Masking in Multi-hop QA: An Analysis of How Language Models Perform with Context Permutation

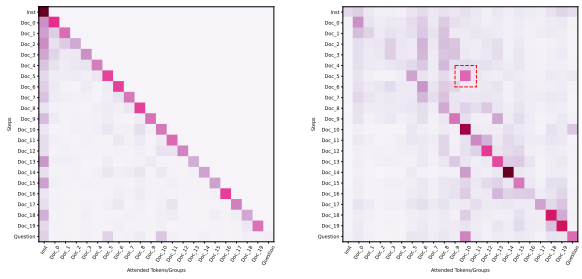
Wenyu Huang[1], **Pavlos Vougiouklis**[2], **Mirella Lapata**[1], **Jeff Z. Pan**[1,2]*
[1]School of Informatics, University of Edinburgh
[2]Huawei Edinburgh Research Centre, Poisson Lab, CSI, UK
w.huang@ed.ac.uk, pavlos.vougiouklis@huawei.com, mlap@inf.ed.ac.uk,
http://knowledge-representation.org/j.z.pan/

Abstract

Multi-hop Question Answering (MHQA) adds layers of complexity to question answering, making it more challenging. When Language Models (LMs) are prompted with multiple search results, they are tasked not only with retrieving relevant information but also employing multi-hop reasoning across the information sources. Although LMs perform well on traditional question-answering tasks, the causal mask can hinder their capacity to reason across complex contexts. In this paper, we explore how LMs respond to multi-hop questions by permuting search results (retrieved documents) under various configurations. Our study reveals interesting findings as follows: 1) Encoder-decoder models, such as the ones in the Flan-T5 family, generally outperform causal decoder-only LMs in MHQA tasks, despite being significantly smaller in size; 2) altering the order of gold documents reveals distinct trends in both Flan T5 models and fine-tuned decoder-only models, with optimal performance observed when the document order aligns with the reasoning chain order; 3) enhancing causal decoder-only models with bi-directional attention by modifying the causal mask can effectively boost their end performance. In addition to the above, we conduct a thorough investigation of the distribution of LM attention weights in the context of MHQA. Our experiments reveal that attention weights tend to peak at higher values when the resulting answer is correct. We leverage this finding to heuristically improve LMs' performance on this task. Our code is publicly available at https://github.com/hwy9855/MultiHopQA-Reasoning.

1 Introduction

Language Models (LMs) based on Transformer architectures (Vaswani et al., 2017) have become essential tools for a wide variety of tasks, including question answering and conversational search (Zhuang et al., 2023; He et al., 2023; Yi et al., 2024; Mo et al., 2024; Wu et al., 2024; Huang et al., 2025b; Wang et al., 2025). Among the different architectures, causal decoder-only configurations have become a popular choice for many of the most widely known LM families (Llama Team, 2024; Qwen, 2025).



(a) Qwen 2.5, head 3 (b) FlanT5, head 1

Figure 1: Context attention distribution of Qwen 2.5 1.5B and FlanT5 large for the same question, both captured from the last layer (last encoder layer for FlanT5). The gold documents are Doc_5 and Doc_10, where the reasoning direction is Doc_10 → Doc_5. With bi-directional attention, FlanT5 allows Doc_5 to assign attention and "see" Doc_10 (red dashed box), whereas Qwen 2.5 with causal mask cannot.

LMs exhibit a strong ability to reason within their input context, allowing for adaptability and effective generalisation across a wide range of tasks (OpenAI, 2024). Albeit these capabilities, previous studies have highlighted significant challenges regarding *the extent to which LMs can reason across different input contexts* (Zhang et al., 2023; Kaddour et al., 2023). An exemplary case is the "lost in the middle" problem (Liu et al., 2024), where crucial information positioned in the middle of the context may be overlooked by LMs. As LMs are increasingly utilised in complex scenarios, their ability to reason across different contexts becomes critically important. The general Retrieval-augmented Generation (RAG) framework (Lewis et al., 2020;

*Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17781–17795
July 27 - August 1, 2025

Li et al., 2024; Shen et al., 2024) has recently become the cornerstone of many search-based conversational agents, such as Copilot and Doubao. In this setting, LMs frequently need to synthesise information from multiple retrieved search results to provide coherent answers.

This study conducts an in-depth analysis of how LMs reason across various contexts, specifically focusing on multi-hop question answering (MHQA). Within the RAG framework, MHQA necessitates synthesising knowledge from multiple documents, presenting a more complex level of information integration than other more trivial question-answering tasks. To successfully perform MHQA, LMs must not only identify the most relevant documents within a given context but also reason with the information from these documents to determine the correct answer.

A critical research question arises from the architecture of modern causal decoder-only Transformer-based LMs, which use a causal mask during both training and inference. This constraint hinders these models from performing bi-directional encoding (Raffel et al., 2020), as opposed to traditional encoder-decoder architectures that can capture interactions between documents more effectively (cf. Figure 1). This leads us to examine whether these widely used causal decoder-only LMs are limited by this constraint. Further, what might be the impact if we replace the causal mask with a bi-directional (i.e. prefix) mask?

To address these inquiries, we conduct a comprehensive investigation about how LMs model the MHQA task. Initially, we evaluate the MHQA performance of three widely adopted open-source LM families: the Flan-T5 (Chung et al., 2024) family, representing the traditional encoder-decoder architecture, and the Qwen2.5 (Qwen, 2025) and Llama 3.x (Llama Team, 2024) families, exemplifying two popular causal decoder-only LMs. We design three types of document permutations to investigate the LMs' behaviour in terms of: 1) the order of gold documents[1], 2) the distance between them, and 3) their completeness. Our observations include:

1. The encoder-decoder model (Flan T5) is a superior MHQA solver when no fine-tuning takes place.

2. Fine-tuned LMs tend to favour forward-placed documents (Chen et al., 2024) (i.e., when the order of gold documents in the context mirrors the order of the reasoning chain, cf. Figure 2), a trend also observed in the Flan T5 models.

3. Bi-directional attention with prefix mask can benefit LMs in MHQA tasks and offers better robustness when the order of gold documents is altered.

4. The distance between gold documents significantly affects performance.

5. While the removal of the first hop document reduces MHQA performance, a relatively high level of correctness is still maintained.

Building on the above observations, we delve deeper to analyse how LMs perform MHQA by examining the attention distribution across layers. Our findings reveal that LMs typically assign higher attention score weights to at least one document when they correctly answer a multi-hop question. By sampling answers with different input document permutations, and retaining the answers for the inputs for which the LM assigned the largest peak attention score, we increased the accuracy of Qwen 7B from 28.6% to 33.7%.

