AT-RAG: An Adaptive RAG Model Enhancing Query Efficiency with Topic Filtering and Iterative Reasoning

Anonymous ACL submission

Abstract

Recent advancements in Question Answering (QA) with Large Language Models (LLMs) like GPT-4 have shown limitations in handling complex multi-hop queries. We propose Adaptive Topic RAG (AT-RAG), a novel multi-step Retrieval Augmented Generation (RAG), which incorporates topic modeling for efficient document retrieval and reasoning. Using BERTopic, our model dynamically assigns topics to queries, improving retrieval accuracy and efficiency. We evaluated AT-RAG on multihop benchmark datasets (QA) and a medical case study QA. Results show significant improvements in correctness, completeness, and relevance compared to existing methods. AT-RAG reduces retrieval time while maintaining high precision, making it suitable for general tasks QA and complex domain-specific challenges such as medical QA. The integration of topic filtering and iterative reasoning enables our model to handle intricate queries efficiently, which makes it suitable for applications that require nuanced information retrieval and decision-making.

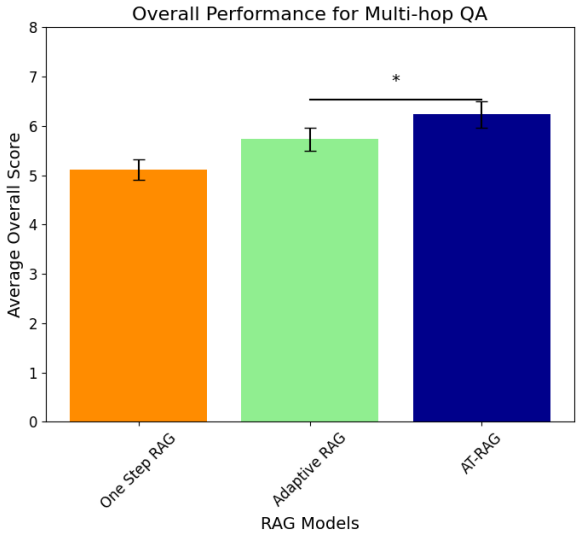

Figure 1: Comparison of the average overall score across multiple datasets for different RAG models (One Step RAG, Adaptive RAG, AT-RAG with GPT40). Error bars depict the standard deviations for each model. An ANOVA test (St et al., 1989) (with $p<0.05$) reveals a statistically significant difference between the AT-RAG and Adaptive RAG, denoted by an asterisk (*). For further details, refer to Table 1

1 Introduction

LLMs have transformed natural language processing, particularly in QA tasks, by generating coherent and contextually relevant responses using their vast pre-trained knowledge (Achiam et al., 2023; Team et al., 2023; Jiang et al., 2024). Although models like GPT-4 demonstrate impressive capabilities, they face significant challenges in responding to queries that require external information or reasoning across multiple documents(Raiaan et al., 2024; Kwiatkowski et al., 2019). These limitations are especially evident in multi-hop QA scenarios, where extracting and synthesizing information from various sources is essential for producing accurate answers(Press et al., 2022; Tang and Yang, 2024). To address these challenges, RAG models have been developed to enable LLMs to access relevant external knowledge and enhance the quality of responses.

In this paper, we introduce AT-RAG, a novel multi-step retrieval-based QA framework that enhances the retrieval process by incorporating a topic assignment model. This model filters external knowledge in QA tasks by assigning relevant topics to each query, ensuring retrieval focuses on contextually significant information. This approach improves retrieval accuracy and reduces the computational overhead associated with multi-step retrieval processes. Furthermore, AT-RAG integrates Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Wang and Zhou, 2024), allowing iterative document retrieval and reasoning to handle complex multi-hop queries better.

We evaluated AT-RAG on several challeng-

ing multi-hop QA datasets, including **HotpotQA** (Yang et al., 2018), **MuSiQue** (Trivedi et al., 2022), and **2WikiMultiHopQA** (Ho et al., 2020). Our results show that AT-RAG outperforms existing RAG models in terms of accuracy; see Figure 1. This work's main contribution is the introduction of topic-guided retrieval aimed at enhancing multi-step reasoning. The model undergoes evaluation on intricate datasets, with a thorough analysis of improvements in answer quality.

