CompQA: Investigating the Weakness of Multihop QA on Comparison Questions

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) models are increasingly employed in Multihop Question Answering (MHQA). However, we identify a critical limitation: existing methods exhibit suboptimal performance on comparison-type questions, with the performance decline being notably greater than that observed for bridge-type questions. Empirical analysis reveals that existing methods consistently underperform relative to LLM-only baselines, particularly as the number of hops increases. Moreover, they require significantly more inference and retriever calls without delivering equivalent performance gains. To demonstrate, we introduced the CompQA dataset, which includes questions with a higher number of hops, alongside the MuSiQue benchmark. Finally, we discuss our findings, examine potential underlying causes, and highlight the limitations of RAG strategies in reasoning over complex question types.

1 Introduction

MHQA requires sequential inference by aggregating evidence from multiple sources, unlike single hop QA that relies on a single piece of evidence. This makes MHQA a key benchmark for assessing LLM reasoning. Datasets like HotpotQA, 2WikiMQA, and MuSiQue, (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022), have been developed for this purpose. RAG models combine LLMs' internal (parametric) knowledge with external (non-parametric) retrieved data, deciding whether to use retrieved evidence before answer generation. This approach usually outperforms LLM-only methods, emphasizing the critical role of retrieval and aggregate information from multiple sources.

The best performing MHQA models adopt an iterative RAG approach. In MHQA, where answers require synthesizing evidence across multiple steps rather than a single one, this distinction is crucial. For example, the Self-Ask (Press et al., 2023) method iteratively decomposes complex queries into simpler sub-questions, retrieving relevant context for each piece of non-parametric knowledge. IRCoT (Trivedi et al., 2023) further enhances performance by iteratively combining retrieval and Chain-of-Thought (CoT) (Wei et al., 2023) reasoning at each step, refining responses beyond a simple retrieve-and-read approach. Recent studies (Fan et al., 2024) address these intricate reasoning challenges effectively.

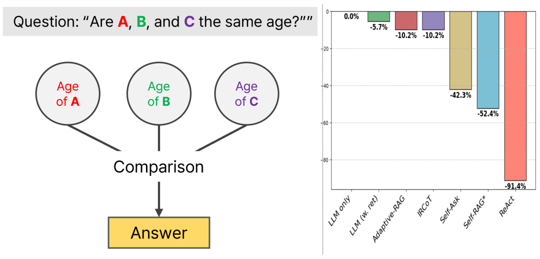


Figure 1: Example of comparison type question and its required process for answer generation **(left)** and performance reduction in F1 score relative to LLM for various RAG methods on our CompQA dataset **(right)**.

We hypothesize that comparison questions are especially challenging for MHQA. Our preliminary study shows that iterative methods underperform on these queries. Comparison questions require retrieving, aggregating, and contrasting independent entities—a process more demanding than sequential inference. For instance, "Are Elon Musk and Mark Zuckerberg the same age?" is common in MHQA and appears in HotpotQA alongside bridge questions. Retrieving and processing information for Elon and Mark from independent sources can further increase complexity, especially as the number of hops grows. However, benchmarks like 2WikiMultihopQA and MuSiQue primarily feature bridge questions, resulting in comparison questions insufficiently represented.

To better test MHQA on comparison questions, we introduce CompQA, a dataset designed to evaluate models that retrieve, aggregate, and compare information across entities. CompQA features 2hop to 6hop yes/no questions, composed of 1,000 questions (built from 4,000 sub-questions) sampled from the RetrievalQA (Zhang et al., 2024) dataset. Figure 1 shows that recent RAG models have lower F1 than an LLM-only baseline, highlighting that iterative methods struggle to integrate and reason over retrieved information for comparison questions.

According to experiments with recent MHQA models on CompQA, they struggle with comparison-type questions, showing performance declines of up to 39.8% compared to the LLM-only baseline, despite increasing LLM call counts by a factor of 2.5 to 9.9. Details can be found in Figure. 3 and Table. 2. Moreover, qualitative analysis reveals that an excessive dependency on retrieval, even when parametric knowledge is queried, results in suboptimal integration of entities. Furthermore, this dependency increases the number of LLM inference calls.