2 Related Works

2.1 RAG and MHQA

The RAG framework has been widely used to inject external knowledge (Pan et al., 2023) into LMs (Gao et al., 2024), and mitigate their hallucination tendencies (Huang et al., 2025a). In this framework, LMs generate a response by conditioning it on the top search results, which are provided as input context. This setup has shown promising performance in a variety of tasks, including knowledge graph question answering (Luo et al., 2024; Huang et al., 2024), open-domain dialogue generation (Wang et al., 2024), and multi-hop question answering (Trivedi et al., 2022). Some studies have focused on investigating how LMs use the input context, identifying discrepancies in how different parts of the input are processed. For instance, Mallen et al. (2023) find that retrieval may sometimes harm model predictions and Liu et al. (2024) report that LMs suffer from a "lost in the middle" predicament. However, these works focus on simpler, one-hop question-answering tasks that do not require reasoning across distant contextual hops. Going a step further, Chen et al. (2024)

[1] In the context of MHQA, they usually refer to the documents needed to answer the single-hop questions into which the original multi-hop question is decomposed.

demonstrate that the order of the premise matters when logical reasoning is expected by LMs. Shen et al. (2024) present GeAR, which advances RAG performance for multi-hop QA through graph expansion and an agent framework that incorporates graph expansion, achieving the state-of-the-art performance of established benchmarks for multi-hop QA. Around the same time as our work, Baker et al. (2024) also identify the "lost in between" issue of long context LLMs. In this work, we systematically explore how language models reason over their input context to address multi-hop questions.

2.2 Drawbacks of Causal Language Model

A causal language model is equipped with a causal mask that prevents tokens in the context to *see* future context (Raffel et al., 2020). This design harms the performance on complex tasks that require rich contextualized representations (Li et al., 2023; Qorib et al., 2024). A common way to mitigate this issue is to prompt the model by repeating the context in the input. Xu et al. (2024) show that repeating the context improves the LMs ability in reasoning. Springer et al. (2024) prove that repeating the context can also improve embedding quality. To sidestep the unnecessary cost associated with repeating the input context, researchers have started to introduce bi-directional attention into decoder-only models. BehnamGhader et al. (2024) and Muennighoff et al. (2024) successfully apply bidirectional attention in decoder-only models to generate text embeddings, and observe better quality in a variety of tasks. In this work, we explore how bi-directional attention can enhance language models' performance in the MHQA task, uncovering valuable heuristics that could improve model outcomes.

3 Preliminary

3.1 Multihop Question Answering

We formally define the MHQA task $\mathcal{T}$. The input of $\mathcal{T}$ has two parts, a question q, and n documents $\mathcal{D} = \{d_1, \ldots, d_n\}$, some of which are relevant to q and some are not. For answering q, the information from m documents, s.t. $m < n$, is mandatory. We accomplish $\mathcal{T}$ by prompting LMs with the concatenation of q and $\mathcal{D}$ in the input context.

3.2 Grouped Attention Weight

For better investigating which documents are more *heavily attended* by LMs, we compute grouped attention weights between token blocks. There are several blocks, including the instruction block, document blocks, question block, and prediction blocks. Please note that for prediction blocks, we directly use the prediction tokens that are not grouped, since the prediction blocks may contain other tokens besides the answer tokens. For attention between block X and block Y, the grouped attention weight is computed as:

$$\text{GA}_{l,h}(X, Y) = \frac{1}{|X|} \sum_{t_X \in X} \sum_{t_Y \in Y} \text{Attention}_{l,h}(t_X, t_Y) \quad (1)$$

where l, h denote the decoder layer and the attention head, $|X|$ is used to normalize grouped attention values, t_X and t_Y are tokens of block X and Y. By grouping attention with (1), we make sure:

$$\sum_{Y} \text{GA}_{l,h}(X, Y) = 1 \quad (2)$$

Using Eq. (1), we can understand which context part contributes more to the next token prediction.

3.3 Information Contribution Score

To investigate how much information from each document is captured during the MHQA task, we introduce the Information Contribution (IC) score based on the grouped attention scores.

For layer l, document token group d, the Information Contribution (IC) score is by defined:

$$\text{IC}_l(d) = \frac{1}{|A||H|} \sum_{h \in H} \sum_{a \in A} \text{GA}_{l,h}(a, d) \quad (3)$$

where H is the set of attention heads and A is the set of answer tokens in the prediction.

4 Experimental Setup

We conduct experiments using both encoder-decoder LMs (from the Flan-T5 family) and causal decoder-only LMs (Qwen 2.5 family and Llama 3.x family). For the Qwen 2.5 family, we select to use five LMs with sizes that range from 0.5B to 14B. For the Llama 3.x family, we use 1B and 3B from Llama 3.2 and 8B from Llama 3.1. The specifications of these models are included in Table 4. We use the instruction-tuned variants of all the models included in this study. We investigate 4 setups for inference:

Answer Only The model is forced to directly generate the answer in the following format: `\box{⟨answer⟩}`.

CoT Zero-shot Chain of Thought prompting is used to ask the model to first generate reasoning steps and then provide the final answer in the same format: `\box{⟨answer⟩}`.

Finetuned We use the MuSiQue training set to train the models.

Finetuned + Bi We replace the original causal mask of the model with bi-directional attention facilitated by a 2D mask, as follows:

$$M_{i,j} = \begin{cases} 0 & \text{if } i \geq j \\ 0 & \text{if } i \leq c, j \leq c \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

where c is the context length. With the new mask, the model is then converted to a prefix LM (Raffel et al., 2020) The model is trained with the same data as in the **Finetuned** setup.

For all finetuning experiments, we use LoRA with $r = 8$ and $\alpha = 16$ and train for 5 epochs with a learning rate $2e-5$ and a batch size 1.

4.1 Dataset

We instantiate the MHQA task with the MuSiQue (Trivedi et al., 2022) dataset, which contains multi-hop questions from 2-hop to 4-hop. In the original dataset, for each question, $2-4$ gold documents are provided, each containing evidence for each decomposed question (hop). Additionally, distractor documents are included to add noise, forming a context with up to 20 documents in total, presented in no specific order. We use the answerable set in all the experiments, and keep the data split unchanged in terms of the training and development set. For the finetuning experiments, we use the original training set with 19,938 queries. For all the experiments, we report performance on the development set which consists of 2,417 queries.

4.2 Metrics

We use the accuracy metric (**Acc**) to measure MHQA performance. For the **Answer Only** setup, we treat the answer inside `\boxed{⟨answer⟩}` as the prediction. For **CoT** setup, we treat the last `\boxed{⟨answer⟩}` as the prediction. Please note that some models do not follow the CoT instruction well[2]. In such cases, we compute **Acc** by finding if the reference answer is included in the last line of the prediction. For finetuning-based models and encoder-decoder models, we compute the exact match as the accuracy of the prediction.

[2]Small Qwen models (0.5B and 1.5B) generally do not apply CoT reasoning. Llama models do not always follow the instruction to place the final answer in `\boxed{⟨answer⟩}`.

