck

Meiru Zhang
University of Cambridge
mz468@cam.ac.uk

Zaiqiao Meng
University of Cambridge
zm324@cam.ac.uk

Nigel Collier
University of Cambridge
nhc30@cam.ac.uk

Abstract

Despite scaling to massive context windows, Large Language Models (LLMs) struggle with multi-hop reasoning due to inherent position bias, which causes them to overlook information at certain positions. Whether these failures stem from an inability to *locate* evidence (recognition failure) or *integrate* it (synthesis failure) is unclear. We introduce Multi-Focus Attention Instruction (MFAI), a semantic probe to disentangle these mechanisms by explicitly steering attention towards selected positions. Across 5 LLMs on two multi-hop QA tasks (MuSiQue and NeoQA), we identify the "Weakest Link Effect": in our 18-document, 3-bucket setting, multi-hop reasoning performance collapses to the level of the least visible evidence, governed by absolute position rather than the linear distance between facts. While matched MFAI resolves recognition bottlenecks, improving accuracy by up to 11.49% in low-visibility positions, misleading MFAI yields divergent effects modulated by task topology: entity-centric tasks with vertical reasoning chains are vulnerable, whereas event-centric tasks with horizontal evidence structures are more resilient. Finally, we demonstrate that "thinking" models utilizing System-2 reasoning effectively locate and integrate the required information, matching gold-only baselines even in noisy, long-context settings. Supplementary experiments on 2WikiMultiHopQA, extended 3–4 hop counts, and a 32B model confirm these findings generalize across datasets, reasoning depths, and model scales.

1 Introduction

While the theoretical capacity of Large Language Models (LLMs) has expanded exponentially, with context windows scaling from 4k to millions of tokens, the *effective* utilization of this context remains fundamentally constrained (Kamradt, 2023; An et al., 2025; Hsieh et al., 2024). Previous studies have characterized this limitation as position bias due to attention failures, such as the Lost-in-the-Middle phenomenon (Liu et al., 2024b), primacy bias (Liu et al., 2024b), and recency bias (Press et al., 2021; Sun et al., 2021), where information at specific positions is systematically overlooked. This bias is often attributed to model mechanisms such as position embedding (Wang et al., 2025c), attention sinks (Xiao et al., 2024), where initial tokens monopolize attention mass, and the failure of induction heads (Olsson et al., 2022) to effectively copy information from mid-context positions.

Recent research extends position bias analysis to Multi-Hop Question Answering (MHQA), where models must synthesize disconnected evidence (Yu et al., 2025a; Huang et al., 2025; Baker et al., 2024). While prior work suggests performance degrades linearly with the distance between facts (Baker et al., 2024; Huang et al., 2025) and the number of document splits (Levy et al., 2025), we challenge this view by analyzing evidence topology. By partitioning the context into positional "buckets", we observe a step-function behavior: the between-bucket gap on MuSiQue is roughly $4\times$ larger than the within-bucket variation (mean 8.31% vs. 1.87%; up to 14.75% for Ministral-8B-Instruct), indicating that attention operates at bucket-level granularity rather than fine-grained distance. Instead, multi-hop reasoning follows a "Weakest Link Effect": performance is governed by the absolute bucket position of the least visible evidence. If a single reasoning hop falls into an under-attended region, the entire chain collapses.

Existing position bias mitigation methods are predominantly resource-intensive, relying on fine-tuning for data augmentation (Li et al., 2024c; He et al., 2024) or modifying inference compute via mechanistic modifications (Wang et al., 2025c; Chen et al., 2025). In contrast, Zhang et al. (2024a) propose Attention Instruction, a training-free intervention that uses natural language to explicitly direct the model's focus. They demonstrated that

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41823–41846

absolute indexing (e.g., "Document 1") can effectively override position bias, allowing retrieval from the typically lost middle. While this validation was limited to one-hop question answering, the ability to steer attention provides a powerful control variable. It allows us to artificially restore evidence visibility, potentially isolating the **recognition failures** (overlooked evidence) from **synthesis failures** (inability to connect facts).

Building on these observations, we introduce `Multi-Focus Attention Instruction (MFAI)`.[1] MFAI acts as a semantic probe, explicitly indexing evidence locations to simulate successful recognition within a factorial experiment (details in Section 3). By comparing performance with and without these instructions, we aim to isolate recognition failures from synthesis failures. This distinction offers insights for Retrieval-Augmented Generation (RAG) architectures (Lewis et al., 2020; Yu et al., 2024a).

By deploying MFAI on the two MHQA datasets, MuSiQue (Trivedi et al., 2022) and NeoQA (Glockner et al., 2025), we address three research questions: (1) Does linear distance between supporting facts or their absolute position within the context govern multi-hop performance? (2) Can MFAI effectively mitigate position bias in MHQA and isolate recognition errors from intrinsic synthesis limitations? (3) Does extended test-time compute (System-2 reasoning) confer robustness against misleading attention cues?

Contributions In summary, this work offers the following contributions:

- **The "Weakest Link Effect":** We provide a granular analysis of position bias in MHQA, showing that performance collapses to the level of the least visible evidence bucket, regardless of the distance between facts. This pattern generalizes across an additional dataset (2WikiMultiHopQA; Ho et al., 2020), longer reasoning chains (3–4 hops), and a 4× larger model (Qwen2.5-32B; Qwen Team and Others, 2025).
- **Mechanistic Disentanglement:** We demonstrate that steering attention with semantic instructions resolves the recognition bottleneck, restoring accuracy in low-visibility positions. This confirms that performance drops are largely recognition-based rather than reasoning-based.
- **The Differential Impact of Attention Steering:** We reveal that while matched MFAI neutralizes position bias, unmatched (misleading) MFAI yields divergent effects modulated by task topology. Misleading cues degrade entity-centric tasks with vertical reasoning chains (MuSiQue) but have limited impact on event-centric tasks with horizontal evidence structures (NeoQA).
- **System-2 Reasoning Robustness:** We show that models utilizing extended test-time compute (e.g., Qwen3-8B-Think) override context retrieval artifacts, exhibiting superior robustness to both inherent position bias and adversarial prompts.

2 Related Work

Multi-Hop Reasoning and Retrieval-Augmented Challenges. Benchmarks for multi-hop question answering evolved from entity-centric tasks like HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) to complex challenges requiring event bridging (Li et al., 2024b; Glockner et al., 2025). Despite this evolution, LLMs frequently exhibited "disconnected reasoning", arriving at answers via shortcuts rather than valid logical chains (Min et al., 2019; Li et al., 2024a; Schnitzler et al., 2024). Recent taxonomies categorized the failures of LLMs on MHQA into two distinct dimensions: evidence selection failures and synthesis failures (Xu et al., 2025). Yang et al. (2024) and Song et al. (2025) validated this split, observing that models often recalled entities successfully yet failed to utilize them for subsequent reasoning. While recent works addressed synthesis via specialized training (Wang et al., 2025b), long-context robustness remains under-explored.

RAG aims to bridge the knowledge gaps, though retrieved context often requires extensive denoising to handle conflicting evidence (Liu et al., 2024a; Wang et al., 2025a). Cuconasu et al. (2024) revealed the "noise paradox", where distractors paradoxically improved model performance. LLM-based interventions, such as generating sequential reasoning notes (Yu et al., 2024b) or re-structuring raw information (Li et al., 2025b), have shown promise in improving robustness of LLMs. We

[1] Code and data are available at https://github.com/cambridgeltl/weakest-link-effect.

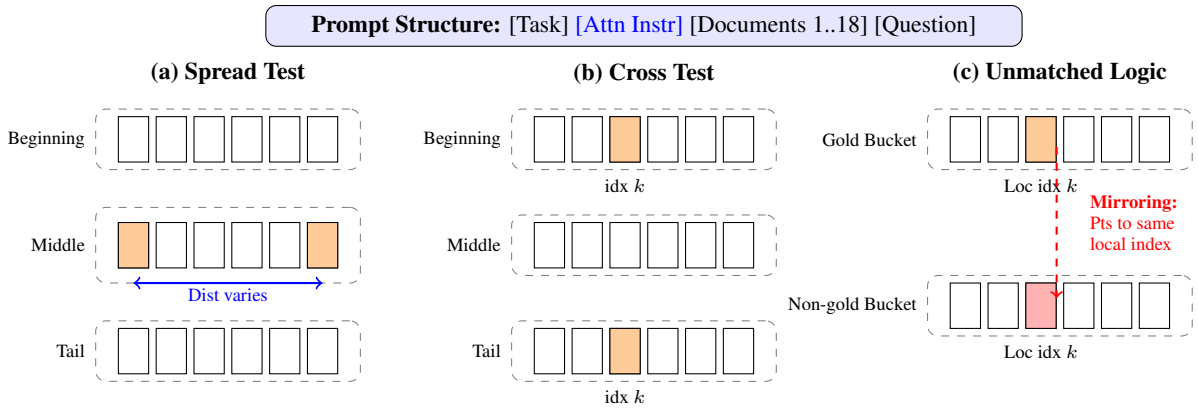


Figure 1: Experimental setups shown in vertical columns. (a) **Spread Test:** Gold documents (orange) are placed in the Middle bucket, and the distance between them varies. (b) **Cross Test:** Gold documents are split between the Beginning and Tail buckets, maintaining the same local index k. (c) **Unmatched Logic:** This illustrates how a misleading (unmatched) instruction points to non-gold documents (red) in an unselected bucket by mirroring the local index k of the true gold documents.

extend this line of work by using MFAI to mechanically isolate how evidence topology interacts with recognition and synthesis failures.

Mechanisms and Mitigation of Position Bias. Position bias (e.g., Lost-in-the-Middle, recency bias) represents a systematic failure in RAG, where models fail to utilize retrieved information located in certain positions (Liu et al., 2024b; Wang et al., 2023; Sun et al., 2021). Mechanistically, this bias is linked to attention sinks (Xiao et al., 2024) and the decay of induction head efficacy in deep context (Olsson et al., 2022). While prior studies modeled performance degradation as a linear decay relative to the distance between supporting facts (Baker et al., 2024; Huang et al., 2025), our analysis challenges this continuous view, proposing a discrete step-wise "Weakest Link Effect".

Existing mitigation strategies predominantly relied on training data augmentation (Li et al., 2024c; He et al., 2024), architectural modification (Chen et al., 2023; Zhang et al., 2024b; Wang et al., 2025c; Adiga et al., 2025; Yu et al., 2025b), or inference-time reordering (Yu et al., 2025a; Yi et al., 2025). In contrast, we adopt a training-free semantic steering approach. By building on Attention Instruction (Zhang et al., 2024a), a natural language prompt to anchor attention towards a single document, we introduce MFAI to explicitly probe LLMs and mitigate position bias.

3 Experimental Setup

We design a factorial experiment to disentangle position bias mechanisms using a fixed context of $N = 18$ documents, denoted as $\mathcal{D} = \{d_0, \ldots, d_{17}\}$. The context is partitioned into three virtual buckets of six documents each: **Beginning** (d_0 to d_5), **Middle** (d_6 to d_{11}), and **Tail** (d_{12} to d_{17}). We enforce a two-gold document set ($|\mathcal{G}| = 2$) to isolate the effects of absolute position and inter-gold-document distance.

3.1 Multi-Focus Attention Instruction

Building on the single-document attention steering (Zhang et al., 2024a), we implement MFAI to probe multi-hop scenarios. MFAI explicitly directs model focus to two documents, d_X and d_Y, using the template: "The answer is in Document X and Document Y. Use the information from Document X and Document Y as the main reference." We design three experimental conditions based on the relationship between the instructed indices $\{X, Y\}$ and the gold set $\mathcal{G}$:

- **No MFAI (NA):** The baseline condition with no MFAI, requiring the model to perform both recognition and synthesis of evidence unaided.
- **Matched MFAI:** The instruction references the true global gold document indices ($X, Y \in \mathcal{G}$), simulating successful recognition.
- **Unmatched MFAI:** The instruction deliberately points to indices in a non-gold bucket ($X, Y \notin \mathcal{G}$) to test robustness against misleading recognition signals. We select adversarial indices by **mirroring the local bucket index** of the gold documents in a non-gold bucket (Figure 1(c)), isolating cross-bucket position bias from local-index effects.

Diagnostic Scope. MFAI is designed as a diagnostic probe rather than a deployable technique: it relies on oracle knowledge of gold document positions to create controlled experimental conditions. We note that directing the model to gold documents not only steers attention but also implicitly reduces the effective search space from 18 candidates to 2 highlighted documents, potentially easing synthesis as well. The performance gain from Matched MFAI therefore provides an *upper-bound estimate* of the recognition bottleneck rather than a precise isolation of it. We use the gold-only ablation (Appendix A.6), where models receive only the two gold documents with no distractors, as a ceiling reference that quantifies the remaining gap attributable to search-space difficulty.