The experimental results reveal that AT-RAG significantly improves accuracy compared to existing methods like Adaptive-RAG (Lewis et al., 2020). These improvements are particularly pronounced in multi-hop reasoning tasks, where prior adaptive strategies often incur high computational costs. By narrowing the search space through a stochastic topic assignment process, AT-RAG reduces the number of documents needed for retrieval while improving the accuracy of query resolution. We evaluated AT-RAG using various state-of-the-art LLMs, including GPT-4 and Mixtral8x7B (Jiang et al., 2024). AT-RAG demonstrated superior performance when leveraging GPT-4 as the LLM, proving particularly effective in addressing complex multi-hop queries with optimized retrieval and precise reasoning.

Furthermore, we conducted a case study to assess the effectiveness of AT-RAG in addressing multi-hop queries in the real world, focusing on answering medical questions. Medical records, which are longitudinal and comprise various documents such as doctor's notes, lab results, diagnostics, and medications, present a unique challenge for the RAG frameworks. Time-based queries, such as *'What are the abnormal laboratory results of the patient in the last year?'*, often yield suboptimal results when processed by a naive RAG. These approaches struggle to retrieve relevant information from potentially thousands of document chunks within a vector database. Furthermore, naive RAG may lack the reasoning ability to identify the correct time-stamped data that meet both the time range and the specific condition. As illustrated in Fig. 4, our method consistently outperforms naive RAG approaches across all six evaluated cases.

2 Related Work

Recent advancements in multi-hop QA systems have emphasized enhancing accuracy, efficiency, and reasoning capabilities. This section reviews three main categories: question decomposition, CoT with iterative retrieval, and adaptive retrieval.

2.1 Question Decomposition

Question decomposition breaks down complex queries into simpler sub-questions for a more straightforward resolution. Khattab et al. (Shao et al., 2023) proposed the Iterative Retriever, Reader, and Reranker framework, which decomposes queries, retrieves relevant passages, and synthesizes information to generate answers (Zhang et al., 2024). Press et al. (2023) introduced "Decomposed Prompting," a technique that leverages large language models to simplify complex queries into more manageable sub-questions(Schulhoff et al., 2024).

2.2 Chain-of-Thought with Iterative Retrieval

This approach combines CoT reasoning with iterative document retrieval. Yao et al. (Sun et al., 2022) introduced "ReCite," where a large language model generates reasoning steps while retrieving relevant documents iteratively. Generating intermediate reasoning steps improves the model's ability to handle complex reasoning tasks. Furthermore, such reasoning capabilities naturally emerge in large models through CoT prompting (Wei et al., 2022).

2.3 Adaptive Retrieval

Adaptive retrieval methods dynamically adjust the retrieval process based on the specific needs of each query (Fan et al., 2024). (Asai et al., 2023) introduced a system that allows a language model to iteratively formulate and resolve subsequent queries as needed. Another notable approach is IRCoT, which integrates the retrieval and reasoning phases, improving multi-step question answering on datasets like HotpotQA and 2WikiMultiHopQA (Trivedi et al., 2022). The Adaptive-RAG framework takes this further by selecting the most appropriate retrieval strategy based on the complexity of the query.

Despite these advancements, challenges remain. Current systems lack a fully flexible approach that dynamically adapts retrieval and reasoning processes in response to query complexity. Balancing efficiency and thoroughness, especially for queries with varying difficulty levels, continues to be a critical area for improvement. Future work must address these issues to enhance the adaptability and performance of multi-hop QA systems.

3 Method

This section introduces the proposed RAG model, AT-RAG, which combines single-step and multi-step retrieval strategies enhanced by topic assignment to tackle complex QA challenges. By integrating topic modeling with multi-step reasoning, the model improves both retrieval precision and efficiency, leading to more accurate and well-reasoned answers.

3.1 Background

LLMs is designed to process an input sequence of tokens and generate an output sequence. Formally, given an input sequence $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, the LLM generates an output sequence $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T$, expressed as $\mathbf{y} = LLM(\mathbf{x})$ where n and m are the number of tokens for each sequence. In the context of QA, the input sequence $\mathbf{x}$ corresponds to the user's query $\mathbf{q}$, and the output sequence $\mathbf{y}$ corresponds to the generated answer $\hat{\mathbf{a}}$, defined as $\hat{\mathbf{a}} = LLM(\mathbf{q})$. Ideally, $\hat{\mathbf{a}}$ should match the correct answer $\mathbf{a}$.