Our main contributions are as follows:

- CompQA is introduced as a benchmark dataset specifically designed to evaluate the comparative and aggregative capabilities of RAG models using comparison questions with up to 6hop complexity.

- Quantitative and qualitative analysis show that existing methods are inadequately evaluated on comparison questions and emphasize that effective integration of non-parametric with parametric knowledge is critical.

2 Related Work

2.1 Multihop Question Answering

MHQA datasets, like HotpotQA and MuSiQue are designed to assess a model's capacity to integrate evidence from multiple sources and inference steps. In particular, MuSiQue transforms singlehop questions into multihop challenges by synthesizing sub-questions through step-wise reasoning, often involving non-linear inference chains that extend up to four hops.

2.2 Retrieval-Augmented Generation

LLMs have demonstrated robust performance in reasoning, generation, and comprehension tasks (Zhao et al., 2024). However, their heavy reliance on pre-training data leads to gaps in domains not well represented in the training corpus, limits the integration of the latest knowledge (Lewis et al., 2020), and makes them susceptible to hallucinations and factual inaccuracies (Jiang et al., 2023). RAG addresses these issues by retrieving external knowledge for informed generation. Simple RAG with single retrieval struggles with MHQA datasets like HotpotQA and MuSiQue, which demand information integration.

Iterative RAG models like IRCoT and Adaptive-RAG (Jeong et al., 2024), while improving MHQA, are weak on comparison questions. IRCoT's independent, step-wise retrieval hinders cross-fact analysis. Adaptive-RAG inherits this limitation, struggling to synthesize and compare information. Similarly, Self-Ask and ReAct (Yao et al., 2023), though adept at fact retrieval, lack explicit comparison mechanisms. Self-RAG (Asai et al., 2024)'s sequential design, with a one-time retrieval decision and single-point generation input, also limits multihop comparison reasoning.

Our work extends this line of research by specifically focusing on the complex integration challenges of comparison type questions, an underexplored area in existing MHQA benchmarks.

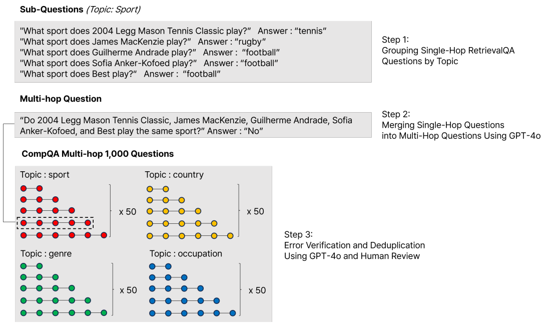
3 CompQA Dataset

CompQA dataset was constructed using MuSiQue, which is a benchmark MHQA dataset for evaluating multihop question answering[1]. It includes 2hop to 4hop questions in six categories, including hop1 (*bridge*), hop2 (*concurrent*), and hop3 (*hybrid*) types. Hop1 use sub-question answers as placeholders. Hop2 questions require simultaneous answers. Hop3 combine hop1 and hop2 characteristics.

CompQA covers 2hop to 6hop comparison questions, focusing on longerhop comparisons (e.g., 5hop, 6hop). Figure 1 shows a 6hop example. Each question includes sub-question pairs and yes/no answers. CompQA comprises 1,000 questions and 4,000 sub-question sets.

Dataset Construction Process CompQA was constructed by organizing RetrievalQA (Zhang et al., 2024) questions into sub-questions. It contains 1,000 questions (200 per hop level, 2hop to

[1] Sampled 400 questions/hop type (2hop, 3hop1, 3hop2, 4hop1, 4hop2, 4hop3) from MuSiQue training set, focusing on largest categories, with minor adjustments for 2hop and 4hop2 due to data constraints.



