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

Anonymous ACL submission

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 establish the "Weakest Link Law": multi-hop reasoning performance collapses to the performance level of the least visible evidence. Crucially, this failure is governed by absolute position rather than the linear distance between facts (performance variance $< 3\%$). We further identify a duality in attention steering: while matched MFAI resolves recognition bottlenecks, improving accuracy by up to 11.5% in low-visibility positions, misleading MFAI triggers confusion in real-world tasks but is successfully filtered in synthetic tasks. Finally, we demonstrate that "thinking" models that utilize System-2 reasoning, effectively locate and integrate the required information, matching gold-only baselines even in noisy, long-context settings.

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 contest this view by analyzing evidence topology. By partitioning the context into positional "buckets", we reveal a step-function behavior: varying the distance between supporting facts yields negligible performance variance ($< 3\%$). Instead, multi-hop reasoning follows a "Weakest Link Law": 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 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 questions answering, the ability to steer attention provides a powerful control variable. It allows us to artificially restore evidence visibility, potentially isolating the **recog-**

nition failures (overlooked evidence) from **synthesis failures** (inability to connect facts).

Building on these observations, we introduce `Multi-Focus Attention Instruction (MFAI)`. 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 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 Law":** We provide a granular analysis of position bias in MHQA, establishing that performance collapses to the level of the least visible evidence bucket, regardless of the distance between facts.
- **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 a dataset-dependent behavior. Misleading cues degrade performance on real-world tasks (MuSiQue) yet unexpectedly trigger effective reasoning on synthetic tasks (NeoQA) due to structural alignment.
- **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 (MHQA) 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 these 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. In this work, we extend these findings by using MFAI not merely to mitigate noise, but to mechanically isolate the interaction between evidence topology and these distinct failure modes (Recognition vs. Synthesis).

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

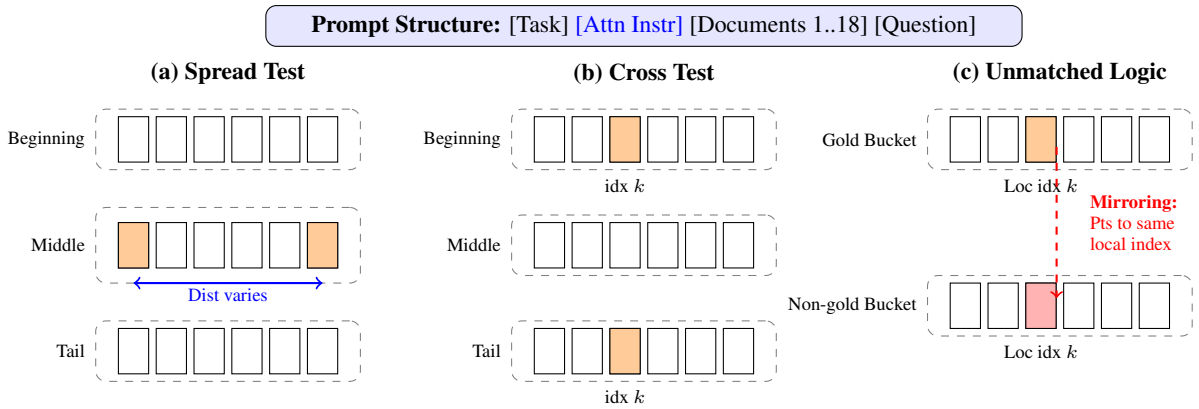


Figure 1: Experimental setups shown in vertical columns. (a) **Spread Test:** Gold documents (orange) are placed in the Middle bucket, and the distance between them varies. (b) **Cross Test:** Gold documents are split between the Beginning and Tail buckets, maintaining the same local index k. (c) **Unmatched 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.

et al., 2024; Huang et al., 2025), our analysis challenges this continuous view, proposing a discrete step-wise "Weakest Link Law".

Existing mitigation strategies predominantly relied on training data augmentation (Li et al., 2024c; He et al., 2024), architectural adjustment 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. To ensure a rigorous control, we select the adversarial indices by **mirroring the local bucket index** of the gold documents in a non-gold bucket (see Figure 1(c)). This isolates the position bias across buckets from the in-bucket local positions effects.

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.3).

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). 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.).

