applied sciences

MDPI

Article

Layered Query Retrieval: An Adaptive Framework for Retrieval-Augmented Generation in Complex Question Answering for Large Language Models

Jie Huang [1,2], Mo Wang [1,2,*], Yunpeng Cui [1,2], Juan Liu [1,2], Li Chen [1,2], Ting Wang [1,2], Huan Li [1,2] and Jinming Wu [1,2]

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2] Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China
* Correspondence: wangmo@caas.cn

Featured Application: This work is being used to develop the QA application of agricultural planting and livestock breeding technologies.

Abstract: Retrieval-augmented generation (RAG) addresses the problem of knowledge cutoff and overcomes the inherent limitations of pre-trained language models by retrieving relevant information in real time. However, challenges related to efficiency and accuracy persist in current RAG strategies. A key issue is how to select appropriate methods for user queries of varying complexity dynamically. This study introduces a novel adaptive retrieval-augmented generation framework termed Layered Query Retrieval (LQR). The LQR framework focuses on query complexity classification, retrieval strategies, and relevance analysis, utilizing a custom-built training dataset to train smaller models that aid the large language model (LLM) in efficiently retrieving relevant information. A central technique in LQR is a semantic rule-based approach to distinguish between different levels of multi-hop queries. The process begins by parsing the user's query for keywords, followed by a keyword-based document retrieval. Subsequently, we employ a natural language inference (NLI) model to assess whether the retrieved document is relevant to the query. We validated our approach on multiple single-hop and multi-hop datasets, demonstrating significant improvements in both accuracy and efficiency compared to existing single-step, multi-step, and adaptive methods. Our method exhibits high accuracy and efficiency, particularly on the HotpotQA dataset, where it outperforms the Adaptive-RAG method by improving accuracy by 9.4% and the F1 score by 16.14%. The proposed approach carefully balances retrieval efficiency with the accuracy of the LLM's responses.

Keywords: retrieval-augmented generation; question answer; adaptive retrieval; complex classification

check for updates

Citation: Huang, J.; Wang, M.; Cui, Y.; Liu, J.; Chen, L.; Wang, T.; Li, H.; Wu, J. Layered Query Retrieval: An Adaptive Framework for Retrieval-Augmented Generation in Complex Question Answering for Large Language Models. *Appl. Sci.* **2024**, *14*, 11014. https://doi.org/10.3390/app142311014

Academic Editor: Jose María Alvarez Rodríguez

Received: 17 October 2024
Revised: 22 November 2024
Accepted: 25 November 2024
Published: 27 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent large language models (LLMs), e.g., GPT-4 [1], Llama 3 [2], and Gemini [3], have demonstrated remarkable capabilities in natural language understanding, generation, and various downstream applications, such as machine translation, summarization, and question-answering. However, LLMs still encounter notable limitations, particularly in tasks that are domain-specific or knowledge-intensive [4,5]. One major issue is the production of "hallucinations" [6], where the models generate inaccurate information when faced with queries beyond their training data or requiring up-to-date knowledge. Retrieval-augmented generation (RAG) tackles these challenges by enhancing LLMs through the retrieval of relevant document fragments from external knowledge bases, leveraging semantic similarity measures [7]. RAG is designed to dynamically query and retrieve relevant documents or data from external knowledge sources, such as structured databases, web resources, or specialized corpora, during the inference process. By incorporating external knowledge, RAG significantly reduces the generation of factually incorrect content. Its integration into LLMs has led to widespread adoption, positioning RAG as a crucial

technology for advancing chatbots and enhancing the real-world applicability of LLMs. By retrieving relevant information in real time, RAG mitigates the problem of knowledge cutoff and helps overcome the inherent limitations of pre-trained language models that rely solely on the static data they were trained on [8].

The architecture of RAG is composed of two key components: a retriever model and a generator model. The retriever selects relevant pieces of external knowledge in response to a query, while the generator combines the retrieved information with the language model's internal knowledge to generate more accurate and contextually appropriate responses. This architecture enables LLMs to better handle tasks that require up-to-date knowledge or specialized expertise that may not be fully captured during training [9].