Sub-Questions *(Topic: Sport)*

"What sport does 2004 Legg Mason Tennis Classic play?" Answer : "tennis"
"What sport does James MacKenzie play?" Answer : "rugby"
"What sport does Guilherme Andrade play?" Answer : "football"
"What sport does Sofia Anker-Kofoed play?" Answer : "football"
"What sport does Best play?" Answer : "football"

Step 1:
Grouping Single-Hop RetrievalQA Questions by Topic

Multi-hop Question

"Do 2004 Legg Mason Tennis Classic, James MacKenzie, Guilherme Andrade, Sofia Anker-Kofoed, and Best play the same sport?" Answer : "No"

Step 2:
Merging Single-Hop Questions into Multi-Hop Questions Using GPT-4o

CompQA Multi-hop 1,000 Questions

Step 3:
Error Verification and Deduplication Using GPT-4o and Human Review

Figure 2: Pipeline of CompQA dataset construction

6hop), designed to compare attributes across similar topics. As shown in Figure 2, we generated 50 questions per topic (Country, Occupation, Sport, Genre) for each hop level, totaling 200 questions per hop and 1,000 overall. GPT-4o generated the dataset, guided by topic and format examples, followed by human review and GPT-4o verification for quality. Dataset creation occurred between December 18, 2024, and January 5, 2025, including human review and revisions.

Labeling Sub-questions in MuSiQue and CompQA were labeled as parametric (*answerable by internal knowledge*) or non-parametric (*requiring retrieval*), model-specifically. For each sub-question, we generated an answer and calculated F1 score. F1=0 indicated non-parametric (0), F1>0 parametric (1), following RetrievalQA. Questions were grouped by parametric ratio: **High Parametric (p)**: >0.5 parametric sub-questions; **High Non-parametric (np)**: >0.5 non-parametric sub-questions. 0.5 ratio questions were excluded.

4 Experiment & Analysis

4.1 Experimental Setup

We evaluated five representative RAG models (IRCoT, Adaptive-RAG, Self-RAG, Self-Ask, ReAct), focusing on Recursive and Adaptive types. Our setup replicated original configurations to analyze each model's inherent behavior. GPT-3.5-turbo[2] served as the generator LLM for all models except Self-RAG[3]. BM25 (Robertson et al., 1995) was used as retriever for LLM-only, RAG, IRCoT, and Adaptive-RAG, with Wikipedia (Karpukhin et al., 2020) and MuSiQue corpora for CompQA and MuSiQue respectively, following (Trivedi et al., 2023). Self-Ask used Google Search, and ReAct used the Wikipedia API. We used F1 score and LLM call count for performance and efficiency evaluation, respectively.

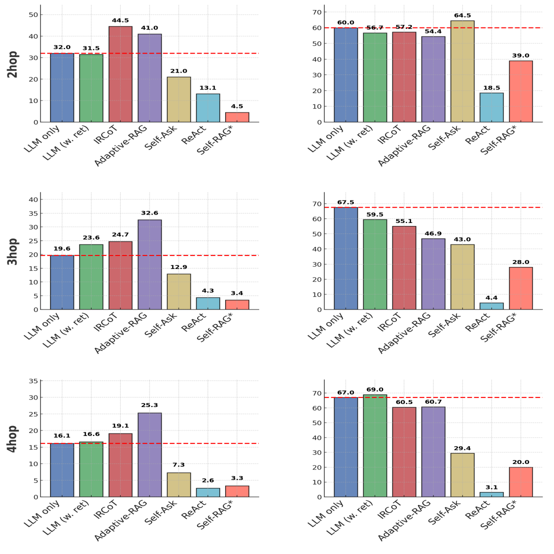


Figure 3: Comparison of F1 score between MuSiQue bridge type questions on the **left** and our CompQA dataset on the **right**. The red dashed line represents the F1 score of the LLM as a baseline.