3.2 Research Questions

We investigate three core questions:

- **RQ1 (Topology):** Does the linear distance between distributed facts or the absolute position govern multi-hop reasoning? We address this by comparing the **Spread protocol** (distance variation within buckets) against the **Cross protocol** (split across buckets), as detailed in Section 3.3.
- **RQ2 (Mechanism):** Are failures driven by recognition deficits or intrinsic synthesis limitations? We probe this by applying **Matched MFAI** to isolate recognition failures from synthesis failures.
- **RQ3 (Robustness):** Does System-2 reasoning improve robustness against misleading signals? We test **Unmatched MFAI** (misleading cues) on standard instruction-following models vs. thinking models (e.g., Qwen3-8B-Think) to evaluate test-time verification capabilities.

3.3 Spread and Cross Protocols

To characterize reasoning failures, we employ two topological protocols (visualized in Figure 1). We evaluate both protocols under all three MFAI conditions (No MFAI, Matched, Unmatched) and report the **Unmatched** performance as the average accuracy across specific adversarial variants (detailed in Appendix A.9).

Spread Test (Distance within Bucket): As shown in Figure 1(a), we fix one gold document at the bucket start (local idx 0) and vary the second gold document within the same bucket. This tests the model's ability to synthesize information over varying inter-gold-document distances while holding the absolute bucket position constant.

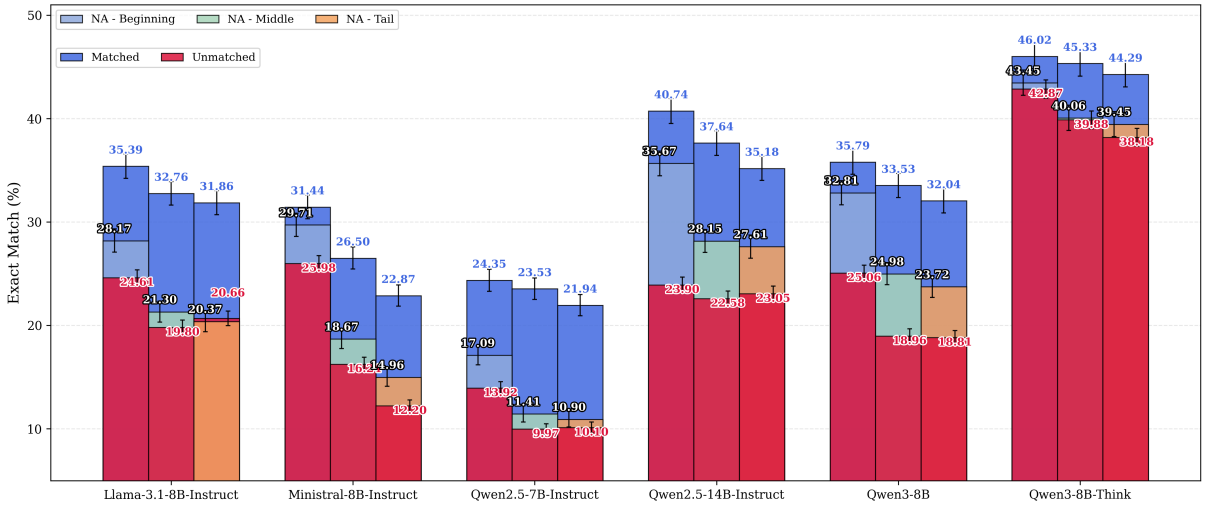
Cross Test (Split across Buckets): As shown in Figure 1(b), we place gold documents in *different* buckets (e.g., Beginning and Tail) while maintaining the same local index k. This tests the impact of crossing bucket boundaries.

3.4 Datasets

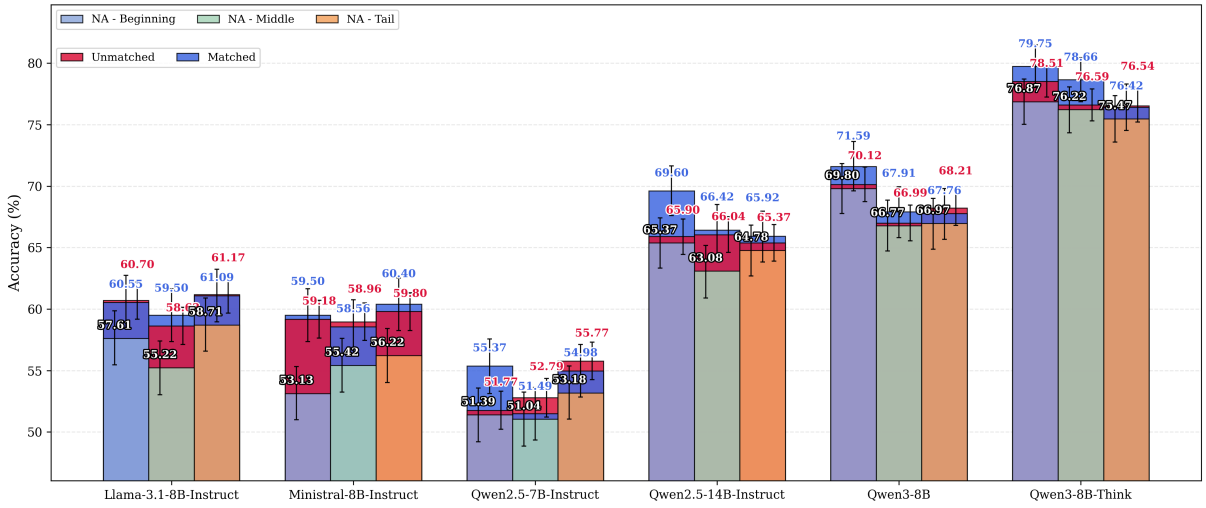
We evaluate our methods on two MHQA benchmarks representing distinct reasoning paradigms: MuSiQue (Trivedi et al., 2022) and NeoQA (Glockner et al., 2025) (examples in Appendix A.3). MuSiQue is an entity-based benchmark with questions derived from real-world Wikipedia documents, requiring reasoning to bridge connections between named entities. It serves as a proxy for standard RAG tasks. In contrast, NeoQA is a fully synthesized dataset constructed from fictional event timelines. It demands complex reasoning over entities involved in multiple sequential events, effectively eliminating parametric knowledge shortcuts. We filter for questions requiring exactly two gold documents and use a fixed context size of 18 documents (2 gold, 16 distractors). The final evaluation set comprises 1,246 MuSiQue examples (evaluated via Exact Match (EM)) and 402 NeoQA examples (evaluated via Accuracy (Acc.)). We further validate on the Compositional and Inference subsets of 2WikiMultiHopQA (Ho et al., 2020), MuSiQue 3-hop and 4-hop subsets, and NeoQA with alternative distractor settings (Appendix A.7).

3.5 Models

We evaluate five state-of-the-art Large Language Models: Qwen2.5-7B-Instruct (Qwen Team and Others, 2025), Qwen2.5-14B-Instruct (Qwen Team and Others, 2025), Llama-3.1-8B-Instruct (Meta Llama Team, 2024), Ministral-8B-Instruct (Mistral AI, 2024), and Qwen3-8B (Yang et al., 2025). Qwen3-8B is assessed in both thinking (triggered via `<think>`) and non-thinking (triggered via `</no_think>`) modes to determine the impact of test-time compute on robustness against position bias and misleading MFAI cues. All models are evaluated at a temperature of 0.0 for reproducibility, consistent with prior work (Levy et al., 2025). We additionally evaluate Qwen2.5-32B-Instruct-



(a) MuSiQue dataset



(b) NeoQA dataset

Figure 2: Model performance across three MFAI conditions (No MFAI, Matched, Unmatched) on MuSiQue and NeoQA in the Spread Test. Bars show mean accuracy (%) for No MFAI (bucket-colored), Matched (dark blue), and Unmatched (red), pooled across five distances within each bucket. Error bars are 95% bootstrap CIs. Unmatched averages over adversarial variants (Appendix A.9).

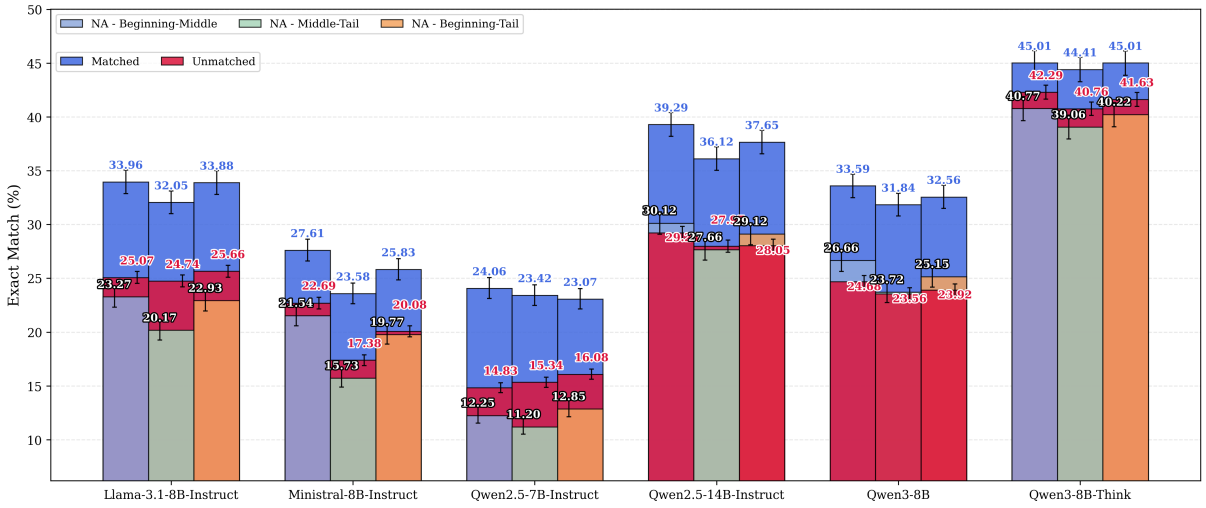
GPTQ-Int8 to assess scalability (Appendix A.7).

4 Results and Discussion

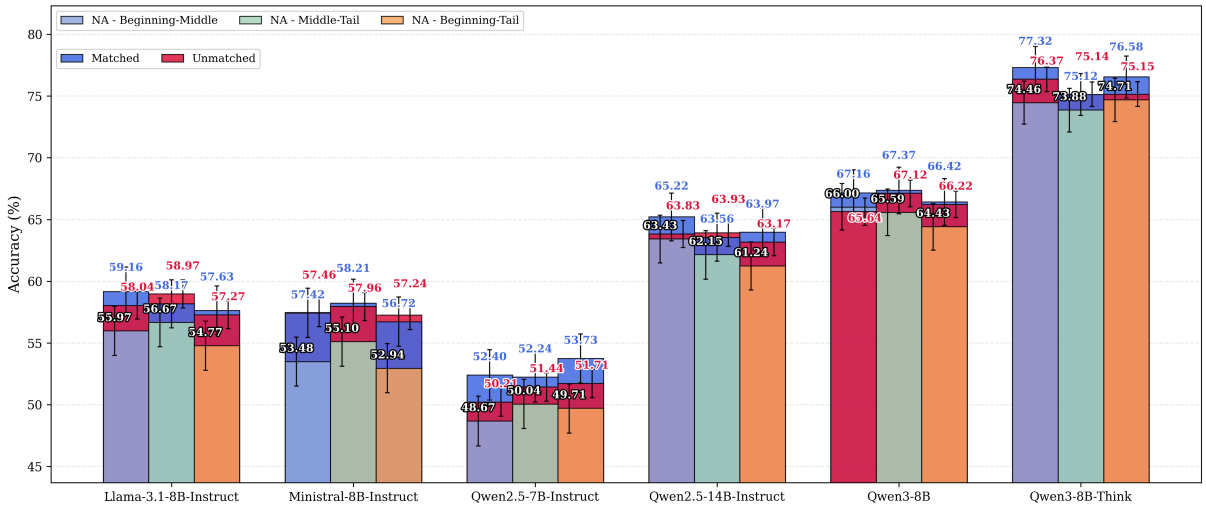
4.1 Evidence Topology: Absolute Position Governs Performance (RQ1)

We first investigate whether multi-hop reasoning performance is determined by the linear distance between supporting facts or their absolute position (RQ1). Focusing on the baseline (No MFAI) results, we observe that position bias is dataset-sensitive: the severity of position bias is much larger on MuSiQue as compared to NeoQA and has different patterns. For example, Figure 2 shows that Ministral-8B-Instruct experiences significant recency bias on MuSiQue but less obvious primacy bias on NeoQA.