3.5 Models

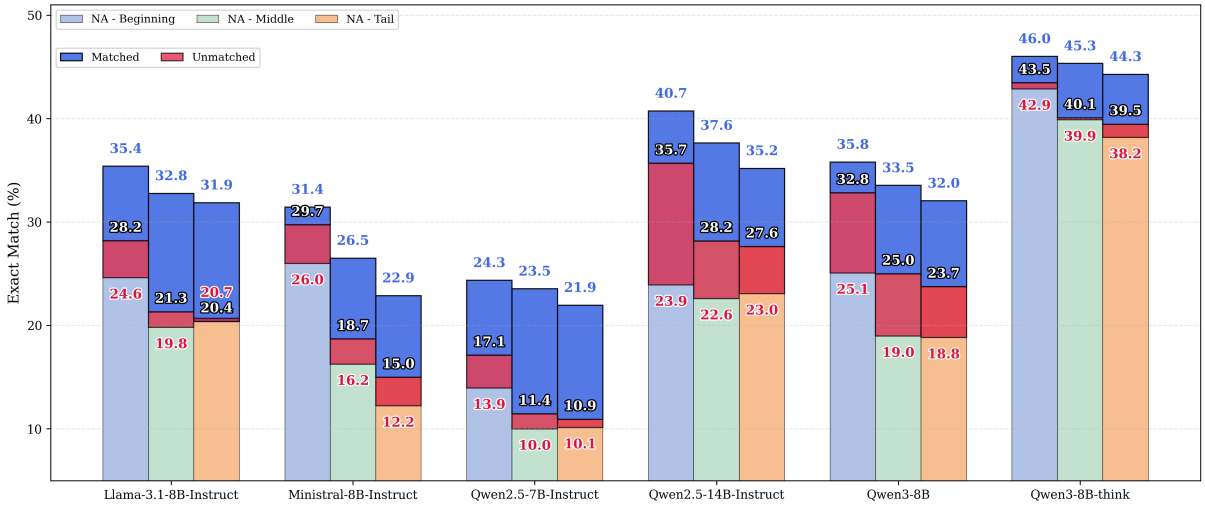
We evaluate five state-of-the-art Large Language Models: Qwen2.5-7B-Instruct (Qwen Team and Others, 2025), Qwen2.5-14B-Instruct (Qwen Team and Others, 2025), Llama-3.1-8B-Instruct (Meta Llama Team, 2024), Ministral-8B-Instruct (Mistral AI, 2024), and Qwen3-8B (Yang et al., 2025). Qwen3-8B is assessed in both thinking (triggered via `<think>`) and non-thinking (triggered via `</no_think>`) modes to determine the impact of test-time compute in 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).

4 Results and Discussion

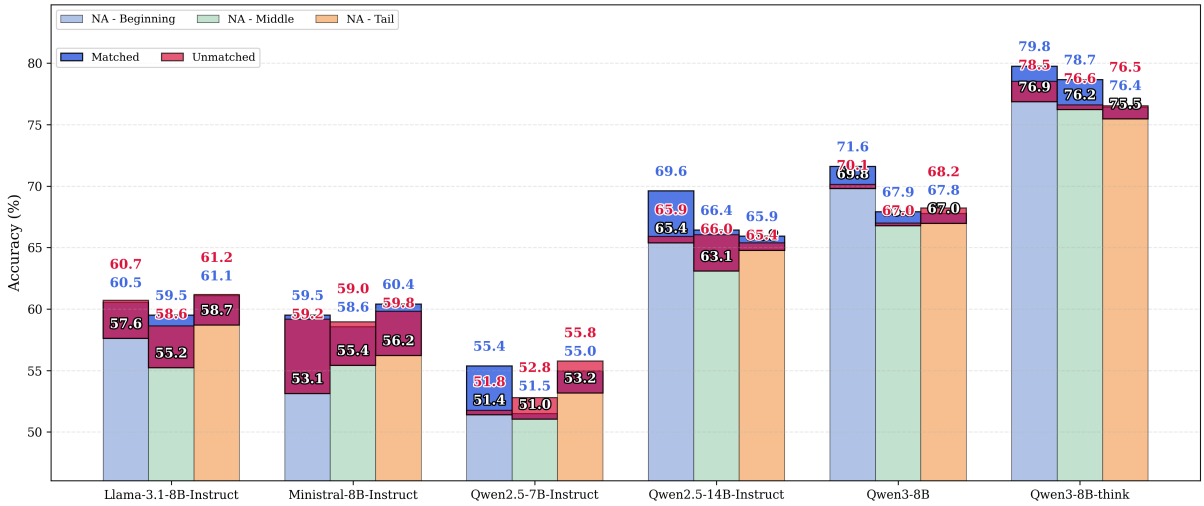
4.1 Evidence Topology: Absolute Position Governs Performance (RQ1)