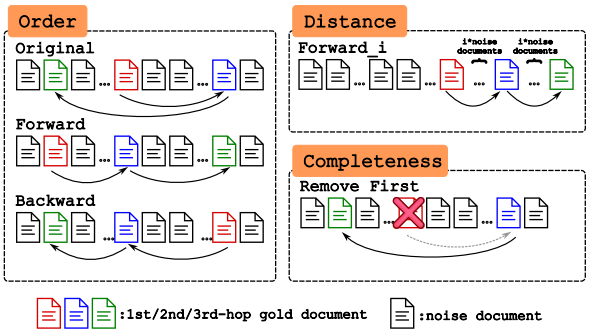


Figure 2: Context permutation design. Take 3-hop question for example. Forward setup follows the reasoning chain while Backward setup is the reverse of Forward setup. Forward_i control the distance between gold documents, and Remove First removing the document that support first hop question.

5 Does Permutation Change LM's Mind?

We designed three distinct context permutation settings to assess LMs' abilities to comprehend different aspects of the context, as shown in Figure 2.

Order of gold documents In an idealised scenario, we would expect the gold documents to be provided in the same order as the reasoning hops. Our hypothesis is that the effect of the causal mask will be minimised by providing the gold documents with the reasoning chain order. We refer to this setting as **Forward**. In contrast to this setting, we test a setting in which the documents are placed in the reverse order of Forward, hypothesising that the counter-intuitive reasoning consistency should make the task more challenging for the involved LMs. The reversed setting is called **Backward**. Finally, we have a setting where we keep the order of the gold documents as in the original dataset, to which we refer as **Original**.

Distance of gold documents Besides the order, the distance between the input gold documents is not guaranteed in real-world MHQA tasks. To investigate how the distance between gold documents affects the LM, we design a series of **Forward**_i settings, where we fix the order of documents to be Forward, and, subsequently, ensure that the final hop document is placed at the end of the context. After that, between each gold document, we inject i noise documents (i.e. the set of noisy documents remain the same as in the previous settings). We

select $i = \{0, 1, 2, 3, 4, 5\}$, where $i = 0$ stands for no noise between gold documents.

Completeness of gold documents To better understand how LMs answer multi-hop questions, we want to know if they are really doing multi-hop reasoning, or just *guess* an answer. To evaluate this, we design a setting in which the knowledge in the input context is incomplete in terms of answering the original question: **Remove First**, which removes the first hop document.[3]

5.1 Order of Documents Matters

Table 1 shows performance across different setups.

Non-finetuned Generally, the **Answer Only (AO)** results in the worst performance. In this setup, changing the order of the supporting documents leads to unstable performance differences, where forward-placed gold documents do not offer overall performance improvement consistently. With CoT prompting, most models get performance improvements. We note that Qwen2.5 0.5B, 1.5B and Llama3.2 1B do not follow the CoT instruction well, as the inclusion of the CoT instruction does not often lead to changes in the models' output. Overall, decoder-only models with zero-shot CoT perform similarly to when they are directly prompted, and the order of documents appears to be invariant to their MHQA performance.

Finetuned (FT) By finetuning the models on the MuSiQue training set, the performance of all models improves. Interestingly, finetuned variants seem to benefit from the forward gold document setting, even though the training data are provided in the original order of the training split. To ensure that the order of the documents in the training set is not sequentially correlated with the forward (and backward) setting, we compute the average Spearman's rank correlation and Kendall's τ coefficients: 0.0013 (-0.0013) and 0.0016 (-0.0016) respectively, indicating that the benefits of the forward setup are *emergent* through finetuning.

Finetuned + Bi (FT+Bi) When finetuning the models modified with bi-directional attention, we observe further performance improvements. Moreover, the models are more robust to document permutations, showing less variation in performance across the three inference setups.

[3]Please note that, in a formal definition, removing an element does not qualify as context permutation. We use the term "permutation" here for convenience in this context.

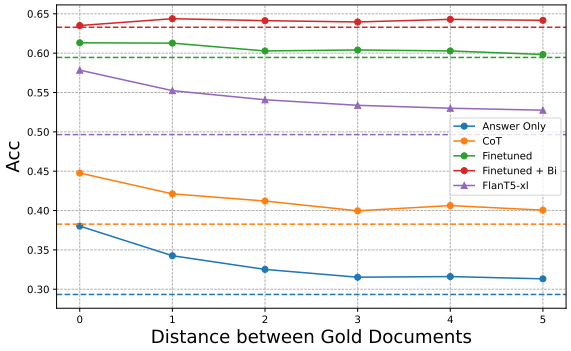


Figure 3: Acc of Qwen2.5 7B across different gold document settings. Each dashed line refers the Forward result of the setup with the same colour. The performance of non-finetuned models generally drops as the distance increases, while finetuned models show better robustness to the increase of distance.

Encoder-Decoder Table 2 shows the results from the non-finetuned encoder-decoder models. Generally, the Flan T5 family performs much better than non-finetuned decoder-only models for approximately similar parameter numbers. Flan T5 xl with 3B parameters already outperforms all decoder-only models under 8B with Answer Only or CoT, and achieves competitive performance compared to Qwen 2.5 14B. Interestingly, we can observe a very clear trend that the forward setting performs the best while the backward one performs the worst. The trend is clearer than in the Finetuned + Bi setting of the decoder-only models. When we tested other encoder-decoder models, the same trend was not observed, while the involved models continue to outperform equally-sized decoder-only models (see Table 5 in Appendix). We believe that this emerging ability of the Flan T5 models can be attributed to the selection of data used for their training (Longpre et al., 2023).

5.2 Distance between Documents Matters

We explore how the distance between gold documents affects the LMs' performance on MHQA using Qwen2.5 7B and FlanT5 xl. The results are shown in Figure 3. Generally, the models' performance drops as the distance of the gold document increases. Notably, placing forward-ordered documents on the last positions of the input context brings significant performance improvement. This is because LMs generally favour documents close to border regions of the input prompt instead of the middle regions (Liu et al., 2024). Notably, finetuned models (Finetuned and Finetuned + Bi) re-

Model	Answer Only			CoT			Finetuned			Finetuned + Bi		
	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F
Qwen2.5 0.5B	0.21	8.94	-0.58	-0.21	12.91	-0.79	-4.34	27.14	3.72	-0.91	30.30	0.41
Qwen2.5 1.5B	-1.08	20.36	-0.70	0.0	22.76	-1.03	-2.61	44.06	1.86	0.04	44.78	1.20
Qwen2.5 3B	-0.87	19.78	0.74	-2.03	24.82	-0.46	-1.37	50.23	2.98	-1.74	52.15	1.70
Qwen2.5 7B	-1.78	28.59	0.74	-2.03	36.24	2.03	-2.36	58.05	1.41	-1.45	62.96	0.33
Qwen2.5 14B	0.12	37.07	-0.29	1.08	39.22	0.62	-1.65	64.34	1.03	0.29	64.88	0.08
Llama3.2 1B	0.83	11.21	-0.83	-0.04	11.96	-0.21	-1.99	33.06	1.74	-0.41	40.85	0.62
Llama3.2 3B	-1.61	25.73	0.79	-0.62	31.65	1.12	-1.99	54.57	1.70	-0.91	59.60	0.58
Llama3.1 8B	0.54	36.37	-0.95	-0.62	44.60	-0.25	-2.11	63.51	1.24	-1.20	65.48	1.41

Table 1: Overall MHQA performance on the MuSiQue development set. Δ_B and Δ_F are performance differences between original documents and re-ordered backward and forward documents respectively. Green cells indicate performance improvement while red cells indicate performance drop.