Performance follows a Step-Function, not Linear Decay. The Spread Test results (Figure 4) demonstrate that reasoning performance is largely independent of the distance between evidence documents. Within any fixed bucket (e.g., Beginning), varying the distance between two gold documents from 1 to 5 results in negligible performance variance (typically $\pm 3\%$). For instance, Qwen2.5-14B-Instruct maintains $\sim 28\%$ in the Middle bucket regardless of spread on MuSiQue. However, shifting the entire evidence set from a high-visibility zone (Beginning) to a low-visibility zone (Tail) causes significant drops (e.g., Ministral-8B-Instruct drops by 14.75% on MuSiQue). This confirms that at-



(a) MuSiQue dataset

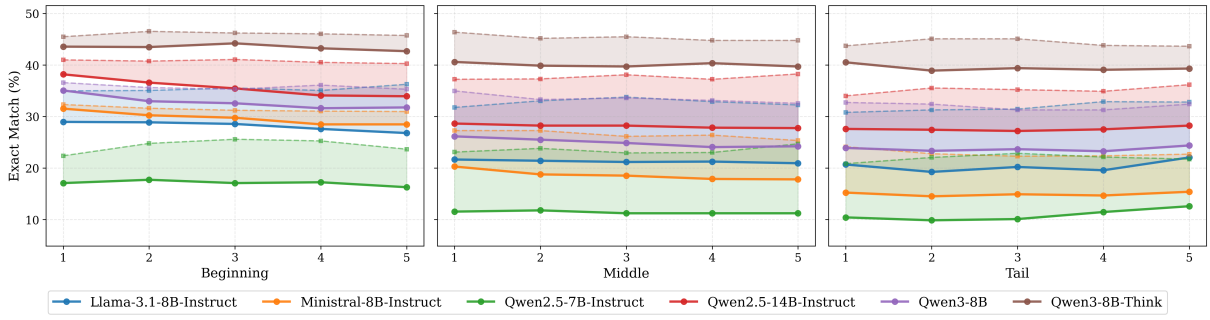


(b) NeoQA dataset

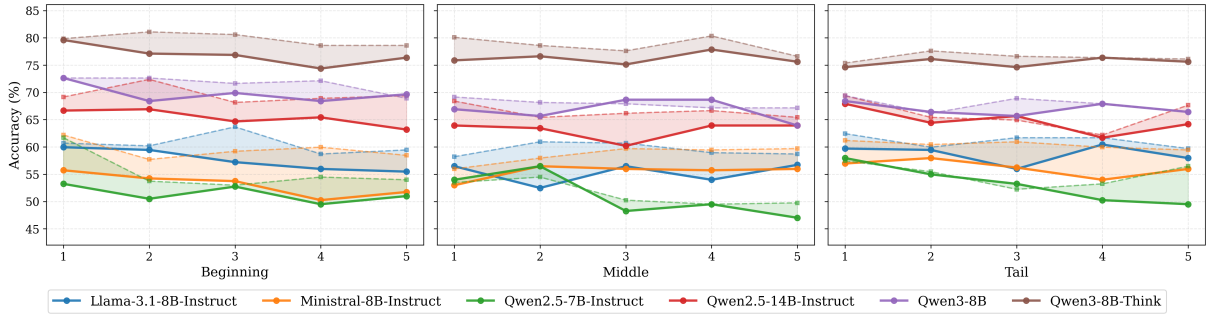
Figure 3: Model performance across three MFAI conditions (No MFAI, Matched, Unmatched) on MuSiQue and NeoQA in the Cross Test. Bars show mean accuracy (%) for No MFAI (pair-colored), Matched (dark blue), and Unmatched (red), pooled across six local indices within each bucket pair. Error bars are 95% bootstrap CIs. Unmatched averages over adversarial variants (Appendix A.9).

tention functions as a step-function governed by coarse-grained bucket location. The Cross Test (Figure 3) shows much smaller variation across the three bucket-pairs (B+M, B+T, M+T) than across single buckets: for example, on MuSiQue, Qwen2.5-14B-Instruct varies by only 2.46% across the bucket-pairs (30.12%, 29.12%, 27.66%), compared to 8.06% across the corresponding single buckets (35.67%, 28.15%, 27.61%). Furthermore, Figure 6 shows that performance remains flat across local indices (0 to 5) as long as the bucket-pair is constant, indicating that the model's attentional focus is governed by bucket membership rather than fine-grained offset.

The "Weakest Link Effect" of Multi-Hop Reasoning. The Cross Test (Figure 3) further reveals the fragility of multi-hop reasoning. When evidence is split between a high-visibility bucket and a low-visibility bucket, performance does not average out; it collapses toward the weaker bucket. For instance, Ministral-8B-Instruct on MuSiQue achieves 29.71% when both gold documents are in the Beginning bucket and 18.67% when both are in the Middle; with gold split across Beginning+Middle (B+M), accuracy drops to 21.54% — below the 24.19% naive average and close to the weaker Middle rate. We term this the **Weakest Link Effect**: the probability of successful reasoning is bounded by the minimum recognition

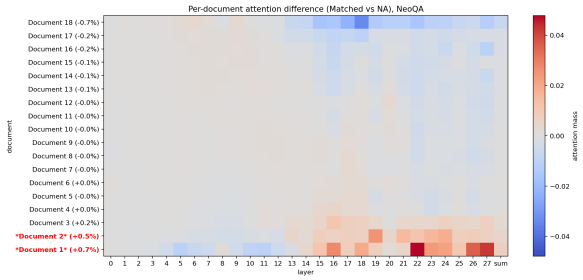


(a) MuSiQue dataset

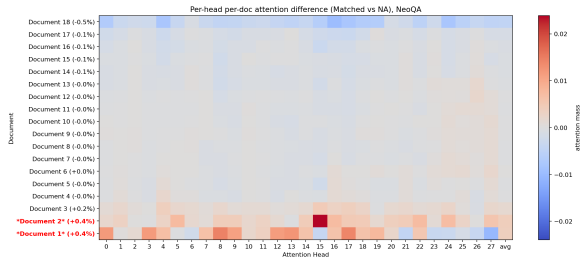


(b) NeoQA dataset

Figure 4: Model performance across various inter-gold-document distances in each positional bucket. The solid line represents the No MFAI accuracy, whereas the dashed line represents the matched instruction accuracy.



(a) Layer-wise Heatmap.



(b) Head-wise Heatmap.

Figure 5: Attention mass difference (Matched − No MFAI) of Qwen2.5-7B-Instruct on NeoQA when gold documents are placed in the Beginning bucket (Doc 1 and Doc 2), averaged over 100 examples. The y-axis lists document spans; the x-axis is layer index in (a) and head index in (b). The color bar on the right shows the percentage-point change in normalized attention mass per token. Blue cells indicate a drop; red indicates an increase.

probability of any supporting fact. If the model cannot robustly attend to the second hop, the entire reasoning chain breaks, regardless of how salient the first hop was. These step-function and Weakest Link patterns replicate on 2WikiMultiHopQA (Table 5, Table 7, Table 12), extend to MuSiQue 3-/4-hop (Table 8, Table 11), and persist at 32B scale (Table 9); full results in Appendix A.7.

4.2 The Nature of Failure: Recognition as the Bottleneck (RQ2)

Having established where failures occur (position bias), we examine why: are these drops driven by recognition deficits or synthesis limitations?

Matched MFAI Restores Performance by Resolving Recognition Deficits. We probe this distinction by applying Matched MFAI, which explicitly indexes the gold documents $\mathcal{G}$. For MuSiQue, as shown in Figure 4(a) and Figure 6(a), matched MFAI (dashed lines) consistently boosts the model performance across all inter-gold-document distances and local indices. It elevates performance in the Tail bucket by 4.83% to 11.49% across models on MuSiQue, effectively bridging the gap to Beginning-bucket performance (Figure 2(a)). In

addition, the performance gap between buckets is significantly lower than the baseline (No MFAI, white text) in Figure 2(a). This provides evidence for the **recognition bottleneck** hypothesis: models possess the inherent ability to synthesize, but this capacity is bottlenecked by attentional failures. NeoQA behaves differently: because the baseline position bias is limited, Matched MFAI provides smaller gains, primarily boosting the Beginning bucket on Spread Test (Figure 2(b)) but shows nearly equal boosting on Cross Test (Figure 3(b)). This suggests that when models can already effectively reason without bias, Matched MFAI acts as a confirmation signal and further boosts the performance. The Matched MFAI rescue also holds on 2WikiMultiHopQA (Table 6) and at larger scale on Qwen2.5-32B-Instruct-GPTQ-Int8 (Table 10, Table 13).

Redistributing Attention Mass. This recovery of model performance is driven by an observable shift in internal attention. As shown in the attention heatmaps (Figure 5), averaged over 100 NeoQA examples where Qwen2.5-7B-Instruct produced a correct answer under Matched MFAI (priority-sampled toward cases in which MFAI causally rescued a No-MFAI failure; see Appendix A.4), we observe that attention mass shifts toward the gold documents among all documents (red cells). This increase is uniform across heads but concentrated in deep layers. The same reallocation pattern replicates on MuSiQue (Figure 7), confirming that the effect generalizes across multi-hop benchmarks. This redistribution confirms that the failure mode is a recognition bottleneck driven by a lack of attention. The fact that Matched MFAI restores performance implies that the "compositionality gap" (Press et al., 2023) is frequently an attention allocation failure rather than a reasoning deficit. McNemar's paired tests (McNemar, 1947) confirm that this improvement is statistically significant: across all 6 model configurations, 2 datasets, 2 protocols, and 3 positional buckets, 62 of 72 comparisons reach $p < 0.05$ and 56 reach $p < 0.01$ (Tables 1 and 2).

4.3 Robustness and Verification: Unmatched MFAI and System-2 Reasoning (RQ3)

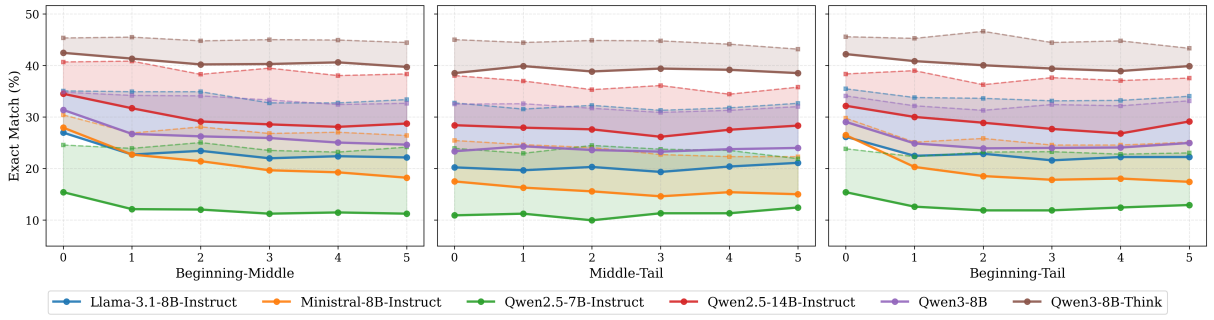
Finally, we investigate model robustness against misleading signals and the role of test-time compute (thinking mode) (RQ3).

Divergent Effects of Unmatched Instructions: Task Topology as a Moderating Factor. We observe a distinct difference in how models handle Unmatched MFAI. On MuSiQue, unmatched instructions degrade performance and, for the most instruction-sensitive model Llama-3.1-8B-Instruct, trigger a $\sim 3\times$ increase in response length on the Spread Test ($\sim$331 to $\sim$1,052 characters), signaling confusion. Conversely, on NeoQA, unmatched instructions maintain or even slightly improve accuracy. We trace this divergence to the underlying *task topology*. MuSiQue forms a strictly *vertical* causal chain: extracting Entity A from Document 1 is the prerequisite to searching for Entity B in Document 2; if attention shifts away from Document 1, the chain breaks entirely. NeoQA, by contrast, is a *horizontal* parallel task in which independent evidence segments (e.g., two dates) can be retrieved separately; even if locally misled, the model can still scan the context to recover both facts. In addition, NeoQA's original distractors come from the same fictional timeline, creating high semantic redundancy that masks positional neglect. When we replaced these same-timeline distractors with documents from random timelines, a clear primacy bias emerged: for Qwen2.5-7B-Instruct, accuracy dropped by 3.83% from Beginning (62.29%) to Tail (58.46%). Matched MFAI reduced this gap to 0.65%, demonstrating that the attention steering mechanism neutralizes positional bias even in horizontal tasks (Table 14). Detailed analysis of specific unmatched variants (Appendix A.9) further shows that partially correct instructions (one gold index provided) outperform random non-gold indices, confirming that even imperfect cues can partially resolve the recognition bottleneck.

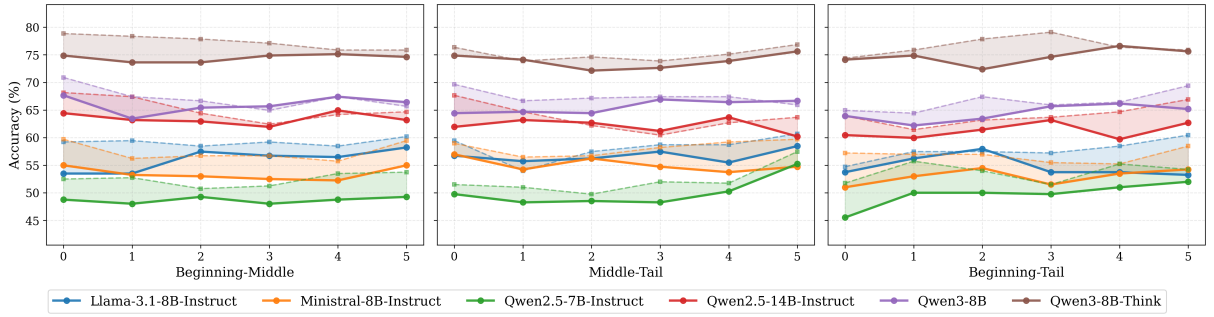
System-2 Reasoning Overrides the Recognition Artifact. Comparing standard instruction-following models with System-2 reasoning models (e.g., Qwen3-8B-Think) reveals that extended test-time compute ($\sim 6\times$ more output tokens than its non-thinking counterpart on NeoQA; see Table 15) functions as a verification filter. While standard models passively follow unmatched instructions, the thinking model maintains high accuracy with low variance.