Methods	MuSiQue			CompQA		
	2hop	*3hop1*	*4hop1*	*2hop*	*3hop*	*4hop*
IRCoT	3.2 (+2.2)	3.6 (+2.6)	4.0 (+3.0)	2.9 (+1.9)	3.1 (+2.1)	3.3 (+2.3)
Adaptive-RAG	2.6 (+1.6)	3.2 (+2.2)	3.7 (+2.7)	2.9 (+1.9)	3.1 (+2.1)	3.3 (+2.3)
Self-Ask	1.1 (+0.1)	1.8 (+0.8)	1.9 (+0.9)	2.5 (+1.5)	3.1 (+2.1)	3.7 (+2.7)
ReAct	8.3 (+7.3)	9.4 (+8.4)	9.9 (+8.9)	7.2 (+6.2)	8.0 (+7.0)	8.6 (+7.6)
Self-RAG*	1.6 (+0.6)	1.5 (+0.5)	1.7 (+0.7)	4.5 (+3.5)	5.1 (+4.1)	4.9 (+3.9)

Table 1: LLM calls per hop for MuSiQue bridge type questions and CompQA. Red numbers indicate the increase based on LLM-only and LLM w. retrieval call counts.

4.2 RAG Performance on Comparison Type Questions

This subsection presents experimental validation showing that RAG models exhibit limitations when addressing comparison type questions.

To experimentally verify this hypothesis, we utilized the CompQA dataset. As a control group for comparison type questions, we extracted bridge type questions (specifically 2hop, 3hop1, and 4hop1 subsets) from the MuSiQue. The experimental results are presented in the subsequent section.

Figure 3 distinctly illustrates the differential performance trends of RAG models across bridge and comparison type questions. Notably, in contrast to LLM-only and LLM with retrieval methodolo-

[2] GPT-3.5-turbo replaced deprecated text-davinci-002 for Self-Ask and ReAct experiments.

[3] Fine-tuned LLaMA-2 7B

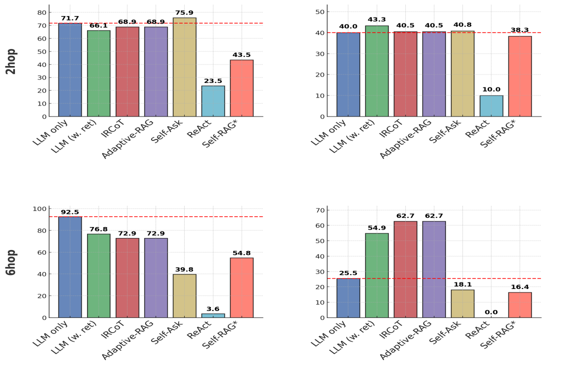


Figure 4: Comparison of F1 score for Parametric Major **(left)** and Non-Parametric Major **(right)** on CompQA. The red dashed line represents the F1 score of LLM as a baseline.

gies, RAG models exhibit a peculiar vulnerability, demonstrating comparatively weaker performance on comparison type questions.

For comparison type questions, the LLM-only and LLM w. retrieval methods consistently maintain robust F1 score even with a single LLM call. This suggests that for comparison type questions, satisfactory performance can be attained without applying RAG approach with complex reasoning process.

Conversely, while RAG models may demonstrate performance enhancements over LLM-only and LLM w. retrieval methods for bridge type questions, they tend to exhibit performance degradation or stagnation at LLM-only levels for comparison type questions. For instance, IRCoT achieves an 57.2% for 2hop comparison type questions, which is comparable to or even slightly lower than LLM-only, despite incurring approximately three times the number of LLM calls. For 4hop comparison type questions, IRCoT's performance further diminishes to 60.5%, significantly lower than LLM-only while maintaining a high LLM call count of 3.3.

Adaptive-RAG and Self-Ask show similar performance trends, while ReAct struggles the most with comparison questions. For 2hop, ReAct scores just 18.5% with 7.2 LLM calls, dropping to 3.1% at 4hop as calls rise to 8.6.