While this non-retrieval-based QA method is efficient and leverages the vast knowledge within the LLM, it struggles with queries requiring precise or up-to-date information, such as details about specific people or events beyond the LLM's internal knowledge. Non-retrieval QA is effective for simple queries but faces limitations with more complex or niche questions.

3.1.1 One Step RAG for QA

To overcome the limitations of nonretrieval methods for queries requiring external knowledge, retrieval-based methods QA can be employed. This method utilizes external knowledge $\mathbf{d}$, retrieved from a knowledge source $\mathbf{D}$ (e.g., Wikipedia (Chen, 2017) or Wikidata (Vrandečić and Krötzsch, 2014)), which contains millions of documents. The retrieval process is formalized as:

$$\mathbf{d} = Retriever(\mathbf{q}; \mathbf{D})$$

where `Retriever` is the model that searches $\mathbf{D}$ for relevant documents based on query $\mathbf{q}$. The retrieved knowledge $\mathbf{d}$ is then incorporated into the input of LLM, enhancing the QA process by generating an answer $\hat{\mathbf{a}}$ based on both the query and the retrieved documents:

$$\hat{\mathbf{a}} = LLM(\mathbf{q}, \mathbf{d})$$

This approach improves the performance of the LLM for queries requiring specific or real-time information, augmenting its pre-trained knowledge with external sources.

3.1.2 Multi-Step RAG for QA

Although single-step RAG is effective for many queries, it has limitations when dealing with complex questions requiring simultaneous processing of information in multiple documents or the reasoning of interconnected knowledge. A, a multi-step RAG approach is introduced to address complex question, where the LLM iteratively interact with the retriever.

At each iteration $i = 1, ..., N$, the retriever fetches new documents $\mathbf{d}_i$ from $\mathbf{D}$, and the LLM incorporates both the newly retrieved document $\mathbf{d}_i$ and the context $\mathbf{c}_i$ (which includes previously retrieved documents and intermediate answers $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_{i-1}$):

$$\hat{\mathbf{a}}_i = LLM(\mathbf{q}, \mathbf{d}_i, \mathbf{c}_i) \quad (1)$$

$$\mathbf{d}_i = Retriever(\mathbf{q}, \mathbf{c}_i; \mathbf{D})$$

This iterative process continues until the LLM constructs a comprehensive understanding of the query, leading to the final answer. This approach is beneficial for complex multi-hop queries, where information needs to be integrated from multiple retrievals. However, it is more resource-intensive due to the increased computational cost of repeated interactions between the retriever and the LLM.

3.2 Topic Filtering for RAG

To improve the efficiency of Multi-Step RAG in QA, we propose the AT-RAG model. Before passing the query to the retriever, it is processed by a topic assignment module that generates a topic based on the input query. This topic reduces the search space by filtering irrelevant information, allowing for a more focused retrieval of relevant documents. This strategy can reduce search time and improve document relevance.

3.2.1 AT-RAG Model

The AT-RAG enhances the LLMs by integrating retrieved documents and performing multi-step reasoning. Formally, let $\mathbf{q}$ represent the query and $\mathbf{D}$ denote the external knowledge base. Using a topic assignment model $f_{\boldsymbol{\theta}}(.)$, the associated query topic is defined as $t = f_{\boldsymbol{\theta}}(\mathbf{q})$, where $\boldsymbol{\theta}$ represents the parameters of the model.

The AT-RAG process begins by passing $\mathbf{q}$ through $f_{\boldsymbol{\theta}}(.)$ to generate the topic t_1. This topic is