4.4 Ablation Analysis: The Cost of Noise

Comparing full-context performance to a gold-only baseline (Table 3 in Appendix A.6), standard models (e.g., Llama-3.1-8B) suffer a significant penalty



(a) MuSiQue dataset



(b) NeoQA dataset

Figure 6: Model performance across various local indices within the selected pair of positional buckets in the cross test. The solid line represents the No MFAI accuracy, dashed line represents the matched instruction accuracy.

in the 18-document setting (e.g., dropping from 40.45% to avg. 23.28% on MuSiQue), confirming that noise filtering is a primary bottleneck for instruction-following models. In contrast, Qwen3-8B-Think defies this trend, matching or exceeding its gold-only performance even with noise. This suggests that distractors may paradoxically aid thinking models by triggering more rigorous verification. The recognition bottleneck and its rescue by Matched MFAI persist at larger scale on Qwen2.5-32B-Instruct-GPTQ-Int8 (Table 10, Table 13). A no-document ablation (Table 4 in Appendix A.6) confirms that performance is driven by the provided context rather than parametric memory, as accuracy drops to near-zero when documents are removed.

5 Conclusion

We analyzed position bias in long-context multi-hop reasoning, distinguishing between recognition and synthesis deficits. We identified the "Weakest Link Effect": multi-hop performance is constrained by the absolute position of the least visible evidence and driven primarily by a recognition bottleneck. This effect replicates across datasets (MuSiQue, NeoQA, 2WikiMultiHopQA), hop counts (2-, 3-, 4-hop), and model scales (7B–32B). Task topology modulates the impact: vertical reasoning chains (MuSiQue) are more vulnerable to misleading attention cues than horizontal tasks (NeoQA), and context homogeneity can mask the underlying bias. Importantly, misleading instructions exhibit a graded effect: partially correct cues (one gold index) still improve over unguided baselines, while fully random cues harm performance, reinforcing that even imperfect recognition signals can partially break the weakest-link bottleneck.

These findings point to several actionable directions. (1) Since absolute position governs performance, RAG reranking should prioritize high-visibility placement of critical evidence; more broadly, training objectives should target recognition under noisy retrieval, not solely reasoning capabilities. (2) Developing lightweight mechanistic probes (e.g., attention patterns, logprob signals) would make position bias diagnosis practical without the extensive factorial evaluation our prompt-level approach requires. (3) Thinking models override position bias at substantial cost ($\sim 6\times$ output tokens; Table 15); distilling this self-verification into standard inference remains open. (4) Whether other reasoning structures (e.g., comparative, temporal) exhibit distinct vulnerability profiles under the vertical–horizontal topology is an open question for task-specific mitigation.

Limitations

Our core experiments cover two datasets (MuSiQue, NeoQA) and five models; supplementary evaluations extend to 2WikiMultiHopQA, 3-/4-hop subsets, and Qwen2.5-32B (Appendix A.7), but these use representative models rather than the full suite, and frontier-scale models (70B+) remain untested. We did not perform a mechanistic analysis using logprobs or perplexity, nor explore prompt variations, different bucket counts, or context sizes beyond 18 documents. A pilot study on MFAI index ordering (Appendix A.8) shows negligible sensitivity (median 2.52% relative change), but was tested on only two models. MuSiQue uses open-ended generation while NeoQA is multiple-choice; the answer options in NeoQA may act as lexical anchors that partially mask position bias independently of task topology, so our attribution of the MuSiQue–NeoQA divergence to topology may be confounded by this format difference. We also report strict Exact Match (EM) for open-ended generation; softer alternatives such as token-level F1 or Contains (normalized substring match) would raise absolute accuracy on MuSiQue by roughly 8–15 percentage points, since they credit sentence-form responses that overlap with (F1) or contain (Contains) the gold answer rather than requiring exact-string equality, though we expect the relative patterns we report to be largely preserved. Finally, we used a fixed distractor order to isolate position bias from semantic similarity effects; real-world scenarios with retrieval rerankers remain future work.

Ethics Statement

We affirm that this research follows the ACL Code of Ethics. Our study uses publicly available models and datasets (MuSiQue (Trivedi et al., 2022) and NeoQA (Glockner et al., 2025)) and does not involve human subjects or private information. By establishing the "Weakest Link Effect" to explain position bias in LLMs in multi-hop question answering, our research contributes to the development of more transparent and reliable reasoning systems. We acknowledge the societal risks associated with reasoning LLMs. Our experiments operate within the existing paradigm of multi-hop reasoning and retrieval-augmented generation. We do not introduce novel risks through our experiments. Code and data are publicly available.[2]

We are committed to the full reproducibility of this study. The complete source code and all curated data artifacts are released under a permissive open-source license. Key implementation details and hyperparameters are described in Appendix A.1.

Disclosure of AI Use: We acknowledge the use of AI assistants for linguistic polishing and grammatical and stylistic revision during the preparation of this manuscript. While these tools were employed to enhance clarity and readability, the authors conducted all experiments and verified all AI-assisted coding. All experimental design, data analysis, scientific reasoning, and final conclusions were produced by the authors, who maintain full responsibility for the content and scientific accuracy of the final paper.

Acknowledgments

This work was supported by the Gates Cambridge Scholarship.

References

Rishabh Adiga, Besmira Nushi, and Varun Chandrasekaran. 2025. Attention speaks volumes: Localizing and mitigating bias in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26403–26423.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2025. Why does the effective context length of llms fall short? In *The Thirteenth International Conference on Learning Representations*.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. 2024. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa. *arXiv preprint arXiv:2412.10079*.

Jiabei Chen, Guang Liu, Shizhu He, Kun Luo, Yao Xu, Jun Zhao, and Kang Liu. 2025. Search-in-context: Efficient multi-hop qa over long contexts via monte carlo tree search with dynamic kv retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26443–26455.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

[2] https://github.com/cambridgeltl/weakest-link-effect

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

Bradley Efron and RJ Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press.

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. 2025. NeoQA: Evidence-based question answering with generated news events. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11842–11926, Vienna, Austria. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Qianguo Sun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642.

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.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Wenyu Huang, Pavlos Vougiouklis, Mirella Lapata, and Jeff Z Pan. 2025. Masking in multi-hop qa: An analysis of how language models perform with context permutation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17781–17795.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. *Github*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. 2025. More documents, same length: Isolating the challenge of multiple documents in RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19539–19547, Suzhou, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024a. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7675–7688.

Ruosen Li, Zimu Wang, Son Quoc Tran, Lei Xia, and Xinya Du. 2024b. Meqa: a benchmark for multi-hop event-centric question answering with explanations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024c. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025a. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2025b. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations*.

Jingyu Liu, Jiaen Lin, and Yong Liu. 2024a. How much can rag help the reasoning of llm? *arXiv preprint arXiv:2410.02338*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meta Llama Team. 2024. Llama 3.1-8b-instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Accessed: 2026-01-01.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257.

Mistral AI. 2024. Ministral-8b-instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410. Accessed: 2026-01-01.

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*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Shortformer: Better language modeling using shorter inputs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5493–5505.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Qwen Team and Others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *arXiv preprint arXiv:2406.13397*.

Maojia Song, Renhang Liu, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, Jingren Zhou, Dorien Herremans, and Soujanya Poria. 2025. Demystifying deep search: a holistic evaluation with hint-free multi-hop questions and factorised metrics. *arXiv preprint arXiv:2510.05137*.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822.

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.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy effect of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115.

Zhengren Wang, Jiayang Yu, Dongsheng Ma, Zhe Chen, Yu Wang, Zhiyu Li, Feiyu Xiong, Yanfeng Wang, Linpeng Tang, Wentao Zhang, and 1 others. 2025b. Rare: Retrieval-augmented reasoning modeling. *arXiv preprint arXiv:2503.23513*.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. 2025c. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2025. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.

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.

Zihao Yi, Delong Zeng, Zhenqing Ling, Haohao Luo, Zhe Xu, Wei Liu, Jian Luan, Wanxia Cao, and Ying Shen. 2025. Attention basin: Why contextual position matters in large language models. *arXiv preprint arXiv:2508.05128*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024a. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.

Sangwon Yu, Ik-hwan Kim, Jongyoon Song, Saehyung Lee, Junsung Park, and Sungroh Yoon. 2025a. Unleashing multi-hop reasoning potential in large language models through repetition of misordered context. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6435–6455.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 14672–14685.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2025b. Mitigate position bias in llms via scaling a single hidden states channel. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6092–6111.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. 2024. Quiet-star: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*.

Meiru Zhang, Zaiqiao Meng, and Nigel Collier. 2024a. Can we instruct llms to compensate for position bias? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12545–12556.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, and 1 others. 2024b. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37:60755–60775.

A Appendix

A.1 Implementation Details

We utilized vLLM (v0.8.5.post1) (Kwon et al., 2023) to perform inference in the default bf16 precision on A6000, 2x3090, and A100 GPUs. For all experiments, the temperature was set to 0 to enforce greedy sampling, with the random seed fixed at 42 for reproducibility. Despite greedy decoding, vLLM inference is not strictly bit-deterministic under concurrent batching (due to kernel-level numerical noise and batch-size-dependent attention computations), so exact per-cell EM values may drift by a fraction of a percentage point on reruns; the directional conclusions we report are robust to this residual stochasticity. Inference on the full dataset required approximately 0.5 hours for standard instruction-following models on 2x3090, faster on A100 and A6000. Qwen3-8B-Think required roughly 3 hours per run. In total, the full Spread and Cross tests for each model involved 150 discrete inference runs.

We selected the tested LLMs based on their maximum context window length and GPU compatibility. This includes Ministral-8B-Instruct, released in 2024, specifically Ministral-8B-Instruct-2410.[3] For Qwen3-8B, we evaluated both "thinking" and "non-thinking" modes following the official vLLM deployment guidelines.[4]

A.2 Additional Related Work on System-2 Reasoning

An emerging paradigm for mitigating reasoning failures involves transitioning from fast System 1 processing to deliberate System-2 reasoning (Li et al., 2025a). Models such as the OpenAI o1 series (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) leverage large-scale reinforcement learning to internalize chain-of-thought verification without supervised fine-tuning. Similarly, the Qwen3 series (Yang et al., 2025) unifies "thinking" and standard "non-thinking" modes within a dynamic compute budget. Beyond explicit reasoning models, techniques like *Quiet-STaR* (Zelikman et al., 2024) and *System-2 Attention* (Weston and Sukhbaatar, 2023) act as implicit context reconstruction mechanisms, generating internal rationales or filtering context before attending. Although Ma et al. (2025) question the strict necessity of explicit thought tokens in low-budget settings, our work empirically demonstrates that this extended test-time compute is crucial for robustness against position bias. We show that "thinking" models actively verify retrieval artifacts, overcom-

[3] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

[4] https://qwen.readthedocs.io/en/latest/deployment/vllm.html

ing the topological limitations of standard LLMs.

A.3 Examples of MuSiQue and NeoQA

MuSiQue Example 1

Question: Which county does Lloyd Dane's birthplace belong to?
Answer: Miller County
Gold Doc 1 (meta): id=para_3; title=Lloyd Dane; paragraph_idx=3
Gold Doc 1 (text): Lloyd Dane (August 19, 1925 – December 11, 2015) was a NASCAR Grand National Series driver from Eldon, Missouri. He participated part-time in the 1951 and 1954 to 1964 seasons, capturing four wins, all in his own car...
Gold Doc 2 (meta): id=para_11; title=Eldon, Missouri; paragraph_idx=11
Gold Doc 2 (text): Eldon is a city in Miller County, Missouri, United States, located thirty miles southwest of Jefferson City...
Logical Connection: Lloyd Dane was born in Eldon; Eldon is in Miller County.
Distractor Example (meta): id=para_19; title=Minsk Region
Distractor Example (text): Minsk Region or Minsk Voblasć or Minsk Oblast ... is one of the regions of Belarus. Its administrative center is Minsk...

MuSiQue Example 2

Question: Who wrote "Turn Me On" by performer of "Happy Pills"?
Answer: John D. Loudermilk
Gold Doc 1 (meta): id=para_0; title=Happy Pills (song); paragraph_idx=0
Gold Doc 1 (text): "Happy Pills" is a song by the American singer-songwriter Norah Jones. It is the lead single from her fifth studio album "Little Broken Hearts"...
Gold Doc 2 (meta): id=para_10; title=Turn Me On (Mark Dinning song); paragraph_idx=10
Gold Doc 2 (text): "Turn Me On" Single by Norah Jones from the album First Sessions ... Songwriter(s) John D. Loudermilk ...
Logical Connection: Performer of "Happy Pills" is Norah Jones; she performs "Turn Me On"; songwriter is John D. Loudermilk.
Distractor Example (meta): id=para_17; title=Birth control
Distractor Example (text): In 1909, Richard Richter developed the first intrauterine device ... Further developments followed in the 1950s...