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 where 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 $\sim 30\%$ 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 14.7% on MuSiQue). This confirms that attention functions as a step-function governed by coarse-grained bucket location. The Cross Test (Figure 3) illustrates a much more stable performance variation when gold documents are placed across buckets, suggesting that the absolute position governs performance. Furthermore, Figure 6) shows that the performance remains flat across lo-



(a) MuSiQue dataset



(b) NeoQA dataset

Figure 2: Model performance across various MFAI (No MFAI, Matched, and Unmatched) on the MuSiQue and NeoQA datasets within the Spread Test. Bars represent the average accuracy across five inter-gold-document distances within each positional bucket. Accuracy values are presented as percentages, with No MFAI in white, Matched in blue, and Unmatched in red.

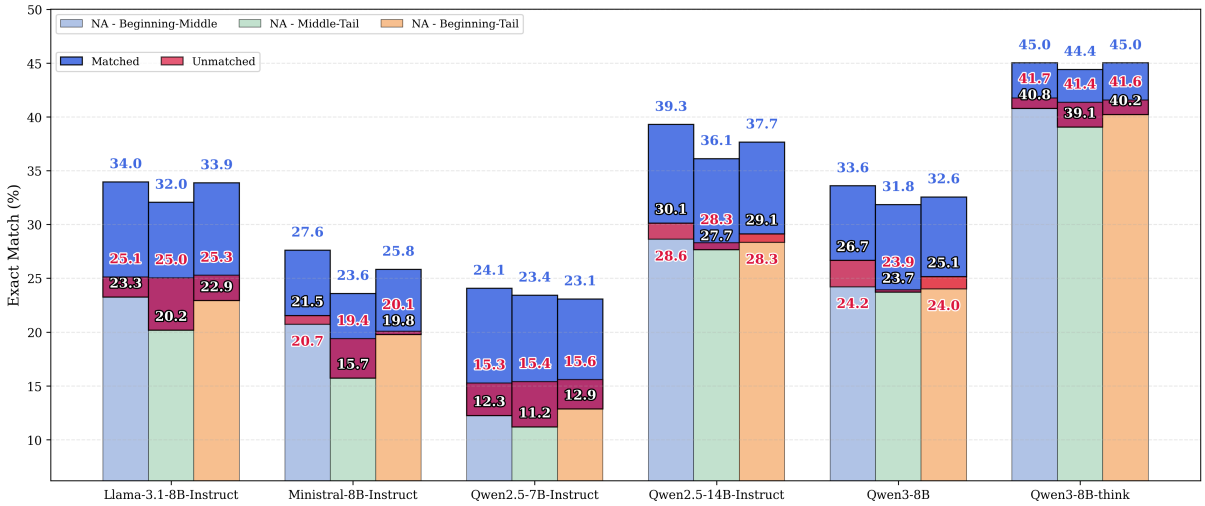
cal indices (0 to 5) as long as the bucket combination is constant, indicating that the model's attentional focus is static.

The "Weakest Link Law" 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 (Beginning) and a low-visibility bucket (Middle), performance does not average out; it collapses to the performance level of the weakest bucket. We term this the **Weakest Link Law**: the probability of successful reasoning is bounded by the minimum recognition 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.