The performance decline of RAG models below LLM-only and LLM w. retrieval baselines on comparison type questions underscores their inefficacy in explicit information comparison. As shown in Table 1, despite increasing LLM call counts to 2.5–9.9, RAG models yield negligible or negative performance gains, highlighting their inefficiency relative to computational cost.

Methods	Parametric			Non-Parametric		
	2hop	*3hop*	*4hop*	*2hop*	*3hop*	*4hop*
IRCoT	3.1 (+2.1)	3.2 (+2.2)	3.4 (+2.4)	2.7 (+1.7)	2.9 (+1.9)	3.2 (+2.2)
Adaptive-RAG	3.1 (+2.1)	3.2 (+2.2)	3.4 (+2.4)	2.7 (+1.7)	2.9 (+1.9)	3.2 (+2.2)
Self-Ask	2.5 (+1.5)	3.2 (+2.2)	3.7 (+2.7)	2.4 (+1.4)	3.1 (+2.1)	4.0 (+3.0)
ReAct	7.1 (+6.1)	8.4 (+7.4)	8.9 (+7.9)	7.4 (+6.4)	7.1 (+6.1)	8.2 (+7.2)
Self-RAG*	4.2 (+3.2)	4.8 (+3.8)	4.2 (+3.2)	5.3 (+4.3)	5.7 (+4.7)	5.6 (+4.6)

Table 2: LLM calls per question for CompQA (Red numbers indicate the increase based on LLM-only and LLM w. retrieval call counts).

In conclusion, the experimental results presented in Figure 3 underscore that RAG models are not universally effective across all question types, particularly revealing their limitations in handling comparison type questions.

We hypothesized that effective comparison relies on consistent information comparison, with retrieval quality affecting accuracy. Analyzing CompQA, we examined parametric (internally answerable) vs. non-parametric (requiring retrieval) sub-questions to assess their impact on RAG performance.

Figure 4 shows parametric questions yield higher F1 score, leveraging internal knowledge, while non-parametric questions suffer performance declines. As expected, RAG outperforms LLM-only on non-parametric questions, whereas LLM-only peaks on parametric ones. This trend intensifies at higher hops (5hop, 6hop). Full results are in Appendix A.3.

Table 2 reveals a counterintuitive trend: IRCoT and Adaptive-RAG increase LLM calls for parametric questions. To explain this, we classify three failure cases of IRCoT on comparison questions requiring parametric knowledge in Appendix A.4.

5 Conclusion

This paper highlights a notable limitation of RAG methods in MHQA, focusing on comparison questions. Our experimental study, using newly introduced CompQA dataset, shows that recent models underperform relative to LLM-only baseline. Furthermore, the results demonstrate significant inefficiency, with RAG methods requiring more inference calls without corresponding performance improvements.

6 Limitations

Our study is limited by the specific RAG models and datasets evaluated, primarily CompQA and MuSiQue. Generalizability may be limited to other RAG architectures and comparison question types beyond our evaluation scope. Further qualitative analysis across all RAG models, beyond quantitative metrics and the IRCoT error analysis, could provide richer insights. Future work should focus on developing question-type-aware RAG models, exploring efficient methods for comparison questions, and broader evaluations.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Wenqi Fan, Yujuan Ding, Liang bo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Knowledge Discovery and Data Mining*.

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.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah 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, Singapore. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang

Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Appendix

A.1 Dataset label statistics

Hop	GPT3.5			Llama2 (self-rag ft)		
	parametric	*non-parametric*	*neutral*	*parametric*	*non-parametric*	*neutral*
2hop	107	90	192	26	230	133
3hop1	189	211	0	79	321	0
3hop2	179	221	0	48	352	0
4hop1	129	119	152	23	273	104
4hop2	56	20	51	9	77	41
4hop3	83	159	158	13	316	71

Table 3: Labeling statistics for MuSiQue, categorized into parametric, non-parametric, and neutral (0.5 ratio). The model Llama2 (self-rag ft) refers to a fine-tuned version of Llama2 based on self-rag.