NeoQA Example 1 (time-span question)

Event: Viroscope app lifecycle and governance changes .
Question: What is the total duration between the inferred start of the Viroscope app's six-month testing phase in Misterine City and the release of the report by HealthStream Tech Solutions about the feasibility of the citizen-led data trust model?
Answer: 1 year, 2 months, and 28 days
Gold Doc 1 (meta): article_id=epidemics_3_0-ev7-3; title=Major Overhaul Planned for Viroscope App...; date=2025-08-12
Gold Doc 1 (text): HealthStream Tech Solutions ... released a detailed report ... about adapting Viroscope to a citizen-led data trust model. The report, released on August 12, 2025, confirms ... a decentralized framework...
Gold Doc 2 (meta): article_id=epidemics_3_0-ev0-9; title=Drenvale Institute Launches "Viroscope"...; date=2024-11-15
Gold Doc 2 (text): The Drenvale Institute ... unveiled its mobile app "Viroscope" on November 15, 2024, in a bold move to revolutionize epidemic tracking ... "Viroscope" underwent six months of testing in Misterine City...
Logical Connection: Testing lasted 6 months; inferred start is 6 months before 2024-11-15, which is 2024-5-15; compute interval to 2025-08-12, the answer is 1 year, 2 months, and 28 days.
Distractor Example (meta): article_id=epidemics_3_0-ev4-10; title=Pilot Study of Viroscope App in Larnwick...
Distractor Example (text): The Drenvale Institute for Public Health is moving forward with the pilot rollout of the Viroscope app in Larnwick, scheduled to begin on July 1, 2025...

NeoQA Example 2

Event: Norhaven rollout delay and privacy oversight debate in the epidemics timeline (event-centric multi-hop).
Question: What specific process confirmed the compliance of the app with data protection standards, as stated by the individual who expressed regret over delays in Norhaven and emphasized public trust?
Answer: Independent security audit by a third-party firm
Gold Doc 1 (meta): article_id=epidemics_3_0-ev0-11; title="Drenvale Institute Launches 'Viroscope' App Amid Privacy Concerns"; date=2024-11-15
Gold Doc 1 (text): The Drenvale Institute for Public Health announced the release of its latest innovation ... addressed privacy concerns around Viroscope. Dr. Elara Tovrin ... stated the app does not collect personal identifiers ... the app underwent an independent security audit conducted by a third-party firm, which confirmed its compliance with data protection standards...
Gold Doc 2 (meta): article_id=epidemics_3_0-ev8-7; title=Privacy Concerns and Oversight Delay Viroscope App Rollout in Norhaven; date=2025-10-05
Gold Doc 2 (text): The Viroscope rollout in Norhaven was delayed due to privacy concerns ... Dr. Elara Tovrin expressed regret over the delay and reaffirmed that public trust remains the institute's top priority...
Logical Connection: Gold Doc 2 identifies the speaker (Dr. Elara Tovrin). Gold Doc 1 states the process she cited as confirming compliance, which is the independent third-party security audit.
Distractor Example (meta): article_id=epidemics_3_0-ev8-1; title=Norhaven Launches Review of Viroscope App Amid Privacy Concerns
Distractor Example (text): Norhaven's Ministry of Citizen Protection announced a comprehensive review ... emphasized compliance with the Data Protection Act of 2023 ... announced oversight and evaluation plans...

A.4 Heatmap Computation and Visualization

We visualize the attention of the first generated answer token, aggregated into semantic spans and differenced across MFAI conditions.

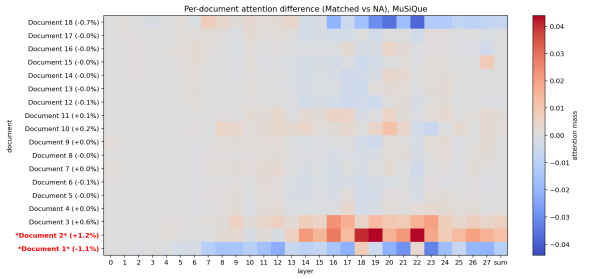
Extraction and aggregation. For each instance we render the prompt with the model's chat template and run the forward pass with the *eager* attention implementation, reusing the KV cache so only the target-token attention is computed. Let $a_{l,h,i}$ be that attention weight at layer l, head h for input token i. The prompt is segmented into S spans: task instruction, attention instruction (when present), question, answer options, and D document blocks. We average over heads to obtain a per-layer span matrix $M_{l,\sigma} = \frac{1}{H|\sigma|} \sum_{h,\, i\in\sigma} a_{l,h,i}$, and over a valid-layer range $L_{\text{valid}} = \{1, \ldots, 26\}$ (excluding layer 0, which is embedding-dominated, and the last output-specialized layers) to obtain a per-head span matrix $H_{\sigma,h} = \frac{1}{|L_{\text{valid}}||\sigma|} \sum_{l\in L_{\text{valid}},\, i\in\sigma} a_{l,h,i}$. We fix one bucket–distance configuration per figure (e.g., `gold_at_b_dist1` for the Beginning-bucket panel) and average the matrices across $N{=}100$ instances selected from that configuration by a priority-fill sampler: it first takes every Matched-correct / NA-incorrect case (causal-lift cases) and fills the remainder from cases where both modes are correct; Matched-incorrect / NA-correct cases are excluded.

Per-document share difference. Figure 7 shows the quantity we report: the difference, between two modes, of per-layer- (or per-head-) normalized document shares. Restricting $M_{l,\sigma}$ to the D document spans and normalizing so that each layer's doc shares sum to one, $\tilde{M}_{l,d} = M_{l,d} / \sum_{d'} M_{l,d'}$, we plot $\tilde{M}^{\text{matched}}_{l,d} - \tilde{M}^{\text{NA}}_{l,d}$. The per-head version uses the analogous column normalization of $H_{d,h}$, with a final `avg` column appended that is the row mean over heads. Normalizing before differencing factors out each layer's (or head's) overall doc-attention budget, so colors reflect *reallocation among documents*: red = larger share under Matched, blue = smaller. Gold documents are marked with an asterisk; instruction-targeted documents are highlighted in red. Per-document labels show the mean of the normalized-share difference over layers (or heads), multiplied by 100 (percentage points); the colorbar reports the raw signed share difference per cell, so a colorbar value of 0.02 corresponds to a $+2$ pp reallocation.

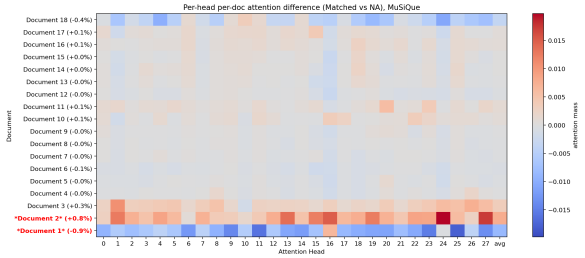
A.5 Statistical Analysis

We apply two statistical tests to validate the reliability of our experimental results: bootstrap confidence intervals for the accuracy estimates reported in the bar charts, and McNemar's test for the significance of performance differences between the No MFAI and Matched MFAI conditions.

Bootstrap Confidence Intervals. Each accuracy value reported in our figures is a point estimate computed over a finite sample of question–answer pairs. To quantify the uncertainty of these estimates, we construct 95% confidence intervals (CIs) using the bootstrap percentile method (Efron and Tibshirani, 1994). For a given experimental condition (model, dataset, protocol, position bucket), let $\mathbf{s} = (s_1, \ldots, s_N)$ denote the vector of N binary scores. We draw $B = 10{,}000$ bootstrap samples of size N with replacement and compute the replicate accuracy for each. The 95% CI is given by the



(a) Layer-wise Heatmap.



(b) Head-wise Heatmap.

Figure 7: Normalized document-share difference (Matched − No MFAI) of Qwen2.5-7B-Instruct on MuSiQue when gold documents are placed in the Beginning bucket (Doc 1 and Doc 2, configuration `gold_at_b_dist1`), averaged over 100 examples drawn with the priority-fill sampler described in Appendix A.4. Y-axis: document spans (with per-document mean over layers/heads, $\times 100$, shown in the label). X-axis: layer index in (a), head index in (b) with a final head-averaged `avg` column. The colorbar reports the raw signed share difference per cell; e.g., 0.02 corresponds to a $+2$ pp reallocation. Red = larger share under Matched; blue = smaller.

2.5th and 97.5th percentiles of the empirical distribution. These confidence intervals are displayed as error bars on the bar charts in Figures 2 and 3.

McNemar's Test. To test whether Matched MFAI significantly improves accuracy relative to the No MFAI baseline, we apply McNemar's test (McNemar, 1947), which is appropriate for comparing paired binary outcomes on the same set of examples. For each setting (model, dataset, protocol, position bucket), we evaluate the same N questions under both conditions, producing paired outcomes $(s_i^{\text{na}}, s_i^{\text{matched}})$ summarized in a 2×2 contingency table. Let b count questions correct under No MFAI but incorrect under Matched (degraded), and c count the reverse (improved). Under the null hypothesis that Matched MFAI has no effect, the test statistic with continuity correction is:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{1}$$

which follows a χ^2 distribution with one degree of freedom. We report significance at both $p < 0.05$ and $p < 0.01$ thresholds.

Tables 1 and 2 report the paired McNemar results for the five core instruction-tuned models plus Qwen3-8B-Think on MuSiQue and NeoQA, under the Spread and Cross protocols respectively. On MuSiQue, Matched MFAI significantly improves over No MFAI in every cell at $p < 0.01$ (36/36 across both protocols), consistent with the recognition-bottleneck interpretation: explicitly steering attention toward gold documents reliably rescues performance. On NeoQA, the effect is smaller and the significance pattern is mixed: most cells still reach $p < 0.01$ or $p < 0.05$, but a few cells with small Δ (e.g., Qwen2.5-7B Spread Middle/Tail, Qwen3-8B Cross B+M) do not reach significance, reflecting the smaller effect sizes on NeoQA's more homogeneous same-timeline context (No-MFAI baseline accuracies of ∼50–70% on NeoQA vs. ∼10–35% on MuSiQue). Per protocol, the Spread test yields 28/36 cells at $p < 0.01$ and 30/36 at $p < 0.05$, while the Cross test yields 28/36 at $p < 0.01$ and 32/36 at $p < 0.05$; aggregated, $56/72$ cells reach $p < 0.01$ and $62/72$ reach $p < 0.05$.

Dataset	Model	Beginning	Middle	Tail
MuSiQue	Llama-3.1-8B	$+7.22^{**}$	$+11.46^{**}$	$+11.49^{**}$
	Ministral-8B	$+1.73^{**}$	$+7.83^{**}$	$+7.91^{**}$
	Qwen2.5-7B	$+7.26^{**}$	$+12.12^{**}$	$+11.04^{**}$
	Qwen2.5-14B	$+5.07^{**}$	$+9.49^{**}$	$+7.58^{**}$
	Qwen3-8B	$+2.99^{**}$	$+8.56^{**}$	$+8.31^{**}$
	Qwen3-8B-Think	$+2.57^{**}$	$+5.26^{**}$	$+4.83^{**}$
NeoQA	Llama-3.1-8B	$+2.94^{**}$	$+4.28^{**}$	$+2.39^{*}$
	Ministral-8B	$+6.37^{**}$	$+3.13^{**}$	$+4.18^{**}$
	Qwen2.5-7B	$+3.98^{**}$	$+0.45$	$+1.79$
	Qwen2.5-14B	$+4.23^{**}$	$+3.33^{**}$	$+1.14$
	Qwen3-8B	$+1.79^{*}$	$+1.14$	$+0.80$
	Qwen3-8B-Think	$+2.89^{**}$	$+2.44^{**}$	$+0.95$

Table 1: McNemar's test: Matched MFAI vs. No MFAI on the **Spread Test** (MuSiQue, NeoQA). Rows are the five core instruction-tuned models plus Qwen3-8B-Think; columns are position buckets. Cell values are Δ accuracy (%) = Matched − NA. Significance markers: * $p < 0.05$, ** $p < 0.01$ (McNemar's test with continuity correction; two-sided, paired binary outcomes).