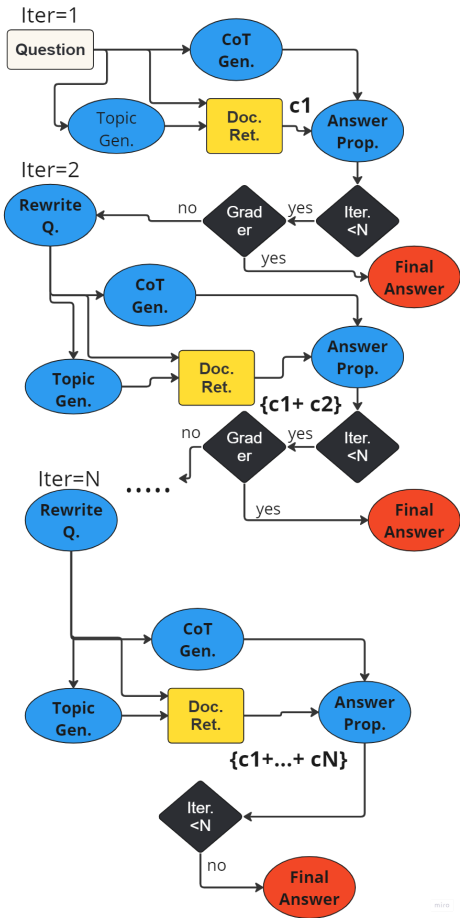

Figure 2: The AT-RAG answering pipeline leverages a topic generator to streamline document retrieval. It iteratively generates reasoning steps through a CoT generator, guiding the formulation of answers. This process alternates between retrieval and reasoning until a predefined maximum number of iterations (N) is reached or the answer passes quality checks by grader nodes.

then used to retrieve a set of relevant documents $\mathbf{d}_1$ from $\mathbf{D}$:

$$\mathbf{d}_1 = \text{Retriever}(\mathbf{q}, t_1; \mathbf{D})$$

Next, the model generates reasoning steps (CoT), denoted as $\mathbf{c}_1$, and integrates these documents into the LLM's input to generate an answer $\hat{\mathbf{a}}$:

$$\hat{\mathbf{a}} = \text{LLM}(\mathbf{q}, \mathbf{d}_1, \mathbf{c}_1)$$

Should the initial retrieval prove inadequate, as determined by the Answer Grader module (refer to section 3.2.5 for further information), the system is capable of iteratively enhancing the query and executing the retrieval process again in subsequent iterations:

$$\hat{\mathbf{a}}_i = \text{LLM}(\mathbf{q}_i, \mathbf{d}_i, \mathbf{c}_{1:i})$$

where $\mathbf{q}_i$ is the refined query at iteration i, and $\mathbf{c}_{1:i}$ is the accumulated reasoning context. Documents are retrieved as follows:

$$\mathbf{d}_i = \text{Retriever}(\mathbf{q}_i, t_i; \mathbf{D})$$

where t_i is generated by $t_i = f_{\boldsymbol{\theta}}(\mathbf{q}_i)$.

3.2.2 Topic Assignment Model

The topic assignment model enhances retrieval accuracy by refining the search space. It predicts the most relevant topic for a given query, enabling the retrieval system to focus on a specific subset of the knowledge base. Let $f_{\boldsymbol{\theta}}(.)$ denote the topic assignment model. For a query at iteration i, $\mathbf{q}_i$, the corresponding topic t_i is generated as:

$$t_i = f_{\boldsymbol{\theta}}(\mathbf{q}_i)$$

The topic t summarizes the query's domain, filtering the document database $\mathbf{D}$ to improve retrieval precision and reduce computational complexity.

We implement this using BERTopic (Grootendorst, 2022), which leverages transformer-based models for advanced topic discovery. BERTopic clusters document embeddings and applies a class-based model to create coherent topic representations, effectively capturing contextual word meanings.

For each multi-hop dataset, we fine-tune a pre-trained BERTopic model and apply it during both inference and data ingestion phases. This fine-tuning process adapts the model to each dataset's unique characteristics, enhancing topic coherence and retrieval accuracy.

3.2.3 Analysis of Topic Distribution Across Datasets

To highlight the importance of topic assignment in AT-RAG, we analyzed the topic distribution for documents in each multi-hop dataset. We fine-tuned the BertTopic model for the datasets and visualized the distribution of the top 5 topics. Figure 3 illustrates how topic assignment helps mitigate dataset bias, which can influence retrieval processes and RAG model performance.

Our analysis reveals distinct differences in topic distribution across datasets. For instance, "film" content is similarly represented in **MuSiQue** and **HotpotQA**, but less prevalent in **2WikiMultiHopQA**. "Music" topics are more prominent in **MuSiQue**, while **HotpotQA** and **2WikiMultiHopQA** show less emphasis on this area.