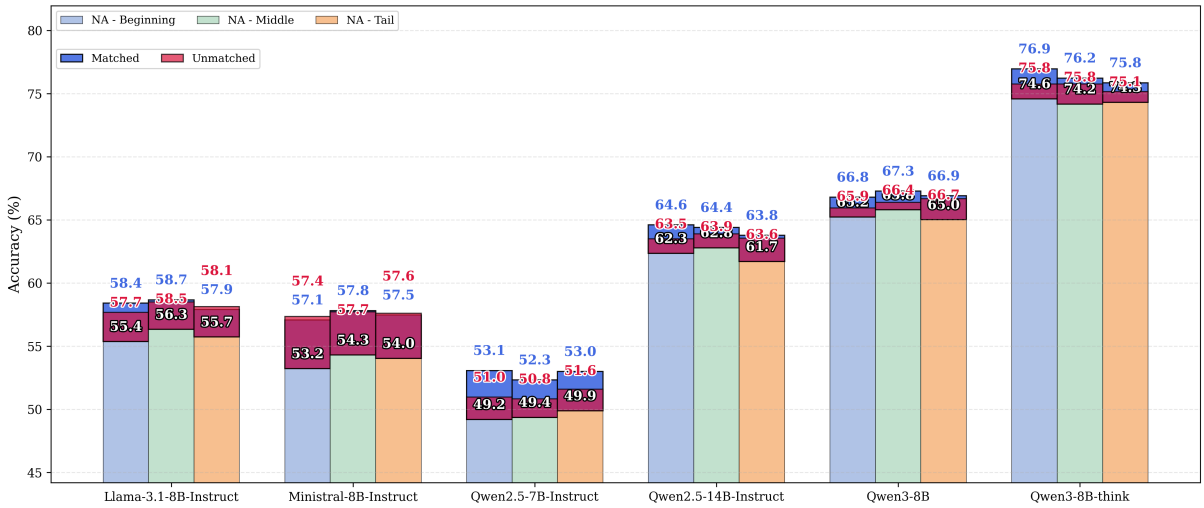
4.2 The Nature of Failure: Recognition as the Bottleneck (RQ2)

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 index 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 dis-



(a) MuSiQue dataset

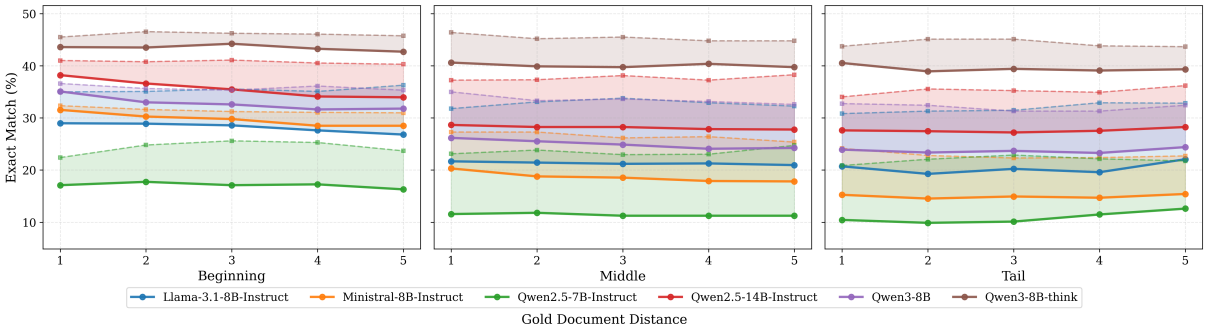


(b) NeoQA dataset

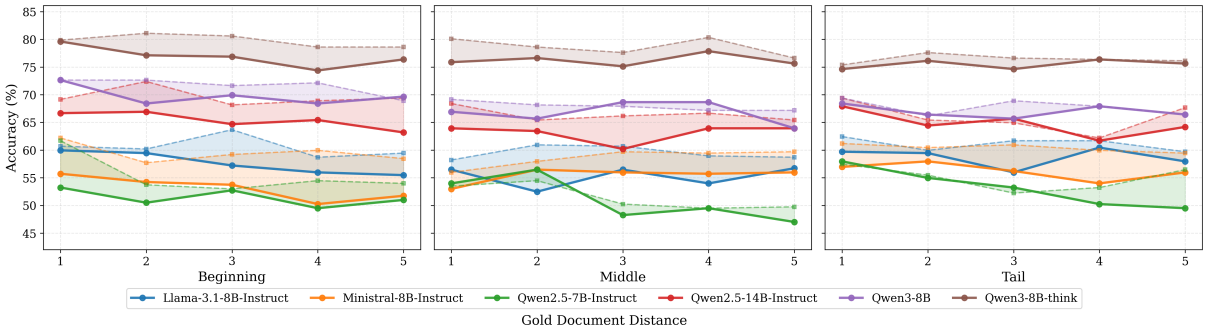
Figure 3: Model performance across various attention instruction conditions (No MFAI, Matched, Unmatched) on the MuSiQue and NeoQA datasets within the Cross Test. Bars represent the average accuracy across six local indexes within the selected pair of positional buckets. Accuracy values are presented as percentages, with No MFAI in white, Matched in blue, and Unmatched in red.