A.6 Ablation

Gold-Only Ablation. To establish an upper bound on model performance and isolate the impact of distractors, we conducted a gold-only ablation where models receive only the two gold documents. Table 3 presents the results. This setting represents

Dataset	Model	B+M	B+T	M+T
MuSiQue	Llama-3.1-8B	$+10.69^{**}$	$+10.96^{**}$	$+11.88^{**}$
	Ministral-8B	$+6.07^{**}$	$+6.06^{**}$	$+7.85^{**}$
	Qwen2.5-7B	$+11.81^{**}$	$+10.22^{**}$	$+12.23^{**}$
	Qwen2.5-14B	$+9.16^{**}$	$+8.53^{**}$	$+8.45^{**}$
	Qwen3-8B	$+6.93^{**}$	$+7.41^{**}$	$+8.12^{**}$
	Qwen3-8B-Think	$+4.24^{**}$	$+4.79^{**}$	$+5.35^{**}$
NeoQA	Llama-3.1-8B	$+3.19^{**}$	$+2.86^{**}$	$+1.49$
	Ministral-8B	$+3.94^{**}$	$+3.77^{**}$	$+3.11^{**}$
	Qwen2.5-7B	$+3.73^{**}$	$+4.02^{**}$	$+2.20^{*}$
	Qwen2.5-14B	$+1.78^{*}$	$+2.74^{**}$	$+1.41$
	Qwen3-8B	$+1.16$	$+1.99^{*}$	$+1.78^{*}$
	Qwen3-8B-Think	$+2.86^{**}$	$+1.87^{**}$	$+1.24$

Table 2: McNemar's test: Matched MFAI vs. No MFAI on the **Cross Test** (MuSiQue, NeoQA). Rows are the five core instruction-tuned models plus Qwen3-8B-Think; columns are bucket pairs (Beginning+Middle, Beginning+Tail, Middle+Tail). Cell values are Δ accuracy (%) = Matched − NA. Significance markers: * $p < 0.05$, ** $p < 0.01$ (McNemar's test with continuity correction; two-sided, paired binary outcomes).

the ideal scenario where recognition is perfect and the model only needs to perform reasoning or synthesis over the minimal necessary context. The performance gap between this gold-only condition and the full 18-document setting (reported in Figure 2 in Section 4.1) quantifies the cost of distraction, revealing the extent to which noise filtering bottlenecks each model.

Model	NeoQA	MuSiQue EM (%)		
	Acc. (%)	2h	3h	4h
Llama-3.1-8B-Instruct	66.42	40.45	35.54	32.84
Ministral-8B-Instruct	63.18	37.40	25.76	34.81
Qwen2.5-7B-Instruct	68.91	19.58	15.19	17.78
Qwen2.5-14B-Instruct	69.15	41.57	37.65	48.15
Qwen3-8B	71.64	42.46	34.21	42.22
Qwen3-8B-Think	71.64	44.70	—	—

Table 3: Gold-only exact-match accuracy (%) for NeoQA and MuSiQue at 2-hop, 3-hop, and 4-hop. Each setting feeds the model only the gold supporting documents (no distractors). "—" marks cells where the experiment was not run for that model. Qwen3-8B-Think has 2-hop only.

Specifically, all models, including Qwen3-8B-Think, exhibit performance degradation when distractors are present compared to the high-visibility Beginning bucket on MuSiQue. However, this degradation is typically less severe than the drop caused by position bias. For example, the performance of Ministral-8B-Instruct drops by 7.69% due to distractors when gold documents are in the Beginning bucket, but moving the gold documents to the Middle bucket leads to 11.04% drop. Performance on NeoQA also declines when distracting documents are added, though at a similar level to the drop caused by position bias. Surprisingly, Qwen3-8B-Think appears to benefit from the noise, which suggests that distractors may trigger more rigorous verification in "thinking" models. The gold-only upper bound generally drops from 2-hop to 3-hop (see 3-hop and 4-hop columns of Table 3), though the trend is non-monotone at 4-hop for several models.

No-Document Ablation. To verify that model performance is driven by the provided context rather than parametric memory, we removed all documents from the input. As shown in Table 4, accuracy drops to near-zero across all models for MuSiQue at all hop counts (2-, 3-, and 4-hop) and for NeoQA, with most models correctly outputting "Unanswerable" (except Qwen2.5-7B-Instruct). This confirms that models are indeed relying on the retrieved documents rather than memorized knowledge, validating the integrity of our experimental setup even as hop count increases.

Model	NeoQA		MuSiQue EM (%)		
	Acc. (%)	Unans. (%)	2h	3h	4h
Llama-3.1-8B-Instruct	0.00	100.00	0.08	0.26	0.25
Ministral-8B-Instruct	0.00	100.00	1.77	0.00	0.00
Qwen2.5-7B-Instruct	14.93	61.44	0.00	0.00	0.00
Qwen2.5-14B-Instruct	0.00	100.00	0.00	0.26	0.25
Qwen3-8B	1.74	94.03	0.16	0.00	0.25
Qwen3-8B-Think	0.00	100.00	0.08	—	—

Table 4: No-document ablation exact-match accuracy (%) for NeoQA and MuSiQue at 2-hop, 3-hop, and 4-hop. Documents are removed from the prompt; NeoQA "Unans." is the rate at which the model selects "Unanswerable". "—" marks cells where the experiment was not run.

A.7 Supplementary Validation: Additional Datasets and Model Scale

To assess the generalizability of the Weakest Link Effect beyond the two primary datasets and the five core models, we conducted supplementary evaluations on additional datasets, hop counts, and a larger model.

A.7.1 2WikiMultiHopQA

We evaluated the Compositional and Inference subsets of 2WikiMultiHopQA (2-hop; 2,685 and 736 examples respectively) on the five core models. The key patterns from MuSiQue replicate across both subsets. To reduce compute load on this supplementary evaluation, we sampled a sparse subset

of the inter-gold distance and cross local-index grid used for MuSiQue and NeoQA: Spread distances $d \in \{1, 3, 5\}$ (vs. 1–5 on MuSiQue/NeoQA) and Cross local indices $\in \{0, 3, 5\}$ (vs. 0–5), which still spans the within-bucket range.[5]

Step-function position bias. Table 5 reports the Spread NA accuracy per bucket across inter-gold distances 1, 3, and 5. Within-bucket variation is generally small (particularly in the Compositional subset), but shifting the evidence set between buckets causes a large drop, replicating the step-function pattern observed on MuSiQue.

Subset	Model	Bucket	$d=1$	$d=3$	$d=5$	Avg
Comp.	Llama-3.1-8B	Beginning	35.20	34.67	33.82	34.56
		Middle	24.88	24.99	24.62	24.83
		Tail	23.02	21.86	25.03	23.30
	Ministral-8B	Beginning	41.08	39.74	38.14	39.65
		Middle	29.91	28.90	27.86	28.89
		Tail	25.74	24.51	24.80	25.02
	Qwen2.5-7B	Beginning	28.23	26.52	27.00	27.25
		Middle	16.95	16.09	15.98	16.34
		Tail	16.57	16.65	19.44	17.55
	Qwen2.5-14B	Beginning	43.17	40.97	40.60	41.58
		Middle	35.83	34.53	34.30	34.89
		Tail	35.08	34.64	34.71	34.81
	Qwen3-8B	Beginning	41.12	39.93	38.47	39.84
		Middle	31.62	31.40	29.98	31.00
		Tail	29.46	29.05	29.68	29.40
Infer.	Llama-3.1-8B	Beginning	30.98	28.80	25.68	28.49
		Middle	13.86	12.77	12.09	12.91
		Tail	12.64	10.87	12.91	12.14
	Ministral-8B	Beginning	31.66	22.83	20.65	25.05
		Middle	9.24	8.56	8.70	8.83
		Tail	7.34	6.25	7.74	7.11
	Qwen2.5-7B	Beginning	7.88	7.07	7.74	7.56
		Middle	5.43	4.76	4.89	5.03
		Tail	5.98	4.62	6.11	5.57
	Qwen2.5-14B	Beginning	36.55	28.53	25.00	30.03
		Middle	13.04	12.64	10.19	11.96
		Tail	10.87	9.51	10.46	10.28
	Qwen3-8B	Beginning	23.51	19.02	17.53	20.02
		Middle	8.56	8.15	7.61	8.11
		Tail	6.25	6.25	6.11	6.20

Table 5: 2WikiMultiHopQA Spread NA accuracy (%) per bucket across inter-gold distances 1, 3, and 5.

MFAI effects on Spread. Table 6 shows that Matched MFAI consistently rescues the low-visibility buckets (positive Δ), while Unmatched MFAI degrades performance (negative Δ), demonstrating the asymmetric response to correct vs. misleading attention steering.

Cross local-index invariance. Table 7 shows that Cross accuracy is relatively stable across local indices (0, 3, 5) within each bucket-pair, indicating that bucket-pair membership rather than fine-grained offset is the primary driver of performance.

[5]Within-bucket variation is small on MuSiQue (cf. Figure 2); sampling every other distance and index was chosen to preserve the near-extreme cases while halving GPU cost.

Subset	Model	Bucket	NA	Matched (Δ)	Unmatched (Δ)
Comp.	Llama-3.1-8B	Beginning	34.56	38.83 (+4.27)	32.45 (-2.12)
		Middle	24.83	37.27 (+12.44)	23.34 (-1.49)
		Tail	23.30	36.82 (+13.52)	21.92 (-1.38)
	Ministral-8B	Beginning	39.65	40.09 (+0.43)	38.04 (-1.61)
		Middle	28.89	33.53 (+4.64)	26.44 (-2.45)
		Tail	25.02	30.75 (+5.74)	21.99 (-3.03)
	Qwen2.5-7B	Beginning	27.25	32.97 (+5.72)	24.54 (-2.71)
		Middle	16.34	28.38 (+12.04)	13.73 (-2.61)
		Tail	17.55	28.23 (+10.68)	15.56 (-2.00)
	Qwen2.5-14B	Beginning	41.58	43.70 (+2.12)	28.00 (-13.58)
		Middle	34.89	42.83 (+7.95)	26.18 (-8.71)
		Tail	34.81	41.01 (+6.19)	29.26 (-5.56)
	Qwen3-8B	Beginning	39.84	39.86 (+0.02)	34.58 (-5.26)
		Middle	31.00	37.01 (+6.01)	26.36 (-4.64)
		Tail	29.40	35.12 (+5.72)	24.63 (-4.77)
Infer.	Llama-3.1-8B	Beginning	28.49	37.14 (+8.65)	22.42 (-6.07)
		Middle	12.91	26.95 (+14.04)	12.79 (-0.11)
		Tail	12.14	24.05 (+11.91)	12.70 (+0.57)
	Ministral-8B	Beginning	25.05	28.58 (+3.53)	21.63 (-3.42)
		Middle	8.83	15.44 (+6.61)	7.93 (-0.91)
		Tail	7.11	13.09 (+5.98)	6.16 (-0.95)
	Qwen2.5-7B	Beginning	7.56	15.44 (+7.88)	7.16 (-0.41)
		Middle	5.03	12.86 (+7.84)	4.82 (-0.20)
		Tail	5.57	11.73 (+6.16)	5.71 (+0.14)
	Qwen2.5-14B	Beginning	30.03	33.42 (+3.40)	17.57 (-12.45)
		Middle	11.96	21.51 (+9.56)	8.88 (-3.08)
		Tail	10.28	19.34 (+9.06)	8.74 (-1.54)
	Qwen3-8B	Beginning	20.02	25.27 (+5.25)	13.99 (-6.02)
		Middle	8.11	17.98 (+9.87)	5.82 (-2.29)
		Tail	6.20	15.35 (+9.15)	5.41 (-0.79)

Table 6: 2WikiMultiHopQA Spread protocol MFAI effects per bucket. Δ is relative to NA in the same cell.

Weakest Link Effect on Cross. Table 12 reports the Cross protocol results with Weakest Link references. Cross NA typically falls between Spread Min and Spread Avg and sits well below the stronger constituent bucket, confirming the Weakest Link Effect on 2WikiMultiHopQA.

A.7.2 MuSiQue 3-Hop and 4-Hop

The MuSiQue 3-hop (757 examples) and 4-hop (405 examples) subsets confirm that the Weakest Link Effect generalizes to longer reasoning chains.

Placement configurations. Each placement label concatenates *bucket-letter* + *local-index* tokens per gold document, with bucket letters B (global 0–5), M (6–11), T (12–17) and local index $\in \{0, \ldots, 5\}$. For example, `b0b1b2` places three golds in Beginning; `b0b1m2` is a "2B+1M" split; 4-hop labels extend the pattern (`b0b1b2b3`, `b0b1b2m3`, etc.). Table 8 uses the all-in-one-bucket configs (`b0b1b2`, `m0m1m2`, `t0t1t2` and 4-hop analogues). Table 11 uses the mixed configs `2B+1M` (`b0b1m2` / `b0b1b2m3`) and `2B+1T` (`b0b1t2` / `b0b1b2t3`). *Spread Max* is the all-Beginning reference; *Spread Min* is the all-one-bucket NA for the weaker constituent (all-M or all-T).

Unmatched mirror. The *Unmatched* column in Table 11 keeps the gold documents in place but redirects the MFAI instruction into an empty (gold-free) bucket at the *same local indices* as the golds

Subset	Model	Pair	idx 0	idx 3	idx 5	Avg
Comp.	Llama-3.1-8B	B+M	33.67	27.15	24.58	28.47
		B+T	32.51	26.07	26.37	28.32
		M+T	24.88	22.57	24.92	24.12
	Ministral-8B	B+M	37.09	29.68	28.12	31.63
		B+T	35.64	26.85	27.52	30.01
		M+T	26.82	25.14	25.07	25.67
	Qwen2.5-7B	B+M	25.92	17.13	15.98	19.68
		B+T	25.70	17.84	20.37	21.30
		M+T	15.72	17.54	19.52	17.59
	Qwen2.5-14B	B+M	40.37	35.83	34.64	36.95
		B+T	39.96	35.31	35.27	36.85
		M+T	33.93	34.08	34.82	34.28
	Qwen3-8B	B+M	37.73	32.81	30.61	33.72
		B+T	36.31	30.91	30.50	32.58
		M+T	29.80	28.60	29.16	29.19
Infer.	Llama-3.1-8B	B+M	24.18	14.81	12.77	17.26
		B+T	22.15	14.27	13.45	16.62
		M+T	12.64	10.33	11.41	11.46
	Ministral-8B	B+M	19.16	10.73	9.24	13.04
		B+T	16.30	9.51	8.97	11.59
		M+T	8.56	7.07	7.34	7.65
	Qwen2.5-7B	B+M	7.61	3.80	5.57	5.66
		B+T	7.74	3.80	5.71	5.75
		M+T	4.89	4.76	5.84	5.16
	Qwen2.5-14B	B+M	25.54	13.32	11.55	16.80
		B+T	20.65	13.18	11.82	15.22
		M+T	11.28	10.19	10.05	10.51
	Qwen3-8B	B+M	17.66	8.97	7.34	11.32
		B+T	14.54	8.70	8.42	10.55
		M+T	7.61	6.79	6.39	6.93

Table 7: 2WikiMultiHopQA Cross NA accuracy (%) at local indices 0, 3, and 5 within each bucket-pair.