Model	Δ_B	Acc	Δ_F	Qwen2.5 Acc
FT5 small/80M	-1.94	20.11	1.86	8.94 (0.5B)
FT5 base/250M	-1.65	28.09	1.90	20.36 (1.5B)
FT5 large/0.8B	0.25	40.01	0.37	19.78 (3B)
FT5 xl/3B	-2.44	47.33	2.19	28.59 (7B)
FT5 xxl/11B	-2.03	56.43	1.65	37.07 (14B)

Table 2: MHQA performance on the MuSiQue development set using Flan T5 models. Qwen2.5 Acc is the Acc score from the Qwen2.5 family for reference.

main more robust since their performance is less affected by both (i) the increase of distance between gold documents and (ii) the placement of forward-ordered documents at the last positions.

Our findings indicate that in a multi-iterative RAG setting ordering documents based on relevance rather than the order of their associated decomposed question (assuming that more than a single document is maintained for each decomposed question), can reduce the distance between relevant documents and effectively increase the end-to-end QA performance. As such, we emphasize the importance of incorporating ranking-based metrics to measure retrieval quality, which currently deviates from the standard practices in RAG-based MHQA, where the focus is primarily on recall (R@n) (Trivedi et al., 2023; Gutierrez et al., 2024).

In Appendix H, we include additional experiments on the 2WikiMultihopQA `Compositional` subset. The findings are in line with the major observations above in terms of the distance and order of the gold documents.

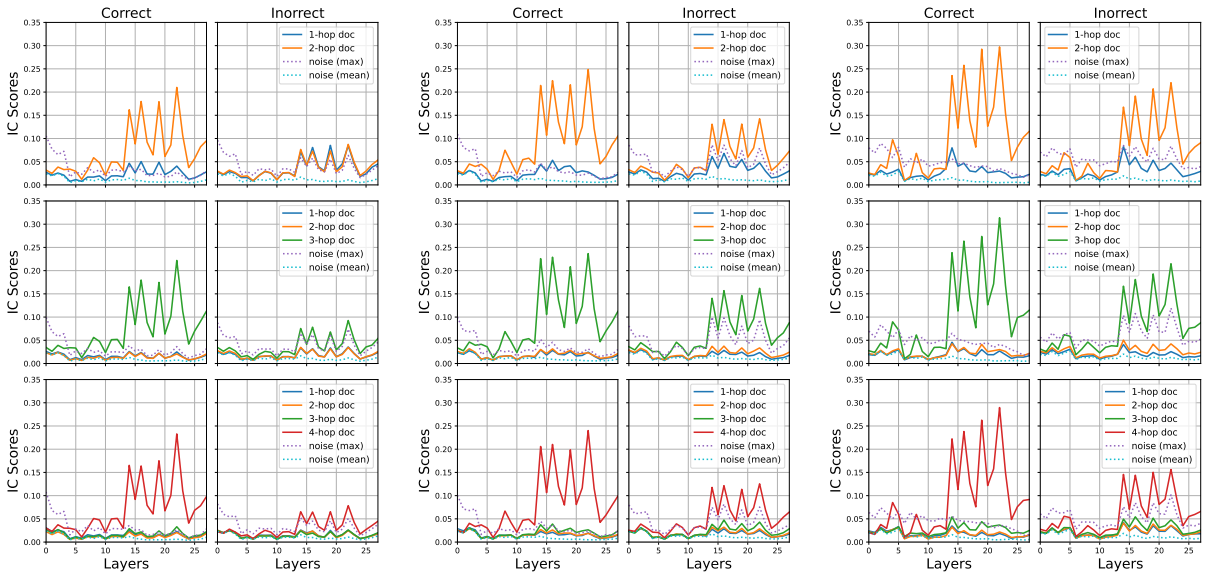
5.3 Do LMs guess the answer?

To identify if LMs can answer questions even when they do not have enough information, we manually remove the first hop document from the context. Since LMs may have the relevant parametric knowledge already, we further design a simple atomic by asking them the first hop question, without any other context. We determine whether they have the required parametric knowledge by evaluating their answer. Table 3 shows the relevant results.

Setup	Accuracy		
	2-Hop	3-Hop	4-Hop
w/o first hop information in parametric knowledge			
AO	26.0→20.8	28.6→29.1	29.7→28.8
CoT	40.3→14.8	38.5→27.1	24.3→24.3
FT	63.9→36.2	57.7→54.4	56.2→57.2
FT+Bi	70.9→37.7	57.0→55.7	64.9→61.7
FT5 xl	53.8→39.1	41.3→39.7	40.1→40.1
w/ first hop information in parametric knowledge			
AO	33.6→33.3	30.8→29.0	22.8→22.8
CoT	40.2→30.0	30.8→39.7	20.7→23.9
FT	57.5→52.2	40.7→41.6	54.3→54.3
FT+Bi	62.1→52.2	39.7→38.3	75.0→78.3
FT5 xl	51.1→46.8	36.6→53.7	39.3→42.9

Table 3: Evaluation results of completeness permutation of Qwen2.5 7B (AO, CoT, FT and FT+Bi) and FlanT5 xl model (FT5 xl). Results are shown as **Original→Remove First**.

Without first-hop information in parametric knowledge For 2-hop questions, since the first hop document is more important, generally all models across all settings drop in accuracy when the information is not stored in their parameters. For 3-hop and 4-hop questions, we see a similar performance drop on non-finetuned models, but the drop is less significant. This means for complex questions, LMs still lack the ability to know what they don't know, and they are expected to refuse to answer the question since the reasoning path of existing evidence is not complete. For finetuned models, the issue is more severe, where the accuracy even increases (4-Hop, FT) as key information



(a) Answer Only (b) Finetuned (c) Finetuned + Bi

Figure 4: IC distribution across different layers of Qwen2.5 7B with different setups for 2-hop, 3-hop and 4-hop questions, all in original order. LMs generally assign higher peak IC scores (in particular for the final hop) when answering the multi-hop questions correctly.

is not provided.

With first-hop information in parametric knowledge Similarly, for 2-hop questions, accuracy drops in most settings, but it is not as significant as when the removed information is known by the model. While for more complex questions, performance generally increased, especially for the CoT and Flan T5 models. Even when we ask models to only use the given context, they still default to parametric knowledge for MHQA. This also suggests that complex MHQA is still a big challenge to LMs, where failure to locate all required information in the context may lead to hallucinations. This also indicates that retrieval may sometimes harm models' performance, even if the external knowledge does not conflict with parametric knowledge.