Understanding these distributions is crucial for identifying dominant themes and thematic focus

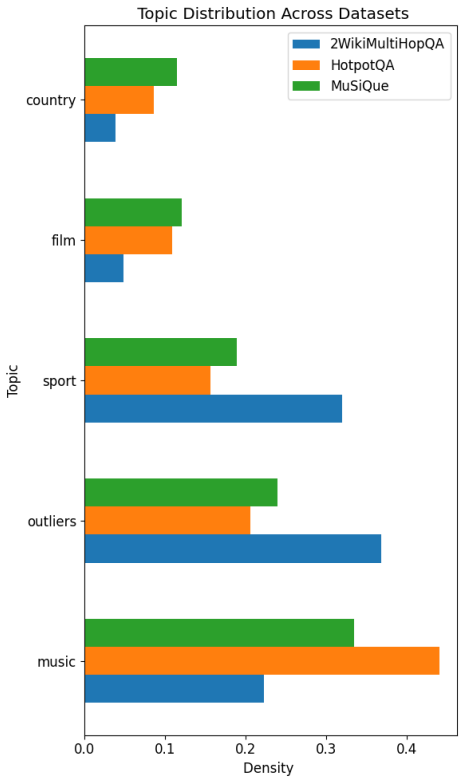

Figure 3: Normalized Topic Distribution Using TopicBERT Across Multi-Hop QA Datasets. The bar plot displays the relative density (proportion of total documents) for each topic, highlighting thematic diversity within each dataset. This visualization emphasizes how topic assignment addresses dataset bias, influencing the retrieval process in QA tasks.

in each dataset. This insight is vital for improving retrieval performance and allows for tailoring the retriever to match the topic distribution, ensuring more relevant document selection for each query.

3.2.4 Data Ingestion for Vector Database

The data ingestion process prepares the dataset for embedding and vector store creation. First, documents and metadata are extracted. Topic assignment is performed using a specialized model, assigning topics and probabilities to each document and adding them to the metadata. Using an embedding model, the vector embedding step converts documents into dense vector embeddings. These embeddings are stored in a vector database (e.g., Chroma) for efficient similarity search. Documents are processed in batches, and the vector store is persisted for future retrieval, enabling content- and topic-based filtering.

3.2.5 Answer Grader

As illustrated in Figure 2, AT-RAG incorporates an Answer Grader module to evaluate the quality and relevance of generated responses. This module consists of two key components:

The Usefulness Grader: Assesses the relevance and value of the answer to the user's query. The Hallucination Grader: Verifies the factual accuracy of the answer by cross-referencing it with the retrieved documents, minimizing the risk of hallucinations.

If discrepancies are detected, the RewriteQuery module is activated to reformulate the original query, addressing information gaps or ambiguities. The reformulated query is then passed back to the retriever for additional context. To prevent endless querying, we implement a maximum iteration limit, N. If reached, the process terminates, outputting the final state answer. All three modules—Usefulness Grader, Hallucination Grader, and RewriteQuery—utilize prompting techniques with LLMs for efficient and effective processing. For a comprehensive understanding of the entire QA process using AT-RAG, refer to Algorithm 1, which provides a detailed step-by-step workflow breakdown.

Algorithm 1 AT-RAG Model Inference

Require: Query $\mathbf{q}$, Knowledge base $\mathbf{D}$, Topic assignment model $f_{\boldsymbol{\theta}}$, Max iterations N
Ensure: Final answer $\hat{\mathbf{a}}$

```
Initialize context q_1 ← q
for i = 1 to N do
    Generate topic: t_i ← f_θ(q_i)
    Retrieve documents: d_i ← Retriever(q_i, t_i; D)
    Generate reasoning steps: c_i ← CoT(q_i, d_i)
    Produce answer: â_i ← LLM(q, d_i, c_{1:i})
    if UsefulnessGrader(â_i) is satisfactory then
        if HallucinationGrader(â_i, d_i) is not hallucinating then
            return â_i
        end if
    end if
    Update query: q_{i+1} ← RewriteQuery(q_i, â_i, c_i)
end for
return â_N
```

4 Experiments

To benchmark the AT-RAG model and rigorously evaluate its effectiveness in handling complex queries, we test it on multi-hop QA datasets. To address more challenging query scenarios, we employ three benchmark multi-hop QA datasets that require sequential reasoning across multiple documents: 1) MuSiQue (Trivedi et al., 2022), 2) HotpotQA (Yang et al., 2018), and 3) 2WikiMultiHopQA (Ho et al., 2020).

4.1 Multi-hop QA Dataset

The evolution of question-answering systems has led to the creation of multi-hop datasets designed to challenge and evaluate more advanced QA models. These datasets, including MuSiQue, HotpotQA, and 2WikiMultiHopQA, feature complex queries that require reasoning across multiple documents. Unlike single-hop QA, where answers can be extracted from a single source, multi-hop QA necessitates synthesizing information from various sources to generate accurate answers.