tances and local indices. It elevates performance in the Tail bucket by 4.8% to 11.5% across models on MuSiQue, effectively bridging the gap to Beginning-bucket performance (Figure 2(a)). In addition, the performance gap of model 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 synthesis, but this capacity is bottlenecked by attentional failures. NeoQA behaves differently: because the baseline position bias is limited, Matched MFAI provides smaller gains and 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.

Redistributing Attention Mass. This recovery of model performance is driven by a observable sshift in internal attention. As shown in the attention heatmaps (Figure 5), averaged over 20 MuSiQue examples where Qwen2.5-7B-Instuct was corrected by MFAI, 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. This redistribution confirms that the failure mode is a

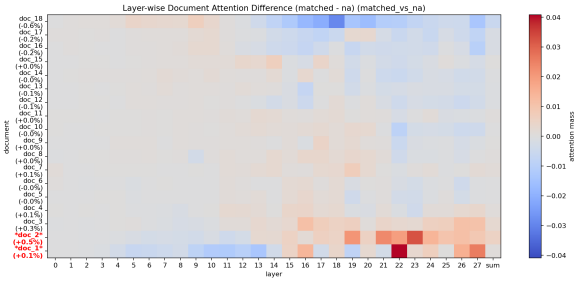


(a) MuSiQue dataset

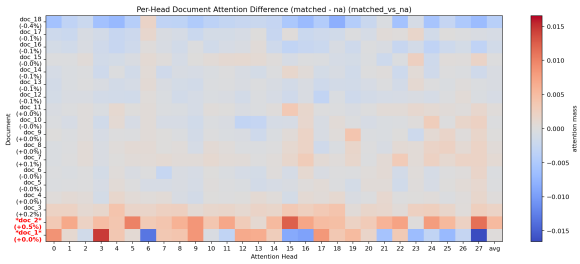


(b) NeoQA dataset

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



(a) Layer-wise Heatmap.



(b) Head-wise Heatmap.

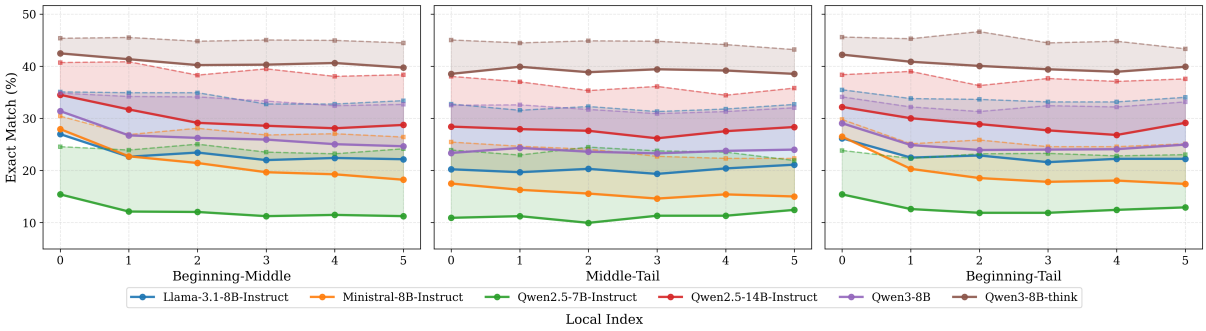
Figure 5: The attention mass difference between matched and No MFAI of Qwen2.5-7B-Instruct on NeoQA when gold documents are placed at the beginning bucket (Doc 1 and Doc 2), averaged over 20 examples. The percentage represents the normalized attention mass per token received on each document over all documents. Blue cells represent a drop in attention mass; red represents an increase.

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.