RAG integrates external knowledge to mitigate the hallucination phenomenon in LLMs, effectively reducing the generation of inaccurate outputs. In QA tasks, queries can be categorized as either single-hop or multi-hop [10]. Early research on retrieval-augmented LLMs predominantly addresses single-hop queries [11,12]. In these cases, answers are generally located within a single document; thus, the retrieval process involves identifying a relevant document based on the query, which is then incorporated into question-answering models to generate a response. However, unlike single-hop question answering, certain queries require linking and combining information from multiple documents. This task t cannot be accomplished through a straightforward, single-step retrieval and response process. For instance, a query such as "What type of fertilizer should I use at the moment for my crop located in Hubei Province?" requires three distinct reasoning steps to answer (what fertilizer, what crop, and what time). Applying a compatible knowledge-retrieving strategy according to the complexity of user queries is a feasible approach. However, issues regarding retrieving efficiency and accuracy still hinder the complex queries using RAG in such a fashion. The following research question arises: how do we maintain fast knowledge retrieval and high accuracy for queries of various complexity using RAG, especially in multi-turn interactions?

This study contributes the following to address the above issues:

(1) This study proposes a Layered Query Retrieval framework that adaptively handles both simple and complex queries, balancing both efficiency and accuracy in responses.
(2) We designed a relevance classifier to filter out irrelevant documents from the retrieved ones, preventing the large model's responses from being influenced by too much irrelevant information.
(3) We conducted experiments on multiple public datasets, and the results show that our model improves accuracy and F1 score by over 10% compared to state-of-the-art methods.

2. Related Work

Large models acquire extensive world knowledge during pre-training. However, since the exact data used in pre-training is unknown, it becomes difficult to assess whether the model can reliably answer user queries. To ensure that responses are well grounded, external knowledge is provided beforehand, allowing the model to search for relevant information within that context. In existing research, questions are typically classified by complexity into single-hop and multi-step categories. Single-hop questions require only a single retrieval from an external knowledge base to generate an answer, whereas multi-step questions are more complex, involving multiple retrievals to iteratively refine the response until the final answer is reached.

Recent research has placed growing emphasis on multi-step and multi-reasoning question answering, which requires iterative interactions between LLMs and retrievers. This approach, while effective for handling complex queries, comes with significant computational costs [13,14].

Conventional retrieval-augmented LLM methods often exhibit inefficiencies when handling queries of varying complexity. They tend to be overly complex for simple queries, leading to unnecessary computational overhead, while being insufficiently robust for more

complex queries. The current issue arises from the use of a one-size-fits-all approach, where all queries are treated as either single-hop or multi-hop, leading to resource inefficiencies [15]. Such a rigid approach is not sufficient to cover the complexity of diverse queries, as simpler queries may waste resources, and more complex ones may not be adequately addressed in a single pass. Additionally, the process of merging retrieved documents can lead to inconsistencies or contradictions in the information presented, which affects the accuracy of the final response. Multiple documents returned after external knowledge retrieval may contain conflicting data or irrelevant information, making it difficult for the model to synthesize a coherent and accurate answer [16].

Some efforts have been made to filter out inconsistent or irrelevant information [17,18], e.g., RAFT [19]. RAFT trains models to ignore irrelevant distractor documents and focus on citing the correct information from relevant documents to answer questions. This, along with its chain-of-thought (COT)-style responses, improves the model's reasoning capabilities. Such a solution might mitigate the inconsistencies in response, but it still lacks a mechanism for processing multi-hop queries. Another direction has focused on refining the retrieval process in multi-step QA [20,21]. Trivedi, Balasubramanian, Khot, and Sabharwal [14] proposed IRCoT, which integrates retrieval with the steps in a CoT, using the CoT to guide the retrieval process and incorporating the retrieved information to enhance the CoT reasoning. However, this iterative method can result in efficiency issues due to excessive branching of possible solutions [22]. Tang and Yang [10] highlighted the limitations of current RAG methods, demonstrating that they struggle with multi-hop queries using datasets specifically designed for such tasks.