6 Are LMs Aware of Context Permutations?

To investigate how LMs make use of the given context to answer multi-hop questions, we compute the information contribution (IC) score with Equation (3) on the MuSiQue development set. Our hypothesis is that by investigating how models *pay attention* to particular documents in the context, we can better explain their behaviour in answering multi-hop questions. In this part, we only consider three settings: **Answer Only**, **Finetuned** and **Finetuned + Bi**. Figure 4 shows the IC distribution of the different setups across different layers of the Qwen 7B model.

6.1 LMs Assign Higher Peak Attention when Correct

For the samples that are answered correctly, the model consistently assigns the largest attention score to the last hop document. This is intuitive since the answer is included in the last hop document. In the case of samples that are answered incorrectly, we observe a smaller gap between the last hop and previous hop documents. Notably, while correct samples always assigned higher attention scores to the gold documents, the IC score of the Answer Only setup with direct prompting is generally lower than the two settings based on finetuning. This is even clearer for the incorrect samples, where, for the non-finetuned variants, the highest IC score of the noise documents (labelled as noise (max)) is almost the same as that of the gold documents. Interestingly, the "retrieval" layers to which most of the fluctuations in the attention weight are attributed remain consistent throughout fine-tuning, even with bi-directional attention.

We find that all models assign much higher attention to the last hop when generating the correct answer, but are less confident and assign more evenly distributed attention scores when generating incorrect answers. More generally, all models assign higher peak attention (i.e., largest attention

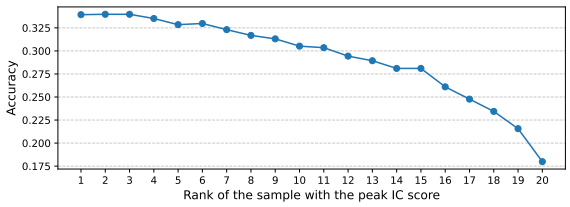


Figure 5: Qwen 2.5 7B performance (Answer Only) peak IC ranking from higher peak IC to lower peak IC. A clear trend can be observed that higher peak IC among the 20 random shuffles brings higher accuracy.

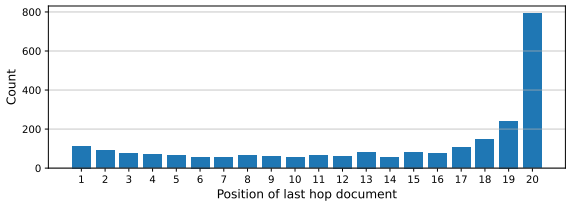


Figure 6: Position of last hop document when the order achieves highest peak IC score in Qwen 7B model (Answer Only). LMs generally assign highest peak IC when the last hop document is close to the end.

to one document in the context) when they predict the answer correctly. To better understand the relation between model predictions and the largest IC score, we randomly shuffle the document order in the context 20 times per question, then perform inference with Qwen2.5 7B in the Answer Only setup. We observe that the median peak IC scores are 2.22 and 1.72 for correct and incorrect samples (cf. Figure 7). IC scores are in general higher for correct than incorrect samples.

In addition, we compute the prediction accuracy with different context shuffles based on the ranked peak IC scores (Figure 5). Our results show that a higher peak IC score among the 20 random shuffles provides higher MHQA performance, showcasing that the peak IC score is a key signal from the LM to identify the optimal context order. This finding underscores the importance of context order for the MHQA task, where the best context order almost doubles performance over the worst order.

6.2 LMs Generally Favour the Last Document

Amongst all noise documents, we capture the most favoured position where the IC score is the highest. We find that the last document receives the most attention in most cases. In Figure 4, the purple dashed line shows the attention assigned to the noise that confuses the model most. Interestingly, we find this is mostly the last document, which seems to have a high probability of being captured in the lower layer of the non-finetuned model. In Figure 4, there is always a local peak of noise document at early layers.

In addition, in all the best samples (with the largest IC score among 20 randomly shuffled contexts) from Section 6.1, we also observe a particular preference for the last document (cf. Figure 6). A similar phenomenon is observed in other causal decoder-only models, even for models like Qwen2.5 0.5B which performs worst on forward order, the trend exists whilst less significant.

7 Discussion

7.1 Optimizing LMs' Use of Context

State-of-the-art RAG methods (Trivedi et al., 2023; Gutierrez et al., 2024) for solving multi-hop question generally split the n-hop complex question q into several decomposed questions q_i and retrieve top-k documents $D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,k}\}$ accordingly. Simply concatenating the documents with the order of decomposed questions ensures that the document order is forward, but the distance between relevant documents is large. According to our observations in Section 5.1 and 5.2, non-finetuned causal language models are not sensitive to the relevant documents' order, but the distance between relevant documents matters. Thus if the selected reader is an off-the-shelf causal language model, gathering additional documents, *ensuring that their in-between distance is minimum*, is essential. On the contrary, if finetuning is an option, then keeping the forward order is more important for getting the best performance. For both situations, it is strongly recommended to place documents with higher relevance close to the end of the context.

7.2 Use Reader with Bidirectional Attention

Causal decoder-only LMs are widely used in the RAG framework as readers. In our experiments, we show that these models are not the ideal choice as MHQA solvers, as they are limited by their causal mask. Our experiments show that by altering the causal mask with a prefix mask, and simply using LoRA finetuning to obtain a non-causal decoder-only model, we can outperform the original causal setup while being more robust against the order of gold documents. Notably, the FlanT5 family shows marvellous off-the-shelf MHQA ability and can serve as a competitive reader alternative to causal decoder-only models within the RAG framework.

7.3 Attribution Is Important

From Section 5.3, we observe that removing the first hop document causes a performance drop to all language models, but still maintain a relative high accuracy, especially for 3-hop and 4-hop questions. In addition, we find that the attention weight assigned to the last hop document does not change when removing the first hop document (Figure 10). These findings further underscore the importance of attribution in RAG, to ensure that the predicted answer is supported by evidence from the context. In knowledge intensive tasks, the ignorance of evidence completeness could be a critical issue that produces hallucinations.

8 Conclusion

In this study, we explore the MHQA capabilities of LMs with different architectures. We find that non-finetuned, causal decoder-only LMs are invariant to the order of gold documents but are affected by their in-between distance in the input context. Incorporating bi-directional attention enhances performance—both in the case of encoder-decoder models, which outperform their decoder-only counterparts and when applied to decoder-only models. Additionally, finetuning instils in the models the bias of forward order and makes them more robust against the distance between gold documents. Our analysis of attention distribution indicates that increased peak attention in the context aligns with accurate predictions. These insights advance our understanding of LMs in MHQA and suggest avenues for future improvements.