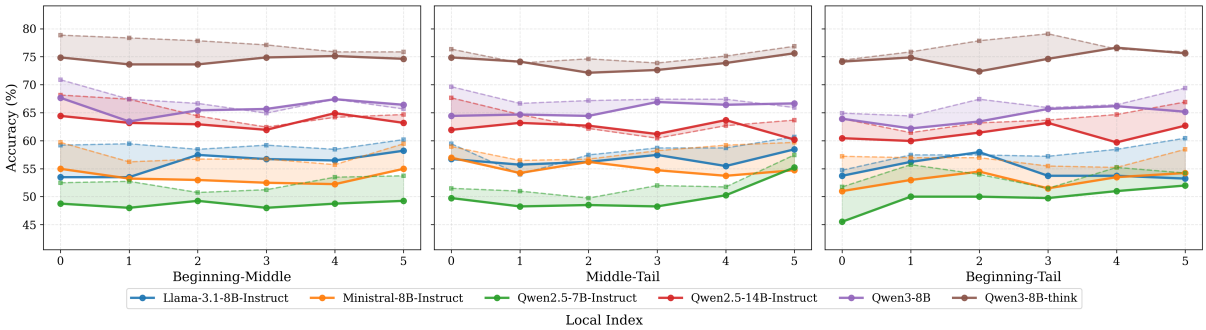
4.3 Robustness and Verification: Unmatched MFAI and System-2 Reasoning (RQ3)

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

The Dual-Edge of Unmatched Instructions. We observe a distinct difference in how models handle Unmatched MFAI. On the real-world MuSiQue dataset, unmatched instructions degrade performance and trigger a $\approx 2\times$ increase in response length across instruction-following models, which signals confusion. Conversely, on the synthesized NeoQA dataset, unmatched instructions unexpectedly maintain or even slightly improve accuracy. We attribute this to NeoQA's synthetic nature. Generated by GPT-4o (Jaech et al., 2024), its structural patterns likely align with model priors, enabling effective reasoning even with misleading cues (Li



(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, dash line represents the matched instruction accuracy.

et al., 2025b). This allows models to locate correct logic even when steered incorrectly. Detailed analysis of specific unmatched variants (Appendix A.5) shows that partially correct instructions (one gold index provided) outperform random non-gold indices, confirming that even imperfect cues can trigger synthesis.

System-2 Reasoning overrides Recognition Artifact. Comparing standard instruction-following models with System-2 reasoning models (e.g., Qwen3-8B-Think) reveals that extended test-time compute ($\sim 5\times$ more tokens than its non-thinking counterpart) 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 1 in Appendix A.4), standard models (e.g., Llama-3.1-8B) suffer a significant penalty in the 18-document setting (e.g., dropping from 40.45% to avg. 23.3% 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.

5 Conclusion

We analysed position bias and disentangled the mechanisms of failure in long-context multi-hop reasoning, distinguishing between recognition and synthesis deficits. We established the "Weakest Link Law", showing that multi-hop performance is constrained by the absolute position of the least visible evidence bucket and driven primarily by a recognition bottleneck. While semantic steering can artificially resolve this bottleneck, it yields contrasting effects: misleading instructions highlight the sensitivity of instruction-following models on the real-world dataset while revealing robustness on the synthetic benchmark. These findings underscored the potential of generating and sketching the summary of documents before reasoning. In addition, we demonstrated that System-2 models utilize test-time compute to override instruction and context noise, suggesting that future long-context architectures must incorporate self-verification strategies to achieve robust reasoning.

Limitations

We focused on multi-hop scenarios and selected entity-based (MuSiQue) and event-based (NeoQA) datasets, but the performance of the models on other benchmarks is left for future work. We did not evaluate other reasoning models and larger LLMs, such as those with 70B parameters, or perform a granular mechanistic analysis using logprobs and perplexity to further verify the internal reasoning processes. Additionally, we did not explore the effect of prompt variations, the number of positional buckets, or larger context sizes beyond 18 documents. Finally, our experiments were conducted under controlled settings with a fixed order of distractors to exclude the impact of semantic similarity reranking common in RAG pipelines. Future work should address real-world scenarios where semantic retrievers or rerankers are applied to select and order documents.

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 Law" to explain position bias of 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 during experiments. We will release code and data used upon acceptance.