— i.e., the local-index pattern is preserved, only the bucket letter is swapped. When more than one bucket is empty, we choose by priority T > M > B. Concretely: `b0b1m2` (B, M occupied) → `t0t1t2`; `b0b1b2` (only B occupied) → `t0t1t2` (T preferred over M); `b0b1b2t3` (B, T occupied) → `m0m1m2m3`. Configurations that fill all three buckets (e.g., `b0m1t2`) are excluded because no empty bucket is available for the mirror.

Single-bucket position bias. Table 8 reports accuracy when all gold documents are placed in a single bucket. Position bias persists at higher hop counts: Beginning yields substantially higher accuracy than Middle or Tail for most models.

Cross Weakest Link. Table 11 shows that when gold documents are split between buckets, Cross performance approaches the weaker single-bucket reference (Spread Min) rather than the average, and Matched MFAI rescues the drop — confirming the Weakest Link Effect at 3-hop and 4-hop.

A.7.3 Model Scale: Qwen2.5-32B

To address whether the Weakest Link Effect persists at larger model scales, we evaluated Qwen2.5-32B-Instruct-GPTQ-Int8 on a subset comprising the first 600 MuSiQue examples. The findings align with the smaller models along three axes.

Model	Hop	All-B	All-M	All-T
Llama-3.1-8B	3h	25.10	19.55	19.68
	4h	21.48	15.31	16.54
Ministral-8B	3h	20.87	13.47	9.64
	4h	19.01	16.30	15.56
Qwen2.5-7B	3h	11.10	12.02	12.55
	4h	16.54	14.32	12.35
Qwen2.5-14B	3h	32.36	23.65	22.85
	4h	33.58	27.41	27.65
Qwen3-8B	3h	29.85	23.25	20.34
	4h	39.26	30.62	28.64

Table 8: MuSiQue multi-hop single-bucket baseline accuracy (%) at 3-hop and 4-hop, with all gold documents placed entirely in the Beginning / Middle / Tail bucket.

Step-function at 32B. Table 9 shows that within-bucket distance variation remains small while between-bucket variation is substantial, replicating the MuSiQue pattern at larger scale.

Bucket	$d = 1$	$d = 3$	$d = 5$	Avg
Beginning	40.33	38.00	36.33	38.22
Middle	32.83	32.17	31.00	32.00
Tail	32.67	31.83	33.83	32.78

Table 9: Qwen2.5-32B-Instruct-GPTQ-Int8 Spread NA accuracy (%) per bucket across inter-gold distances 1, 3, and 5 (first 600 MuSiQue examples).

MFAI at 32B. Table 10 confirms that Matched MFAI rescues low-visibility buckets and Unmatched degrades performance, mirroring the smaller-model pattern.

Bucket	NA	Matched (Δ)	Unmatched (Δ)
Beginning	38.22	47.06 (+8.83)	19.00 (-19.22)
Middle	32.00	44.39 (+12.39)	21.11 (-10.89)
Tail	32.78	44.72 (+11.94)	28.72 (-4.06)

Table 10: Qwen2.5-32B-Instruct-GPTQ-Int8 Spread protocol MFAI effects per bucket, pooled across distances 1, 3, 5. Δ is relative to NA.

Weakest Link at 32B. Table 13 shows that Cross performance falls at or below Spread Min and is fully restored by Matched MFAI, confirming the recognition bottleneck persists at scale.

A.7.4 NeoQA with Random-Timeline Distractors

As discussed in Section 4.3, NeoQA's original same-timeline distractors create context homogene-

Model	Config	Cross NA	Cross Matched	Cross Unmatched	Spread Min	Spread Max	Δ Max
Llama-3.1-8B	3h 2B+1M	21.00	32.63	17.97	19.55	25.10	− 4.10
	3h 2B+1T	19.42	30.52	17.44	19.68	25.10	− 5.68
	4h 3B+1M	18.27	31.11	20.49	15.31	21.48	− 3.21
	4h 3B+1T	15.80	29.38	17.04	16.54	21.48	− 5.68
Ministral-8B	3h 2B+1M	14.13	16.51	12.68	13.47	20.87	− 6.74
	3h 2B+1T	9.64	13.87	8.45	9.64	20.87	−11.23
	4h 3B+1M	14.81	15.56	14.07	16.30	19.01	− 4.20
	4h 3B+1T	14.57	15.56	13.83	15.56	19.01	− 4.44
Qwen2.5-7B	3h 2B+1M	8.72	18.23	5.81	12.02	11.10	− 2.38
	3h 2B+1T	8.98	16.78	7.27	12.55	11.10	− 2.11
	4h 3B+1M	13.58	22.96	10.62	14.32	16.54	− 2.96
	4h 3B+1T	13.33	21.23	12.59	12.35	16.54	− 3.21
Qwen2.5-14B	3h 2B+1M	26.29	33.03	15.72	23.65	32.36	− 6.08
	3h 2B+1T	23.51	28.27	14.93	22.85	32.36	− 8.85
	4h 3B+1M	31.60	39.75	20.74	27.41	33.58	− 1.98
	4h 3B+1T	30.12	36.54	18.77	27.65	33.58	− 3.46
Qwen3-8B	3h 2B+1M	24.17	28.93	16.51	23.25	29.85	− 5.68
	3h 2B+1T	21.66	28.01	15.06	20.34	29.85	− 8.19
	4h 3B+1M	32.59	38.52	27.90	30.62	39.26	− 6.67
	4h 3B+1T	30.12	34.57	24.69	28.64	39.26	− 9.14

Table 11: MuSiQue multi-hop Cross Weakest Link accuracy (%) for mixed-bucket gold placements. Config denotes gold-document distribution (e.g., 2B+1M = 2 in Beginning, 1 in Middle). Cross NA / Matched / Unmatched come from the mixed-placement files; Spread Min is the weaker single-bucket reference (M or T), Spread Max is the all-B single-bucket reference. Δ Max = Cross NA − Spread Max.

Subset	Model	Pair	Cross NA	Cross Matched	Unm. (mirror)	Unm. (random)	Spread Avg	Spread Min
Comp.	Llama-3.1-8B	B+M	28.47	37.62	33.11	26.51	29.70	24.83
		B+T	28.32	37.90	32.94	26.37	28.93	23.30
		M+T	24.12	36.54	31.16	23.45	24.07	23.30
	Ministral-8B	B+M	31.63	34.77	33.38	29.83	34.27	28.89
		B+T	30.01	33.49	30.92	27.88	32.33	25.02
		M+T	25.67	31.21	28.20	22.88	26.95	25.02
	Qwen2.5-7B	B+M	19.68	30.54	24.46	16.51	21.79	16.34
		B+T	21.30	30.66	26.28	19.73	22.40	17.55
		M+T	17.59	29.00	23.87	15.85	16.95	16.34
	Qwen2.5-14B	B+M	36.95	42.87	38.38	25.41	38.23	34.89
		B+T	36.85	42.21	37.77	29.10	38.19	34.81
		M+T	34.28	41.83	37.21	29.75	34.85	34.81
	Qwen3-8B	B+M	33.72	37.63	34.41	28.22	35.42	31.00
		B+T	32.58	36.82	32.94	27.88	34.62	29.40
		M+T	29.19	35.22	31.39	25.13	30.20	29.40
Infer.	Llama-3.1-8B	B+M	17.26	29.12	22.49	14.45	20.70	12.91
		B+T	16.62	28.89	21.90	13.77	20.31	12.14
		M+T	11.46	23.96	19.38	14.90	12.52	12.14
	Ministral-8B	B+M	13.04	17.98	15.49	10.91	16.94	8.83
		B+T	11.59	16.35	13.81	9.69	16.08	7.11
		M+T	7.65	12.59	10.33	6.30	7.97	7.11
	Qwen2.5-7B	B+M	5.66	14.54	8.99	5.07	6.30	5.03
		B+T	5.75	13.63	9.19	5.30	6.57	5.57
		M+T	5.16	12.64	9.90	5.84	5.30	5.03
	Qwen2.5-14B	B+M	16.80	24.32	20.90	11.50	20.99	11.96
		B+T	15.22	22.96	18.77	10.87	20.15	10.28
		M+T	10.51	19.47	15.76	8.24	11.12	10.28
	Qwen3-8B	B+M	11.32	19.70	16.39	7.56	14.06	8.11
		B+T	10.55	17.35	13.68	7.34	13.11	6.20
		M+T	6.93	15.90	12.59	5.12	7.16	6.20

Table 12: 2WikiMultiHopQA Cross protocol with Weakest Link references. Cross NA / Matched are per-pair accuracies (%); Unmatched is split into (mirror) = mean of the two partial-gold mirror variants, and (random) = the random-distractor variant. Spread Avg / Min are the mean and minimum of the two constituent single-bucket Spread NA values.

Pair	Cross NA	Cross Matched	Unm. (mirror)	Unm. (random)	Spread Avg	Spread Min
B+M	30.50	46.67	35.83	14.00	35.11	32.00
B+T	31.67	44.67	38.00	25.17	35.50	32.78
M+T	31.83	43.50	39.00	27.50	32.39	32.00

Table 13: Qwen2.5-32B-Instruct-GPTQ-Int8 Cross protocol with Weakest Link references. Cross NA / Matched are per-pair accuracies (%); Unmatched is split into (mirror) = mean of the two partial-gold mirror variants, and (random) = random-distractor variant. Spread Avg / Min are the mean and minimum of the two single-bucket Spread NA values.

ity that masks position bias. Table 14 shows the effect of replacing these with random-timeline distractors on Qwen2.5-7B-Instruct. Random distractors make the task easier overall, but a clear primacy bias emerges. Matched MFAI closes this gap, confirming that the attention steering mechanism is effective even when the underlying task topology is horizontal.

Setting	Condition	Beg	Mid	Tail	Pos Bias (B–T)
Same TL	NA	51.39%	51.04%	53.18%	-1.79%
	Matched	55.37%	51.49%	54.98%	+0.40%
	Unmatched	51.77%	52.79%	55.77%	-4.00%
Random TL	NA	62.29%	58.76%	58.46%	+3.83%
	Matched	62.84%	62.59%	62.19%	+0.65%
	Unmatched	62.24%	60.47%	61.84%	+0.40%

Table 14: NeoQA Spread Test: Same-Timeline vs. Random-Timeline distractors. Model: Qwen2.5-7B-Instruct. Averages across distances 1–5. TL = Timeline.

A.7.5 Test-Time Compute Cost: Thinking vs. Non-Thinking

The robustness gains of Qwen3-8B-Think come at a non-trivial inference cost. Table 15 reports the mean number of output tokens per example on NeoQA (same-timeline distractors) for Qwen3-8B in its non-thinking and thinking modes, pooled across all MFAI conditions and bucket positions. The thinking mode emits roughly $6\times$ more output tokens than its non-thinking counterpart on both Spread and Cross protocols, quantifying the budget required for the verification behavior discussed in the main text.

Model	Spread	Cross	Ratio
Qwen3-8B (non-thinking)	243.1	248.8	1.00×
Qwen3-8B-Think	1485.3	1512.7	6.09×

Table 15: Mean output tokens per example on NeoQA (same-timeline distractors) for Qwen3-8B in non-thinking mode vs. Qwen3-8B-Think. Tokens counted over `reasoning_content` + `model_response` using the Qwen3-8B tokenizer, pooled across all MFAI conditions and bucket positions (Spread: n=24, 120; Cross: n=36, 180).

A.8 Attention Instruction Order Sensitivity

Our MFAI injects the target document indices as a natural-language list (e.g., "Document i, Document j"). A potential concern is that the model could latch onto the specific ordering of this list rather than treating it as a set of cited positions, in which case reversing the index order inside the instruction would materially change the answer. We re-run every Cross Test cell with the indices inside the attention instruction reversed and compare against the original ordering. The test covers 96 cells in total (2 datasets $\times$ 2 models $\times$ 3 bucket-pairs $\times$ 2 local indices $\times$ 4 MFAI conditions). Table 16 reports both the absolute percentage-point change $|\Delta|$ and the relative change $|\Delta|$/baseline, since a 4 pp shift means something very different on a 55% baseline than on a 10% baseline.