Acknowledgments

This work is supported by UKRI (grant number EP/S022481/1), The University of Edinburgh and Huawei's Dean's Funding (C-00006589).

Limitations

In this work, with the limitation of computing resources, we consider MuSiQue dataset with its original setup, with at most 20 documents per question. To the best of our knowledge, MuSiQue is one of the most challenging datasets for MHQA, and, therefore, ideal for the experiment we want to conduct in this study. In Appendix H, we include additional results on relevant subsets of 2WikiMultihopQA, demonstrating that our original findings on MuSiQue hold.

Most of our prompts are around or below 4k tokens, which is relatively a short setting with respect to the current long-context language models. We believe that this does not diminish the contribution of this work. For example, we have already noticed a significant effect of distance between gold documents with only 5 noise documents in between, indicating that when considering longer contexts, the context order is even more important for complex reasoning tasks, such as MHQA, which is a crucial issue that needs to be considered in future works.

References

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. *Preprint*, arXiv:2412.10079.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jie He, Simon U, Victor Gutierrez-Basulto, and Jeff Pan. 2023. BUCA: A binary classification approach

to unsupervised commonsense question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 376–387, Toronto, Canada. Association for Computational Linguistics.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z. Pan. 2025b. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *Knowledge-Based Systems*, page 113648.

Wenyu Huang, Guancheng Zhou, Hongru Wang, Pavlos Vougiouklis, Mirella Lapata, and Jeff Z. Pan. 2024. Less is more: Making smaller language models competent subgraph retrievers for multi-hop KGQA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15787–15803, Miami, Florida, USA. Association for Computational Linguistics.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

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

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Linhao Luo, Yuan-Fang Li, Reza Haf, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A survey of conversational search. *Preprint*, arXiv:2410.15576.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Zhili Shen, Chenxin Diao, Pavlos Vougiouklis, Pascual Merita, Shriram Piramanayagam, Damien Graux, Dandan Tu, Zeren Jiang, Ruofei Lai, Yang Ren, and Jeff Z. Pan. 2024. Gear: Graph-enhanced agent for retrieval-augmented generation. *Preprint*, arXiv:2412.18431.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *Preprint*, arXiv:2402.15449.

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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *Preprint*, arXiv:2401.13256.

Hongru Wang, Wenyu Huang, Yufei Wang, Yuanhao Xi, Jianqiao Lu, Huan Zhang, Nan Hu, Zeming Liu, Jeff Z. Pan, and Kam-Fai Wong. 2025. Rethinking stateful tool use in multi-turn dialogues: Benchmarks and challenges. *Preprint*, arXiv:2505.13328.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yike Wu, Yi Huang, Nan Hu, Yuncheng Hua, Guilin Qi, Jiaoyan Chen, and Jeff Z. Pan. 2024. CoTKR: Chain-of-thought enhanced knowledge rewriting for complex knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3501–3520, Miami, Florida, USA. Association for Computational Linguistics.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. 2024. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15549–15575, Miami, Florida, USA. Association for Computational Linguistics.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *Preprint*, arXiv:2402.18013.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 50117–50143. Curran Associates, Inc.

A Experiment Environment

All the experiments mentioned in this paper are undertaken on NVIDIA A100 80GB GPUs. To make

Model	Params	Layers	Dim	Heads
Qwen2.5 0.5B	0.49	24	896	14/2
Qwen2.5 1.5B	1.5	28	1,536	12/2
Qwen2.5 3B	3.1	36	2,048	16/2
Qwen2.5 7B	7.6	28	3,584	28/4
Qwen2.5 14B	14.7	48	5,120	40/8
Llama3.2 1B	1.23	16	2,048	32/8
Llama3.2 3B	3.21	28	3,072	24/8
Llama3.1 8B	8.03	32	4,096	32/8
Flan T5 small	0.08	8+8	512	6
Flan T5 base	0.25	12+12	768	12
Flan T5 large	0.8	24+24	1,024	16
Flan T5 xl	3	24+24	2,048	32
Flan T5 xxl	11	24+24	4,096	64

Table 4: Specification of LMs experimented in this paper.

sure that our experiments are replicable, we utilize greedy decoding for all the inference experiments and set the seed to 42 for all finetuning experiments. We use Huggingface Transformers library (Wolf et al., 2020) to accomplish all the experiments in this study, and utilize PEFT (Mangrulkar et al., 2022) for LoRA finetuning.

B Model Specification

Table 4 shows the specifications of LMs investigated in this study.

C Prompt Details

The prompt used in this work is adapted from Liu et al. (2024) as:

> Answer the question using only the provided search results (some of which might be irrelevant).
>
> ⟨Documents⟩
>
> Question: ⟨Question⟩

Specially, for **Answer Only** setup, we explicitly add `\boxed{` to the end of the prompt to force model following the answer only instruction.

D Detailed statistics of peak IC score

As discussed in Section 6.1, we observe that LMs assign higher peak IC scores when they get the correct answer. Figure 7 shows the box plot of peak IC score when correct and incorrect answers are predicted, with 20 random shuffles for each question.

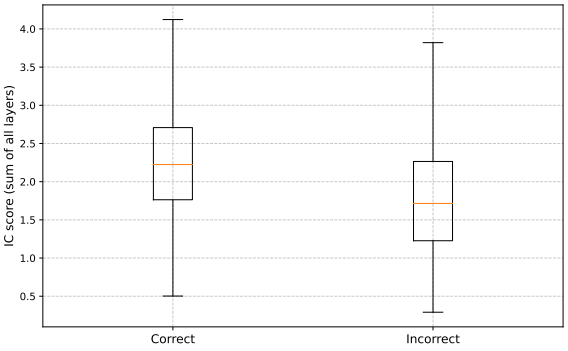


Figure 7: Statistics of peak IC score of correct and incorrect samples.

E Extra analysis of IC plots

E.1 Model behave differently with Backward and Forward

In this section, we illustrate other findings observed in the IC plots. From Figure 8, we can find that in the Forward setup, the model tends to focus more on the first hop document. For instance, in the IC distribution on 4-hop questions of Figure 8a and 8b, the blue plot of attention focusing on a 1-hop document in Forward setup is the highest except the last hop, while in Backward setup the differences between these hops are minor. This behaviour also generalizes to finetuned models. Moreover, this behaviour is mainly observed in 3-hop and 4-hop questions, where the distance between the first-hop document and last-hop document is longer. Though last hop document is able to encoding part of the information of first hop document, longer distance will reduce this encoding and requires model to focus more on it to prevent "forgetting".

E.2 How does distance change the LM's mind?

Figure 9 shows the IC distribution across layers, considering different setups and distances (i.e., i in Forward_i settings) between gold documents. Specifically, when $i = 5$, the model behaves similarly as discussed in Section E.1: the first hop document gets the most attention except the last hop, but when $i = 0$, the attention intensity starts to follow the order of the reasoning hops, which decreases from the last hop to the first hop. This finding supports our hypothesis in Section E.1, suggesting that a longer distance compels the model to allocate more attention to the first-hop document to ensure essential information is retained when gold documents are forward-placed.