As described in the labeling process in Section 3, Table 3 presents the labeling statistics for MuSiQue, categorized into parametric, non-parametric, and neutral (0.5 ratio). The classification was conducted to distinguish answerable sub-questions based on parametric knowledge. The model Llama-2 7B (self-rag ft) refers to a fine-tuned version of Llama2 based on self-rag, used to evaluate the performance differences in handling various question types.

A.2 Dataset label statistics

Hop	GPT3.5			Llama2 (self-rag ft)		
	parametric	*non-parametric*	*neutral*	*parametric*	*non-parametric*	*neutral*
2hop	106	30	64	92	47	61
3hop	152	48	0	121	79	0
4hop	129	44	27	108	53	39
5hop	148	52	0	136	64	0
6hop	146	51	3	124	55	21

Table 4: Labeling statistics for CompQA, categorized into parametric, non-parametric, and neutral (0.5 ratio). The model Llama2 (self-rag ft) refers to a fine-tuned version of Llama2 based on self-rag.

As described in the labeling process in Section 3 Labeling, Table 4 presents the labeling statistics for the CompQA dataset, categorized into parametric, non-parametric, and neutral (0.5 ratio). The labeling process in CompQA was conducted to distinguish whether answering a question requires parametric knowledge. Additionally, the Llama2 (self-rag ft) model refers to a fine-tuned version of Llama2 based on self-rag, used to analyze the performance differences across various question types.

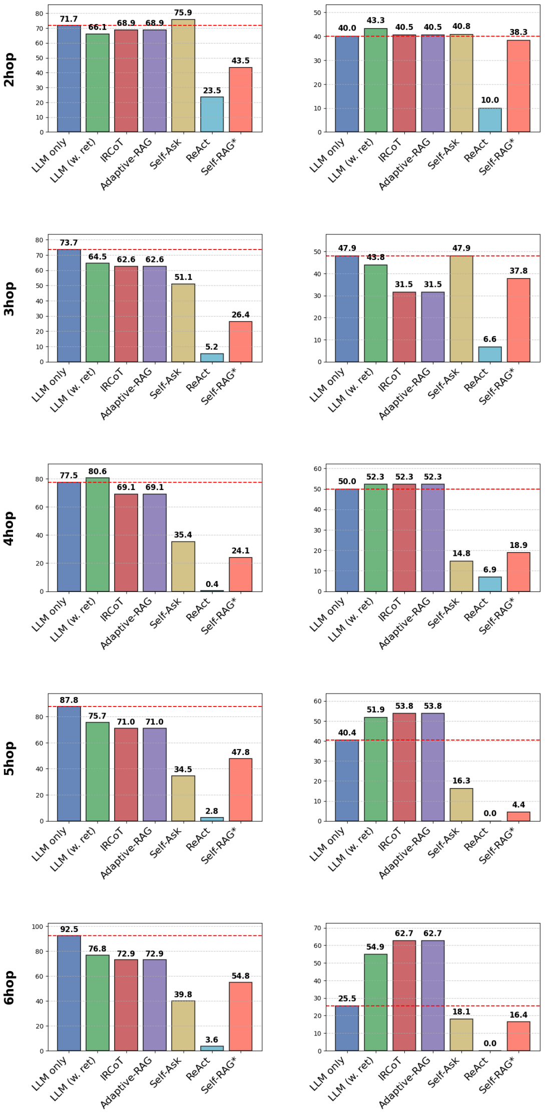


Figure 5: Total Results of comparison with F1 score for Parametric Major **(left)** and Non-Parametric Major **(right)** on our CompQA dataset.

A.3 Total Result on CompQA

Figure 5 shows the total results of Table 4 in Section 4.2. In comparison to the LLM-only baseline, the latter generally achieved peak performance on parametric questions, while RAG models incorporating retrieval mechanisms demonstrated enhanced efficacy on non-parametric questions, aligning with expected trends. This trend is particularly pronounced at higher hop counts, specifically 5hop and 6hop complexities. Furthermore, in the case of ReAct, instances of F1 score dropping to single-digit values are observed when transitioning from 2hop to 3hop and beyond. This underscores that inaccurate or inconsistent external retrieval can precipitate a sharp decline in the ultimate answer quality, particularly for questions that require long-term inference processes.