Across all 96 cells the mean absolute change is 1.15 pp with a median of 0.80 pp, and the median relative change is 2.52%. MuSiQue-Llama is essentially invariant (mean 0.69 pp, max relative 6.32%) and NeoQA-Llama has a median relative change of only 0.89%. The largest relative shifts are concentrated in MuSiQue-Qwen, where the baseline accuracy is close to random (9.71–23.52%) so small absolute wiggles inflate in relative terms; all eight cells exceeding 10% relative change fall within `unmatched_*` control conditions rather than the `matched` cells that drive the main findings. We therefore treat the MFAI index list as order-agnostic throughout the paper.

A.9 Unmatched Instruction Variants and Detailed Analysis

To ensure that our **Unmatched MFAI** condition tests robustness against misleading signals rather than random noise, we generate adversarial indices by mirroring the local structure of the gold evidence. The specific variants for each protocol are:

Dataset	Model	Baseline	Mean $\lvert\Delta\rvert$	Med. $\lvert\Delta\rvert$	Max $\lvert\Delta\rvert$	Med. rel.	Max rel.
MuSiQue	Llama-3.1-8B	18.22–33.71	0.69	0.60	1.77	2.59%	6.32%
MuSiQue	Qwen2.5-7B	9.71–23.52	1.25	1.24	3.13	7.20%	17.18%
NeoQA	Llama-3.1-8B	54.73–60.45	1.11	0.50	4.48	0.89%	8.19%
NeoQA	Qwen2.5-7B	48.26–54.23	1.53	1.24	5.22	2.43%	10.82%
Pooled (96 cells)		9.71–60.45	1.15	0.80	5.22	2.52%	17.18%

Table 16: Effect of reversing the MFAI index order on Cross Test accuracy. Each per-model row aggregates n=24 cells (3 bucket-pairs $\times$ 2 local indices $\times$ 4 MFAI conditions). **Baseline** is the range of original accuracies (percent) over the 24 cells. $|\Delta|$ is the absolute percentage-point change in accuracy when the instruction indices are reversed, reported as the mean, median, and max over the 24 cells. **rel.** is the per-cell relative change $|\Delta_c|/\text{baseline}_c$ (as a percent); **Med. rel.** and **Max rel.** are the median and max of this quantity over the 24 cells.

Spread Test Variants: When the gold set $\mathcal{G}$ resides entirely within one bucket (e.g., Beginning), we generate two unmatched variants:

1. **Middle Mirror:** The instruction points to documents in the Middle bucket that share the same local indices as the gold documents.
2. **Tail Mirror:** The instruction points to documents in the Tail bucket that share the same local indices.

Cross Test Variants: When $\mathcal{G}$ is split across two buckets (e.g., Beginning and Middle), we use three unmatched variants to average out the effects of partial correctness:

1. **Partial Erroneous Mirror (Gold-1 correct):** The instruction correctly points to the gold document in the first bucket (Beginning) but points to a mirrored distractor in the non-gold bucket (Tail).
2. **Partial Erroneous Mirror (Gold-2 correct):** The instruction correctly points to the gold document in the second bucket (Middle) but points to a mirrored distractor in the non-gold bucket (Tail).
3. **Random Distractor:** The instruction points to two randomly selected documents within the non-gold bucket (Tail), ensuring no overlap with the gold indices.

Per-Variant Results. Figure 8 illustrates model performance on the unmatched variants of the Cross Test for MuSiQue and NeoQA. Comparing the variants in each bucket pair (rows), we observe that partially correct MFAI (gold1-mirror and gold2-mirror) remain helpful. The performance is better than with random misleading instructions and even exceeds the baseline (No MFAI), compared to solid lines in Figure 6. This reinforces the recognition bottleneck hypothesis that model failures stem primarily from positional neglect due to attention deficiency, and even partially correct cues can restore the focus.

A.10 Prompts

We used a uniform prompt for MuSiQue across all models. For NeoQA, however, we selected the optimal instruction for each model using gold-only settings, following the methodology established by prior work (Glockner et al., 2025).

A.10.1 The Prompt For MuSiQue

The standard MuSiQue prompt:

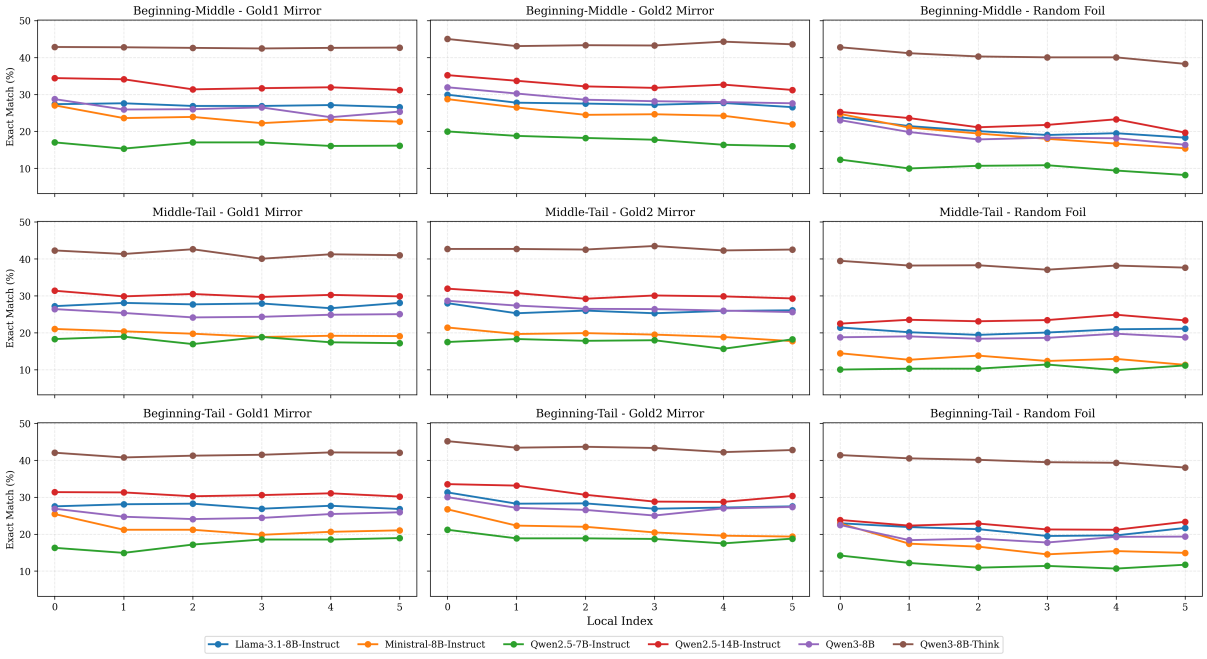
MuSiQue Standard Prompt

```
In this task, you are presented with a
    question, and 18 documents that
    covers the answer to that question.
    Deduce your answer solely from the
    provided documents, avoiding any
    external data sources. Keep the
    answer short and concise, leave
    behind any irrelevant details.
{{ATTENTION_INSTRUCTION}}

Question: {{QUESTION}}

Documents:
{{DOCUMENTS_BLOCK}}

If the documents don't have the answer,
    set "is_answerable" to false in the
    output dictionary. If they do, set "
    is_answerable" to true and put the
    answer in "answer_content".

Please provide your answer in the
    following format:
{"is_answerable": true/false, "
    answer_content": "your answer here"}
```

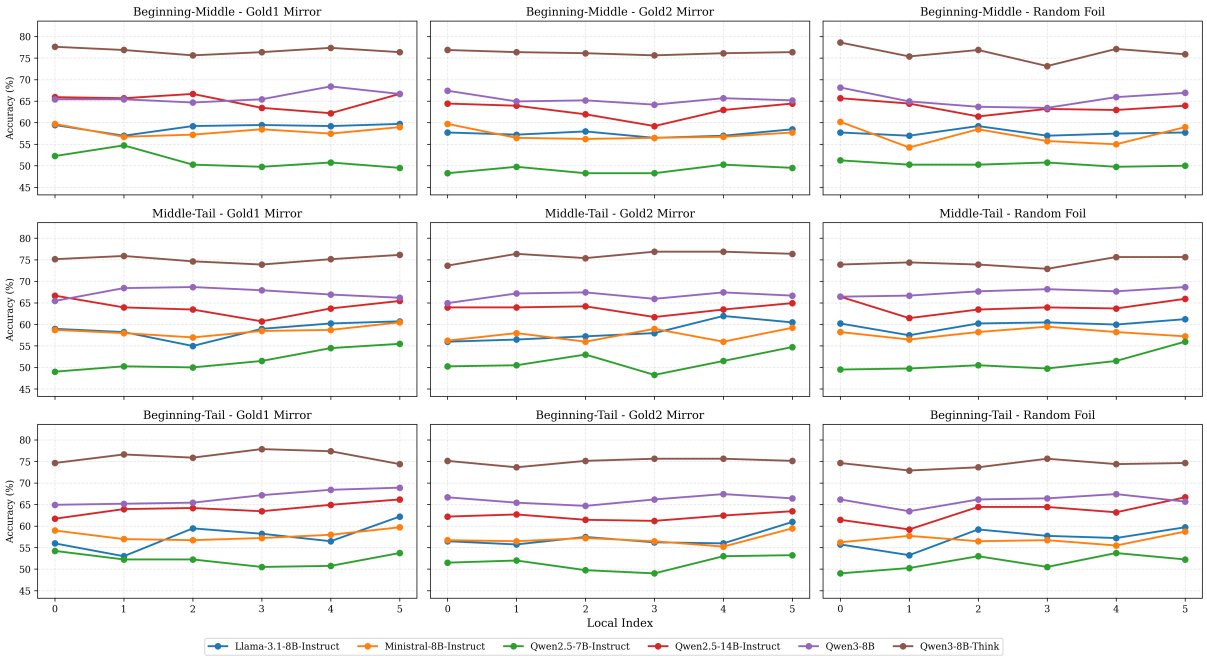
The Formatting of Document Each document is separated by line breaks.

MuSiQue Document Formatting

```
Document 1: Title of Article
Text of the article content...

Document 2: Title of Article
Text of the article content...
...
```



(a) MuSiQue



(b) NeoQA

Figure 8: Plot of the unmatched variants of Cross Test for MuSiQue and NeoQA datasets. Each subplot shows the performance of models in one of the three unmatched variants at the selected bucket pair. The x-axis represents the local index within the selected bucket pair, and the y-axis represents the accuracy. Each row corresponds to a different bucket pair, with the first row showing the Beginning–Middle bucket pair, etc. Each column corresponds to a different unmatched variant.

A.10.2 The Prompt For NeoQA

Model	Instruction Template
Qwen3-8B	last-line-instructions-1
Qwen2.5-7B-Instruct	last-line-instructions-2
Qwen2.5-14B-Instruct	last-line-instructions-1
Llama-3.1-8B-Instruct	last-line-instructions-2
Ministral-8B-Instruct	last-line-instructions-1

Table 17: Mapping of models to NeoQA prompt instruction files.

Prompt: last-line-instructions-1

```
Given the following news articles, the question, and
      the answer options, answer the question.
If the question cannot be answered with certainty
      based on the news articles, select the answer
      option "Unanswerable".

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

Select the answer option that correctly answers the
      question. If the question cannot be answered
      with certainty based on the news articles,
      choose "Unanswerable" (if it is one of the
      options). In the final line of your response,
      provide the number of the correct answer option
       using the format: "Answer: [answer number]" (
      for example, "Answer: X").
```

Prompt: last-line-instructions-2

```
You will receive news articles, a question, a date on
       which the question is asked, and answer
      options.
Your task is to evaluate the articles, determine if
      they provide enough information to answer the
      question based on the date, and choose the
      correct answer.

News Articles:
{{NEWS_ARTICLES}}

Question: {{QUESTION}}

Date of Question: {{DATE}}

Answer options:
{{ANSWERS}}

**Instructions:**
1. **Analyze the news articles:**
- Carefully read all the news articles.
- Compare the information in the articles with the
      question.
- Check if the combined information from the articles
       confirms all the details required to answer
      the question.

2. **Select an Answer:**
- Choose the correct answer if all necessary details
      are provided.
- If the articles lack information or any important
      detail is missing, select the option for "
      Unanswerable".

3. **Submit your Answer**
- Select the answer option that correctly answers the
       question. If the question cannot be answered
      with certainty based on the news articles,
      choose "Unanswerable" (if it is one of the
      options). In the final line of your response,
      provide the number of the correct answer option
       using the format: "Answer: [answer number]" (
      for example, "Answer: X").
```

News Articles Formatting (NEWS_ARTICLES) The documents are formatted using an XML-like structure that includes the title, date, and text content for each article. For each document in the context, the following structure is repeated, separated by two newlines:

Formatting of {{NEWS_ARTICLES}}

```
<article>
<title>Title of the Article</title>
<date>YYYY-MM-DD</date>
<text>Full text of the news article...</text>
</article>
```

Answer Options Formatting (ANSWERS) The multiple-choice options are formatted as a numbered list where each index is enclosed in square brackets. The answer options are provided as a list starting from index 1:

Formatting of {{ANSWERS}}

```
[1] Text for option 1
[2] Text for option 2
[3] Text for option 3
...
[4] Unanswerable
```