Model	Δ_B	Acc	Δ_F
T0	0.29	38.06	-0.87
T0pp	-0.63	43.07	-0.29

Table 5: Performance of other encoder-decoder models on the MuSiQue development set.

E.3 How does removing the first hop change the LM's mind?

From Figure 10, we can find that while 2-hop samples are most affected, removing the first hop document does not affect much the attention weight assigned to the last-hop documents. For 3-hop and 4-hop questions, the IC distribution of the rest documents is nearly the same as when not removing the first hop document. These findings again enhance the importance of attribution of retrieval-augmented generation, to ensure that the predicted answer is supported by the evidence from the context. In knowledge-intensive tasks, the ignorance of evidence completeness could be a critical issue that produces hallucinations.

F Experiment on other encoder-decoder models

This section shows more experiment results of instruction-based encoder-decoder model. T0pp is a 11B model instruction finetuned from T5 11B with P3 datasets (Sanh et al., 2022), where T0 is a weaker version that trained with P3 subset. Noted that the favour of forward order disappeared. Given that P3 is a subset of Flan dataset (Longpre et al., 2023), we then consider that the ability is not obtained from model architecture, but from training data. Future works about which dataset from Flan triggered the favour could be established for a better understanding of the behaviour.

G Closed-book Result

MHQA task is much more challenging than traditional QA tasks, where LMs can not well accomplish it without external knowledge. Here we accomplish a closed-book experiment setup with no context information provided to the LMs discussed in this work. Table 6 shows the evaluation results.

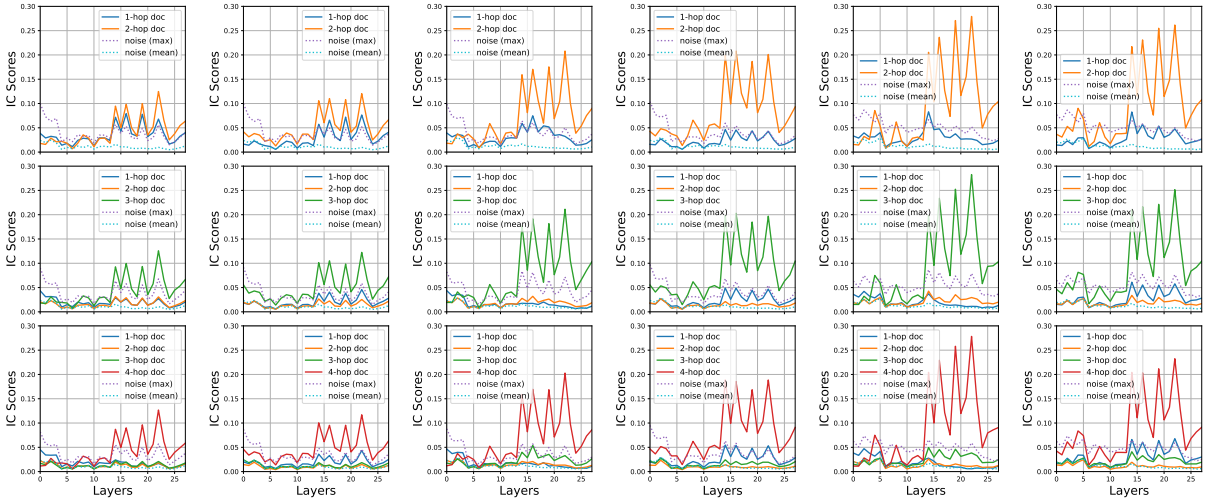
H Experiment Result on Other MHQA dataset

To evaluate if our findings generalize to other MHQA datasets, we run extra experiments on 2WikiMultihopQA dataset (Ho et al., 2020). We use 5,234 questions from the `compositional` subset and 1,549 questions from the `inference` subset, in which all the questions are 2-hop, and the context contains 10 short documents, of which 2 gold documents are provided. To investigate if finetuning generalizes to other datasets, we use the same models finetuned on MuSiQue. In this part, we evaluate LMs including Flan T5 xl, Flan T5 xxl, Qwen 2.5 7B, Llama 3.1 8B, as well as their finetuned and finetuned with bi-directional attention variants if applicable.

Table 7 and 8 shows the evaluation results on 2WikiMultihopQA dataset. It is clear that the favour of forward document still exists in the two sets, and the use of bidirectional attention boosts performance significantly while being more robust with the order of gold documents. In addition, even though not finetuned on 2WikiMultihopQA, the two finetuned setup still obtain competitive performance. Moreover, compared to the finetuned setup with causal attention mask, the bi-directional attention get a better performance, and more robust to the order of gold documents.

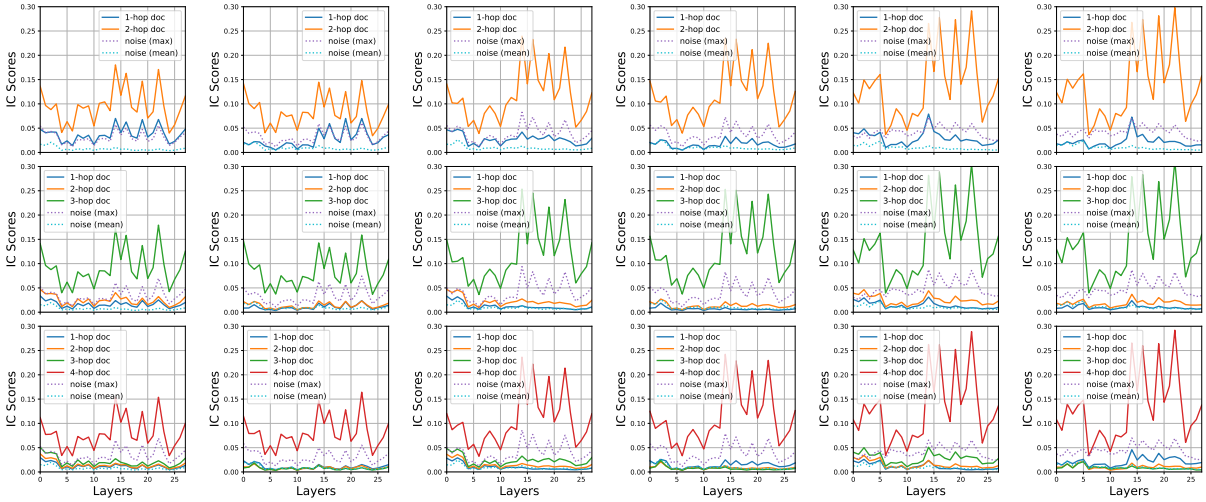
Figure 11 shows the affect from distance on Flan T5 models and Qwen 2.5 7B variants. It is clear that our observation still holds that finetuned models are more robust to the distance of gold document, even not finetuned on the same dataset. While the performance of non-finetuned models decreases as the distance increases, even when the context is much shorter (with shorter documents and only 10 in the context).