A key challenge in improving both efficiency and accuracy lies in dynamically selecting appropriate strategies based on the complexity of user queries. Some approaches, e.g., adaptive-RAG [23], have been developed for query classification to optimize resource allocation. The adaptive-RAG classifies queries to optimize resource allocation by selecting the best strategy for simple and complex questions. Despite its potential, adaptive-RAG still faces unresolved issues, such as refining query classification, filtering inconsistent information during multi-turn interactions, and improving strategies for integrating information from multiple sources to ensure reliability. Additionally, this approach has yet to be tested in specialized domains, such as agricultural technology question answering. The limitations of current RAG technologies are listed in Table 1.

Table 1. Comparison of current RAG methods.

Category	RAG Name	Limitations
Regular RAG	RAG	• Static retrieval processes • Latency issues with large-scale corpora
Knowledge-filtered RAG	RAFT, IRCoT	• Performing poorly on multi-hop queries • Fixed retrieval modules may struggle with out-of-domain queries • Computationally expensive to train and fine-tune
RAG adaptive to query complexity	Adaptive-RAG	• Lacking retrieval robustness • Inconsistent information during multi-turn interactions

In this work, we address the limitations of query and complexity classification in Adaptive-RAG by proposing a framework for categorizing simple and complex queries, termed Layered Query Retrieval (LQR). Specifically, the framework contains a semantic rule-based approach to distinguish between different levels of multi-hop queries. For multi-hop queries, the process begins by parsing the user's query for keywords, followed by a keyword-based document retrieval. Subsequently, we employ a natural language inference (NLI) model to assess whether the retrieved document is relevant to the query.

The system then locates the next-level document based on the initial findings, again using NLI to verify its relevance before combining all relevant results for final input into the model. We test our RAG approach on an agriculture domain-specific dataset for multi-hop queries and compare it with state-of-the-art RAG methods.

3. Method

The proposed LQR framework focuses on three key aspects: dataset construction, query complexity identification, and adaptive multi-retrieval. The framework consists of four stages: dataset construction, query complexity identification, adaptive retrieval, and LLM inference, as illustrated in Figure 1. In the dataset construction stage, parallel dataset generation processes were carried out with the assistance of a state-of-the-art LLM GPT-4o to create training data for both the query complexity classifier and the relevant classifier for multi-retriever results. In the query complexity identification stage, we developed a query complexity classifier to determine whether the user query can be decomposed into multiple questions. In the adaptive retrieval stage, either a single-step strategy or a multi-step strategy is applied based on the complexity of the query for document retrieval. Finally, in the LLM inference stage, the input query and the retrieved documents are combined using a specifically designed prompt. Details of each stage are described in the following sections.