MuSiQue (Trivedi et al., 2022) focuses on questions requiring multistep reasoning by combining multiple facts. This dataset is particularly valuable for assessing models' ability to navigate interconnected information and draw logical conclusions. HotpotQA (Yang et al., 2018) emphasizes both reasoning and fact verification by querying linked documents. It tests a model's ability to find the correct answer and provide supporting facts, thereby assessing comprehension and inference skills simultaneously.

2WikiMultiHopQA (Ho et al., 2020) leverages linked Wikipedia articles to challenge models with reasoning paths that span multiple documents. This dataset is useful for evaluating how well QA systems handle real-world knowledge structures and navigate interlinked information sources.

The primary difference between these multi-hop datasets and their single-hop counterparts is the requirement for information synthesis. Multi-hop questions cannot be answered without combining and reasoning over information from multiple documents or data points. This characteristic makes these datasets essential for evaluating advanced, inference-driven models that emulate human-like reasoning processes.

As QA models continue to advance, these multi-hop datasets play a crucial role in pushing the boundaries of machine comprehension and reasoning, driving the development of more sophisticated QA systems capable of handling real-world complexity.

4.2 LLMs as Autonomous Judges in QA Evaluation

The integration of LLMs as autonomous judges in QA evaluation has transformed the automation of assessing responses based on qualitative metrics. By leveraging the deep comprehension and reasoning capabilities of models like GPT-4o, LLM-based evaluation systems can provide detailed insights into the correctness, completeness, relevance, and clarity of generated answers.

To automatically evaluate QA pairs (both generated and ground truth), an LLM such as GPT-4o is used as the judge. The process begins by presenting the LLM with a question, the ground truth answer, and the generated response(Badshah and Sajjad, 2024). The model is prompted to assess the generated answer based on the following predefined criteria:

1. **Correctness**: Does the generated answer accurately align with the ground truth?

2. **Completeness**: Does the response include all necessary information relevant to the question?

3. **Relevance**: Is the answer relevant to the posed question?

For each criterion, the LLM assigns a score between 0 and 10, with the overall score calculated as the average of the individual scores. This approach enables a nuanced evaluation, going beyond token-level comparison and allowing for a more human-like understanding of the content.

4.3 Experimental Results and Analyses

In this section, we compare the performance of three different RAG approaches: One Step RAG, Adaptive RAG, and our proposed AT-RAG method. The evaluation criteria include correctness, completeness, relevance, and overall score, as shown in Table 1.

As independent evaluators, the experimental results were analyzed using LLMs on a subset of 500 QA samples from each dataset. The answers' correctness, completeness, and relevance were scored using GPT-4, each score ranging from 0 to 10, and the overall score was calculated as the average of these metrics. This approach goes beyond traditional token-level evaluation, leveraging the deep reasoning capabilities of the LLM to simulate human-like judgment.

As demonstrated in Table 1, the AT-RAG method consistently outperforms the other two approaches across all datasets. For example, in the 2WikiMultiHopQA dataset, the AT-RAG method achieved a correctness score of 5.79, a completeness score of 5.72, and a relevance score of 8.18, resulting in an

Table 1: Comparison between different RAG approaches.

	2WikiMultiHopQA				HotpotQA				MuSiQue			
Approach	Corr.	Comp.	Rel.	Overall	Corr.	Comp.	Rel.	Overall	Corr.	Comp.	Rel.	Overall
One Step RAG	3.84	3.94	7.03	4.94	5.93	5.82	7.97	6.57	3.00	3.12	5.38	3.83
Adaptive RAG	**5.99**	5.00	**8.18**	6.39	6.01	5.55	7.72	6.43	3.64	3.63	5.88	4.38
AT-RAG (with GPT-4o)	5.79	**5.72**	**8.18**	**6.57**	**7.27**	**6.98**	**8.56**	**7.61**	**3.65**	**3.88**	**6.02**	**4.52**

overall score of 6.57. This is a significant improvement over both One Step RAG and Adaptive RAG, with the latter scoring 6.39 overall.

A similar trend is observed in the HotpotQA dataset, where the AT-RAG method achieved an overall score of 7.61, compared to 6.57 for One Step RAG and 6.43 for Adaptive RAG. The MuSiQue dataset further highlights the efficacy of the AT-RAG approach, with an overall score of 4.52, outperforming both One Step and Adaptive RAG methods.