Similar to Section 6.1, on the two subset of 2WikiMultihopQA dataset, we also conduct experiment on Qwen 2.5 7B with Answer Only setup with randomly shuffled the document order of each question 10 times and computed the peak IC score for each shuffle. For the `compositional` subset, the average accuracy across these 10 shuffles was 42.14, while the sample with the highest peak IC score achieved an accuracy of 49.29. While for the `inference` subset, the average accuracy across these 10 shuffles was 14.96, while the sample with the highest peak IC score achieved an accuracy of 19.43.



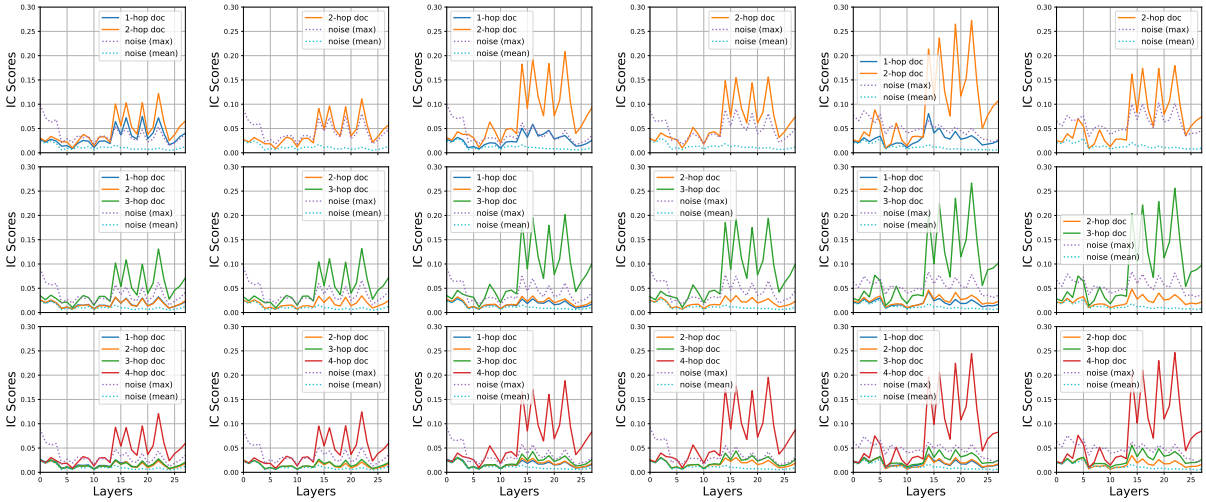
(a) AO, Backward (b) AO, Forward (c) FT, Backward (d) FT, Forward (e) Bi, Backward (f) Bi, Forward

Figure 8: IC distribution across different layers of Qwen2.5 7B with different order of gold document.



(a) AO, i=0 (b) AO, i=5 (c) FT, i=0 (d) FT, i=5 (e) FT+Bi, i=0 (f) FT+Bi, i=5

Figure 9: IC distribution across different layers of Qwen2.5 7B with different distances i between gold documents.



(a) AO, w/ 1 hop (b) AO, w/o 1 hop (c) FT, w/ 1 hop (d) FT, w/o 1 hop (e) FT+Bi, w/ 1 hop (f) FT+Bi, w/o 1 hop

Figure 10: IC distribution across different layers of Qwen2.5 7B with and without first hop document.

Model	Flan T5					Qwen 2.5					Llama 3.x		
	small	base	large	xl	xxl	0.5B	1.5B	3B	7B	14B	1B	3B	8B
Acc	0.54	1.08	2.19	3.10	3.97	2.65	2.48	5.46	8.32	10.18	2.77	5.59	10.88

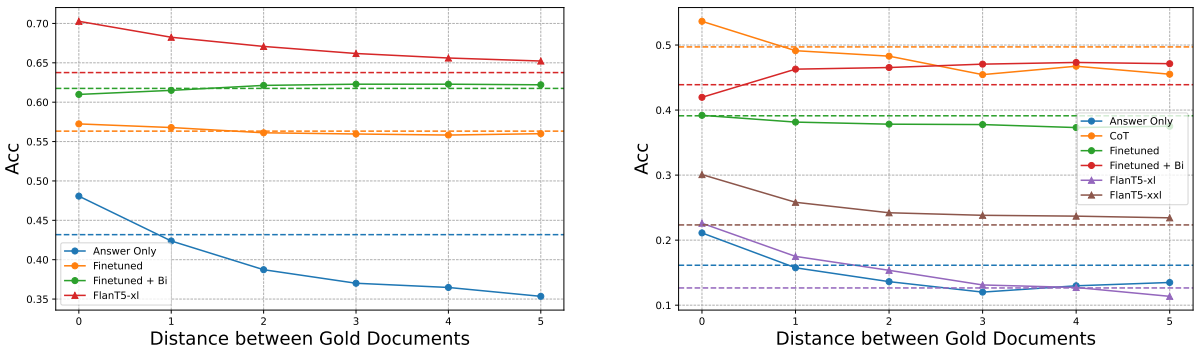
Table 6: Closed-book evaluation results on MuSiQue development set.

Model	Answer Only			Finetuned			Finetuned + Bi		
	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F
Qwen 2.5 7B	-0.44	42.19	0.99	-1.78	54.78	1.55	-0.25	61.04	0.71
Llama 3.1 8B	3.40	51.60	-3.19	-0.99	58.98	0.42	-0.02	62.84	-0.06
Flan T5 xl	-1.38	63.76	1.60	-	-	-	-	-	-

Table 7: MHQA performance on the 2WikiMultihopQA `Compositional` development subset. Δ_B and Δ_F are performance differences between original documents and reordered (backward and forward) documents. Green cells indicate performance improvement while red cells indicate performance drop.

Model	Answer Only			CoT			Finetuned			Finetuned + Bi		
	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F	Δ_B	Acc	Δ_F			
Qwen 2.5 7B	-1.42	13.82	2.39	-1.16	49.52	0.19	-6.00	33.25	5.94	-1.87	40.99	2.97
Llama 3.1 8B	-1.03	24.47	-0.39	0.52	59.85	-0.77	-2.13	46.35	3.74	-1.36	56.10	0.32
Flan T5 xl	-0.13	13.69	-0.58	-	-	-	-	-	-	-	-	-
Flan T5 xxl	-3.16	21.11	2.52	-	-	-	-	-	-	-	-	-

Table 8: MHQA performance on the 2WikiMultihopQA `Inference` development subset. Δ_B and Δ_F are performance differences between original documents and reordered (backward and forward) documents. Green cells indicate performance improvement while red cells indicate performance drop.



(a) `Compositional` (b) `Inference`

Figure 11: Distance results on 2WikiMultihopQA development set.