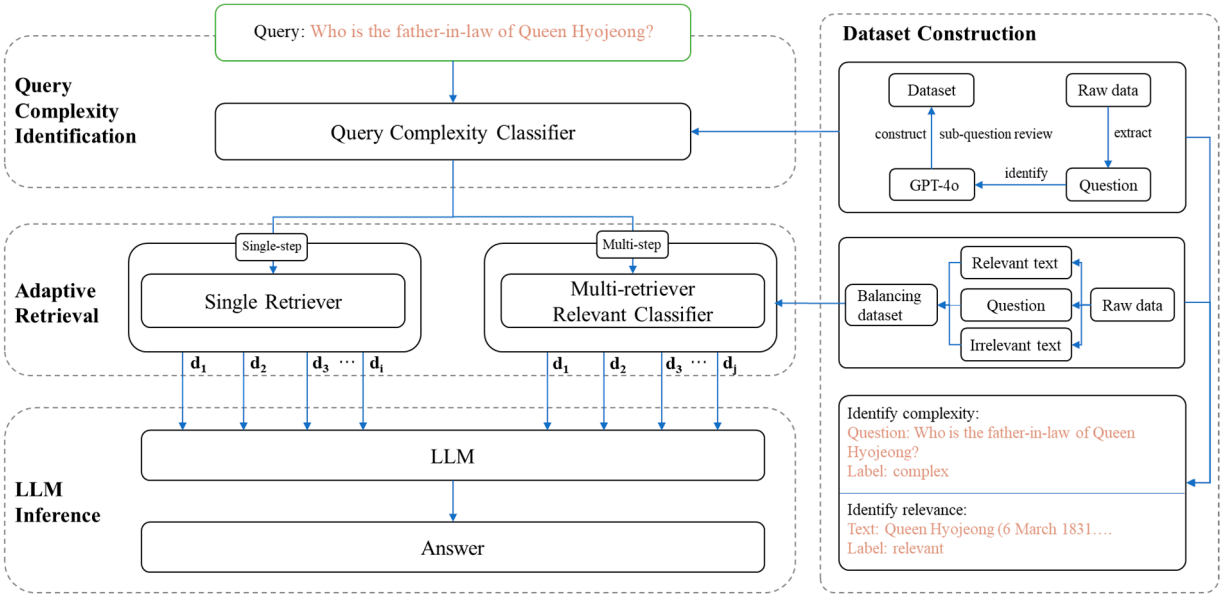


Figure 1. Workflow of the proposed LQR framework.

3.1. Query Complexity Identification

Multi-step questions can typically be decomposed into multiple single-hop questions. To determine the complexity of a user's query, we need to assess whether the query can be broken down into simpler parts. If it can be decomposed into multiple questions, it is classified as a multi-step question; otherwise, it is considered a single-hop question. We train a complexity classifier, where the input is the user query q and the output is the complexity category $c \in \{0, 1\}$. Here, 0 indicates that the query cannot be further decomposed, while 1 indicates that the query requires further breakdown.

$$c = Classifier(q),\ c \in \{0, 1\}$$

The classifier is built on top of the model by adding a linear layer followed by a Sigmoid activation function to assess query complexity.

$$y = \sigma(Wx + b)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The complexity classifier is a smaller model. Since there is currently no publicly available dataset for complexity assessment, we construct the training dataset using a combination of a large model and manual review. We train the complexity classifier on the automatically generated dataset and use it to identify the complexity of queries during inference.

3.2. Adaptive Retrieval

3.2.1. Single-Step Strategy

We adopt different strategies for different types after identifying the complexity of the query. For simple queries, this can be directly retrieved from an external knowledge base using a retriever. Specifically, let us assume the knowledge base is D, the query is q, and the retriever is $Retriever()$. The retriever uses q as input to search the knowledge base and returns relevant documents d_i, as shown in Figure 1.

$$rd = Retriever(q; D)$$

where rd is the set of relevant documents, $rd \in \{d_1, d_2, \ldots, d_i\}$.

The retriever can use any off-the-shelf retrieval model. However, for a fair comparison with other studies, we employed the BM25 retriever. During the retrieving process, documents are ranked according to their BM25 score. For a given document D and a query Q, the BM25 score is calculated as follows:

$$\text{BM25}(D, Q) = \sum_{t \in Q} \text{IDF}(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

where:

- t: A term in the query Q.
- $f(t, D)$: The term frequency of t in document D.
- $|D|$: The length of the document D (in terms of number of words).
- avgdl: The average document length in the corpus.
- k_1: A tunable parameter that controls the term frequency saturation.
- b: A tunable parameter that controls the influence of document length normalization.
- $\text{IDF}(t)$: The inverse document frequency of term t.

The inverse document frequency $\text{IDF}(t)$ adjusts the weight of terms based on their rarity in the document corpus. It is calculated as follows:

$$\text{IDF}(t) = \log\left(\frac{N - n(t) + 0.5}{n(t) + 0.5} + 1\right)$$

where:

- N: The total number of documents in the corpus.
- $n(t)$: The number of documents containing the term t.