We are committed to the full reproducibility of this study. Upon publication, we will release the complete source code and all curated data artifacts under a permissible 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.

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*.

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.

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, LiuYiBo LiuYiBo, Qianguosun Qianguosun,

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.

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. *arXiv preprint arXiv:2505.11754*.

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*.

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, Soujanya Poria, and Jingren Zhou. 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 Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah 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. 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 select 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.[1] For Qwen3-8B, we evaluated both "thinking" and "non-thinking" modes following the official vLLM deployment guidelines.[2]

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, overcoming the topological limitations that of standard LLMs.

A.3 Unmatched Instruction Variants

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:

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.

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

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

A.4 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 1 presents the results. This setting represents 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 inSection 4.1) quantifies the cost of distraction, revealing the extent to which noise filtering bottlenecks each model.

Specifically, all models, including Qwen3-8B-Think, exhibit performance degradation when distrators 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 7.7% on due to distractors when gold documents in the Beginning bucket, but moving the gold documents to the Middle bucket leads to 11% 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 shows to benefit from the noise, which suggest that distractors may trigger more rigorous verification in "thinking" models.

Model	NeoQA Acc. (%)	MuSiQue EM (%)
Llama-3.1-8B-Instruct	66.42	40.45
Ministral-8B-Instruct	63.18	37.40
Qwen2.5-7B-Instruct	68.91	19.58
Qwen2.5-14B-Instruct	69.15	41.57
Qwen3-8B	71.64	42.46
Qwen3-8B-think	**71.64**	**44.70**

Table 1: Gold-only exact-match accuracy for NeoQA and MuSiQue.

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 2, accuracy drops to near-zero across all models for MuSiQue and NeoQA (except Qwen2.5-7B-Instruct), with most models correctly outputting "Unanswerable" for NeoQA (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.

Model	NeoQA		MuSiQue
	Acc. (%)	Unans. (%)	EM (%)
Llama-3.1-8B-Instruct	0.00	100.00	0.08
Ministral-8B-Instruct	0.00	100.00	1.77
Qwen2.5-7B-Instruct	14.93	61.44	0.00
Qwen2.5-14B-Instruct	0.00	100.00	0.00
Qwen3-8B	1.74	94.03	0.16
Qwen3-8B-think	0.00	100.00	0.08

Table 2: No-document ablation exact-match accuracy for NeoQA and MuSiQue. "NeoQA Unans." shows the percentage of examples where the model selected "Unanswerable".

A.5 Detailed Analysis of Unmatched Cross Test Variants

Figure 7 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), compare to solid lines in Figure 6). This reinforces the recognition bottleneck hypothesis that models failures stem primarily from positional neglect due to attention deficiency, and even partially correct cues can restore the focus.

A.6 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).