These results show that the AT-RAG method, which incorporates topic filtering, improves the quality of the answers by focusing on topic-relevant documents. This leads to more accurate, complete, and relevant responses, ultimately improving the overall performance of the QA system.

4.3.1 Ablation Study

To further assess the robustness of our proposed AT-RAG method, we conducted an ablation study using different LLMs, including GPT-4o and Mixtral8x7B (Jiang et al., 2024). This study aims to understand how the performance of the AT-RAG approach varies between different LLMs in multi-hop QA datasets. The results are summarized in Table 2. In all three datasets, GPT-4o consistently outperformed Mixtral8x7B across all evaluation metrics, demonstrating superior performance overall. Although both models performed lower in the MuSiQue dataset, GPT-4o still maintained an edge over Mixtral8x7B (Jiang et al., 2024). The ablation study clearly demonstrates that the choice of LLM significantly impacts the performance of the AT-RAG method.

5 Application on Medical QA

We conducted a case study to assess the effectiveness of our proposed method in answering multi-hop, time-based queries from medical records. The medical records of the patient, which include a variety of longitudinal documents such as doctor's notes, lab results, and diagnostics, pose a challenge for the RAG frameworks. For instance, the query: *What are the abnormal lab results of the patient within the last year?* is difficult for One Step RAG approaches, which must retrieve relevant chunks from vast data and may lack reasoning to select the correct time-stamped information.

We assessed our AT-RAG method by comparing it to the One Step RAG within a dataset of medical records from six patients. As shown in Figure 4, AT-RAG outperformed One Step RAG in all cases. Using Mixtral8x7B (Jiang et al., 2024) as the LLM and GPT-4o as the LLM judge, we queried each case with ten time-based questions. Our method attained an average score of 5.3, nearly two points above the One Step RAG, which had an average score of 3.5.

Table 3 illustrates a query on abnormal HbA1c values over the past two years. Our fine-tuned topic assignment model for this dataset identified the topic as LabResult, narrowing the retrieval to lab data. A CoT prompt further refined the search by focusing on the correct time range and the HbA1c condition, resulting in a correct answer, while naive RAG provided an incorrect one.

Table 4 shows another question about the last doctor visit and its reason. The topic assignment model assigned ClinicalNote to the query, limiting the search to relevant clinical notes. The CoT prompt guided the LLM to identify the date and reason for the visit. Our method retrieved the correct details, while naive RAG mistakenly returned a lab test date with no valid reason for the visit.

6 Conclusion

In this paper, we proposed AT-RAG, a novel RAG model designed to tackle complex multi-hop QA tasks. By integrating topic assignment through BERTopic, we significantly improved the speed and accuracy of multi-step document retrieval. The model effectively combines topic filtering with CoT reasoning to reduce the search space and focus retrieval on the most relevant documents. Experimental results on multi-hop QA demonstrated that AT-RAG outperforms existing RAG approaches in

Table 2: Comparison between different LLMs with AT-RAG model.

Approach	2WikiMultiHopQA				HotpotQA				MuSiQue			
	Corr.	Comp.	Rel.	Overall	Corr.	Comp.	Rel.	Overall	Corr.	Comp.	Rel.	Overall
AT-RAG (with GPT-4o)	**5.79**	**5.72**	**8.18**	**6.57**	**7.27**	**6.98**	**8.56**	**7.61**	**3.65**	**3.88**	**6.02**	**4.52**
AT-RAG (with Mixtral8x7B)	3.41	3.73	6.01	4.38	6.20	6.02	7.57	6.60	2.81	3.11	5.34	3.75