3.2.2. Multi-Step Strategy

Simple queries are relatively straightforward, but complex queries require more unknown information. If the query is used directly for retrieval, it is difficult to find all the relevant documents. Complex queries can be broken down into multiple simple queries. To retrieve useful knowledge and filter out irrelevant information, we use a small model to break down the complex query into two sub-questions, q_1 and q_2, and then perform multi-stage retrieval to gather as many relevant documents as possible.

After decomposing the complex query using a small model, two logically progressive simple queries are formed, namely q_1 and q_2. For each query, as shown in Figure 2, we first

use q_1 to retrieve n relevant documents, i.e., $\{d_1, \ldots, d_n\}$. At this point, it is still insufficient to answer the query, so we perform a second retrieval using $\{q_1, q_2, d_i\}$ as a whole to obtain m relevant documents, i.e., $\{d_1, \ldots, d_m\}$. Compared to the first, the second retrieval introduces more information and retrieves some documents more closely related to the query, but it also returns irrelevant and redundant documents. Irrelevant information can affect the large model's ability to find answer clues and impact the question-answering process.

$$RD^1 = Retriever^1(q_1)$$

$$RD^2 = Retriever^2(q_1, q_2, d_i), d_i \in RD^1$$

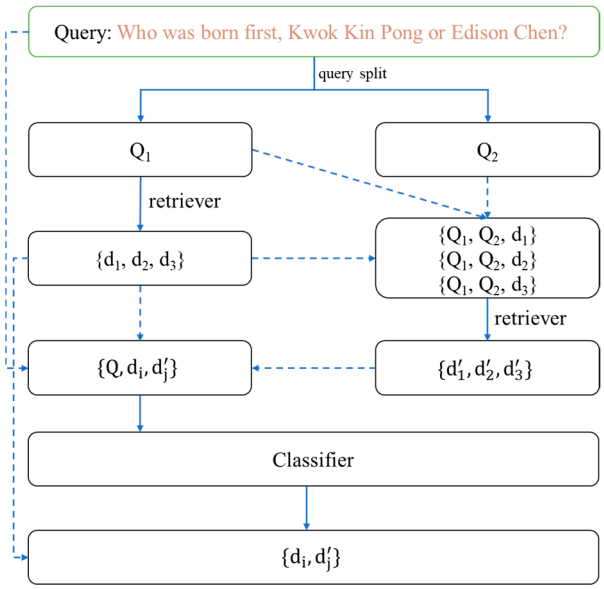


Figure 2. The complex query process structure. Arrows with solid lines denote the processing step; arrows with dotted lines denote data flow.

To remove irrelevant documents, we filter the relevant documents by calculating whether d_j contributes to answering the query. As illustrated in Figure 1, we have developed a balanced dataset in which each entry comprises a question, relevant texts, and irrelevant texts designed to train the relevance model for information filtering.

$$rc = Classifier(q, d_j),\ d_j \in RD^2,\ rc \in \{0, 1\}$$

where 0 indicates that document d_js relevant to q, and 1 indicates it is not relevant. The classifier utilized here adopts the same methodology as the query classifier but is built upon a different base model. Notably, the query classifier and the relevance classifier are designed to fulfill distinct purposes.

3.3. LLM Inference

In the reasoning process, the LLM requires both the input query and the retrieved relevant external documents. Since the LLM is sensitive to the positional information of the text, we place key documents closer to the query, which helps the model find the answer.

As for prompt design, we consider role definition, thought guidance, and output constraints to optimize the prompt, ultimately arriving at a better prompt.

$$answer = LLM(prompt,\ d_1, d_2, \ldots)$$

4. Experiment

This section provides a detailed overview of the datasets, evaluation metrics, experimental parameter settings, and baseline models employed in the experiments.

4.1. Dataset

To validate the effectiveness of our proposed method, we employed two types of datasets: single-hop datasets and multi-hop datasets. The single-hop datasets include SQuAD v1.1 [24], Natural Questions [25], and TriviaQA [26]. The multi-hop datasets include MuSiQue [27], HotpotQA [28], and 2WikiMultiHopQA [29].

In addition, we constructed several domain-specific datasets related to grassland sheep to evaluate the performance of the method on these specialized datasets.