A.4 Why do RAG models fail at comparison type questions?

We conducted a qualitative analysis to understand the reasons behind the observed vulnerability of current RAG models when addressing comparison type questions.

Specifically, we aimed to identify the underlying causes by analyzing the reasoning path of IRCoT, a representative RAG model that serves as a backbone or provides insights for other RAG models. IRCoT is particularly suitable for this analysis because it performs in-depth reasoning to derive answers to multihop questions and explicitly presents this process. Among the 126 questions that LLM answered correctly but IRCoT answered incorrectly, 118 belonged to the P group (high parametric ratio) and 8 to the NP group (high non-parametric ratio). Based on this, we assumed that LLM achieved a perfectly correct answer (EM = 1), while IRCoT was completely incorrect (F1 = 0) to facilitate the evaluation.

Four master's students majoring in natural language processing qualitatively evaluated the 118 questions in the P group, and categorized them into four categories: 'Case 1. Questions where the reasoning process was incorrect, but the answer was semantically correct' (59 questions), 'Case 2. Questions where the reasoning process was incorrect and the answer was also incorrect' (29 questions), 'Case 3. Questions answered as "I don't know" (10 questions), and 'Case 4. Questions where the reasoning process was correct and the answer was semantically correct' (20 questions). (We clarify that although metric filtering was performed, there were semantically correct answers). We will exclude case 4, which presents no issues, and examine qualitative examples for cases 1, 2, and 3.

Case 1. Questions where the reasoning process was incorrect, but the answer was semantically correct

Question: "Are Borysławice and Colonia Nueva Coneta in the same country?"

Analysis: Despite LLM knowing information about all entities, IRCoT incorrectly judged 'Colonia Nueva Coneta' to be in Uruguay. Consequently, a result where the reasoning process was flawed but the answer was correct was derived. This indicates that the model tends to trust the retrieved context more, even when it possesses parametric sub-question knowledge. In other words, even in situations where parametric information could be utilized, the model prioritized the retrieved context and exhibited issues in performing reasoning.

Case 2. Questions where the reasoning process was incorrect and the answer was also incorrect

Question: "Are Tupper-Barnett House, Contest, Studzianka, Podlaskie Voivodeship, Freedom, and Ara in the same country?"

Analysis: Even though LLM-only lacked knowledge about the last entity, IRCoT generated an output stating that all entities are in the same country (Poland). This demonstrates that IRCoT struggles to respond correctly even to questions for which it already possesses knowledge. Notably, even when the model accurately knows the parametric sub-questions, the reasoning process was distorted when the retrieved context was incomplete or incorrect. This implies that rather than utilizing its pre-existing knowledge, the model formed a new reasoning path based on the retrieved context, leading to errors.

Case 3. Questions answered as "I don't know"

Question: "Are Tina, Edmundston, and Valea Seacă River in the same country?"

Analysis: Despite LLM-only lacking knowledge about the first entity, IRCoT did not perform a search for 'Tina' and also failed to utilize parametric information for the known entities 'Edmundston' and 'Valea Seacă River'. This signifies a failure of the model to leverage its pre-existing information. Also, even in situations where only some entities are known, the model showed limitations in combining parametric information with retrieved information when performing reasoning.

Through these qualitative examples, we observed a tendency for the model to fail to perform correct reasoning even for parametric sub-questions that it already knows. This mirrors patterns similar to the phenomenon presented in prior research, which indicates a bias in LLM to accept generated contexts even if they contain inaccurate information, and aligns with the result of a tendency to prefer generated context over parametric information. In particular, IRCoT appeared to prioritize retrieved context, experiencing difficulties in utilizing already learned parametric knowledge.