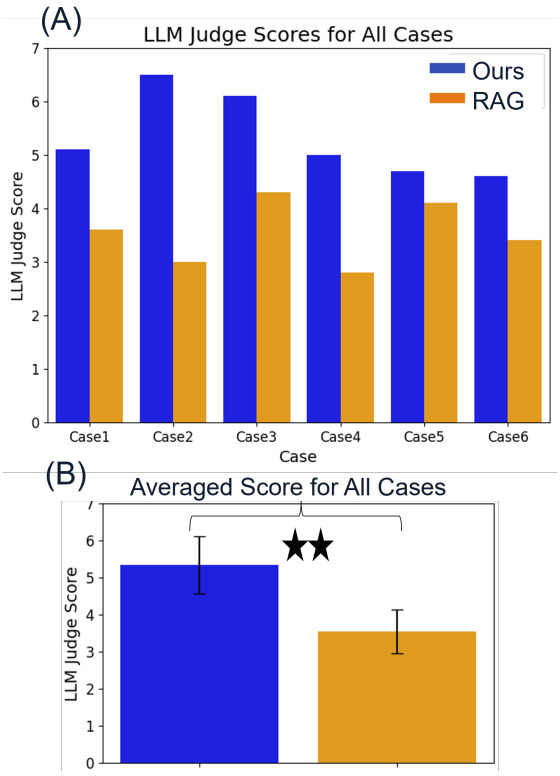

Figure 4: AT-RAG (A) A comparison between our AT-RAG and the One Step RAG in answering time-based questions on the medical records of six patients evaluated by GPT-4. (B) The average scores of our proposed approach and the One Step RAG across the six cases. ** indicates a statistically significant difference between the two bars, with $p < 0.02$ as determined by ANOVA test (St et al., 1989).

terms of correctness, completeness, relevance, and time efficiency. These improvements suggest that our approach is a promising solution for handling intricate QA tasks.

7 Limitations

Despite promising results, AT-RAG has some limitations that warrant further investigation. First, the performance of the model is highly dependent on the quality of the initial topic assignment. If the assigned topic is incorrect or too broad, the retrieval precision may decrease, leading to less accurate answers. Second, while we demonstrated improvements in retrieval efficiency, the multi-step nature of our approach can still be computationally expensive, particularly when applied to large datasets or real-time systems. Future work could explore methods to dynamically adjust the topic assignment model based on query complexity or integrate adaptive topic assigners mechanisms to further optimize the search process.

8 Ethics Statement

The development and deployment of LLMs such as those used in AT-RAG must be guided by strong ethical considerations. While our model aims to improve information retrieval and QA, we recognize the potential risks associated with LLMs, including the possibility of generating incorrect or biased information. In particular, the use of external knowledge sources raises concerns about the credibility and accuracy of retrieved documents, especially in critical domains such as healthcare as investigated here. It is crucial to ensure that the information retrieved by the model is reliable and factually grounded, and that any potential biases in the underlying datasets are mitigated.

Acknowledgements

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *arXiv preprint arXiv:2408.09235*.

D Chen. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Wenqi Fan, Yujuan Ding, Liangbo 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 *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Lars St, Svante Wold, et al. 1989. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chunliang Yang, Rosalind Potts, and David R Shanks. 2018. Enhancing learning and retrieval of new information: a review of the forward testing effect. *NPJ science of learning*, 3(1):8.

Zhihao Zhang, Alan Zhu, Lijie Yang, Yihua Xu, Lanting Li, Phitchaya Mangpo Phothilimthana, and Zhihao Jia. 2024. Accelerating retrieval-augmented language model serving with speculation. *arXiv preprint arXiv:2401.14021*.

A Appendix

A.1 Sample Quality Examples for Medical QA

Aspect	Simple Q&A	Q&A Graph
query	Does the patient have a history of HbA1c>6.0% within the past 2 years?	
LLM Judge Score	0	10
Assigned Topic	-	LabResult
Chain of Thought (CoT)	-	To determine if the patient has a history of HbA1c>6.0 within the past 2 years, we need to search for any lab results related to HbA1c in October 2022 or after. If there is a result with HbA1c>6.0, then the patient has a history of HbA1c>6.0 within the past 2 years.

Table 3: A comparison was made between One Step RAG and AT-RAG concerning the most recent doctor visit. As the QA process was conducted on medical records, we omitted the final answer and focused solely on the LLM Score to evaluate AT-RAG relative to One Step RAG.

Aspect	Simple Q&A	Q&A Graph
query	When was the last doctor visit and the reason behind it?	
LLM Judge Score	0	10
Assigned Topic	-	ClinicalNote
Chain of Thought (CoT)	-	To answer the question, we need to find the most recent clinical note that mentions a doctor visit. We will then extract the date of the visit and the reason behind it from the note.

Table 4: A comparison was made between One Step RAG and AT-RAG concerning the most recent doctor visit. As the QA process was conducted on medical records, we omitted the final answer and focused solely on the LLM Score to evaluate AT-RAG relative to One Step RAG.