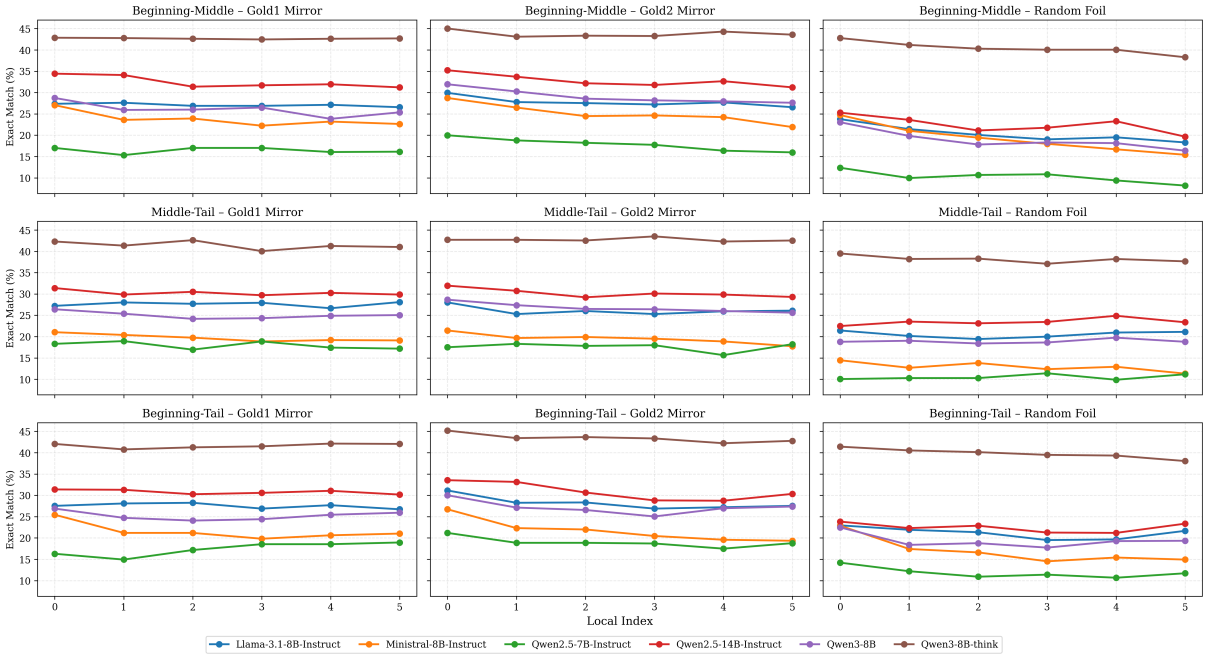
A.6.1 The Prompt For MuSiQue

The standard MuSiQue prompt:

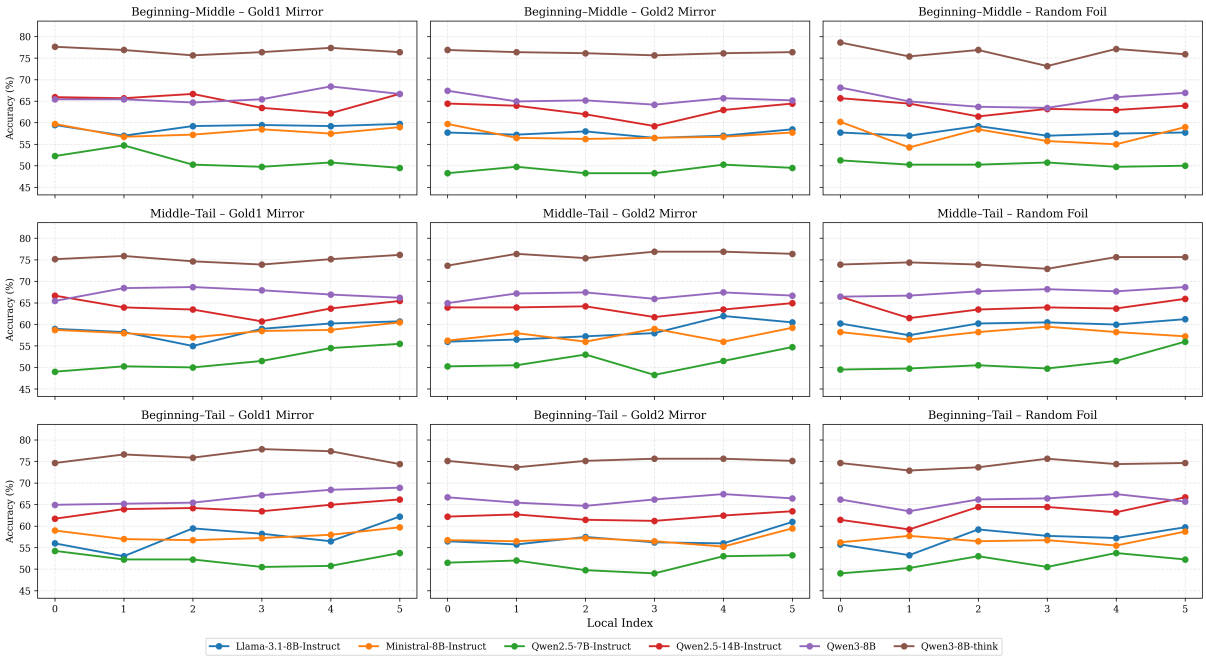
MuSiQue Standard Prompt

```
In this task, you are presented with
    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}}
```



(a) MuSiQue



(b) NeoQA

Figure 7: 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 beginning-middle bucket pair and etc. Each column corresponds to a different unmatched variant.

```
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.6.2 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 3: 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
```