Throughout the proposed process, two classifier models were employed. We selected 3000 entries from public datasets and utilized GPT-4 along with manual review methods to create a complexity dataset of 3000 entries and a relevance dataset of 6000 entries. The relevance dataset was balanced, consisting of 3000 positive and 3000 negative samples.

4.2. Evaluation Metrics

In evaluating the models, both retrieval and generation tasks are considered, requiring the assessment of both accuracy and efficiency. We use accuracy and F1 score to measure the models' effectiveness, while the number of retrievals and the time taken are used to evaluate efficiency. Accuracy reflects whether the predicted results contain the correct answer, while the F1 score assesses the overlap between the predicted results and the correct answers.

4.3. Experiment Details

We follow the experimental protocol outlined by Jeong et al. [23] to ensure a fair comparison with other methods. In our experiments, we use the same retriever, BM25, and maintain consistency between the test set and the retrieval corpus. Since the texts are in English, we conduct the experiments using the Llama3-8B model.

In our overall framework, we employ two classification models. The first classification model is designed to analyze short texts and categorize them based on complexity. For this task, we selected the RoBERTa model, which, after training, achieved accuracy and F1 scores above 85%. The second classification model is focused on processing longer text inputs, for which we selected the Longformer [30] model, capable of handling inputs up to 4096 tokens. After training, this model achieved accuracy and F1 scores exceeding 92%.

This experiment was conducted on a Linux system using an A800 GPU with 80 GB of memory. Both classifiers were trained with a learning rate of 5×10^{-5} using the AdamW optimizer, over 10 epochs. The learning rate warm-up was performed using cosine scheduling, and the batch size was set to 32. The best-performing models were selected from these training runs.

4.4. Baseline

This study proposes a novel retrieval-augmented generation model strategy. To comprehensively compare the effectiveness of this approach, we use four baseline strategies: the non-retrieval strategy, the single-step retrieval strategy, the adaptive retrieval strategy, and the multi-step retrieval strategy.

The non-retrieval strategy refers to answering queries without retrieving external knowledge, relying solely on the LLM for responses. The single-step retrieval strategy enhances the LLM by retrieving external knowledge once for each query. The adaptive retrieval strategy applies different retrieval methods depending on the complexity of the query. Finally, the multi-step retrieval strategy [14] involves multiple rounds of retrieval for a query until a satisfactory answer is obtained.

The adaptive retrieval strategy methods we explored in this study include adaptive retrieval, self-RAG, and adaptive-RAG. The adaptive retrieval strategy [31] sets a threshold

on the validation set to determine whether to employ RAG. The self-RAG strategy [32] introduces special tokens, namely retrieval tokens and reflection tokens, to assess the quality of generated responses and decide whether retrieval is necessary. The adaptive-RAG strategy [23] trains a complexity classifier to evaluate the complexity of the query and, based on this evaluation, selects different processing strategies accordingly.

5. Experimental Results and Analyses

This section is dedicated to the analysis of experimental results, the execution of ablation studies, and the presentation of a case study.

5.1. Main Results

We conducted experiments on three multi-hop datasets and compared the results with those of other methods. We classified previous studies into three groups: simple, complex, and adaptive. Table 2 shows that the accuracy of non-retrieval-augmented methods is low, while multi-hop retrieval methods achieve high accuracy but exhibit low efficiency, requiring considerable processing time. Adaptive retrieval methods, by distinguishing between simple and complex queries, apply different strategies and achieve relatively high accuracy.

Table 2. The results obtained from our proposed method on the multi-hop datasets using the Llama3-8B model are compared with those of other methods. Our results are highlighted in bold for emphasis.

Type	Method	MuSiQue				HotpotQA				2Wiki			
		F1	Acc	Step	Time	F1	Acc	Step	Time	F1	Acc	Step	Time
Simple	Non-Retrieval	10.70	3.20	0.00	0.11	22.71	17.20	0.00	0.11	32.04	27.80	0.00	0.10
	Single-Step Retrieval	22.80	15.20	1.00	1.00	46.15	36.40	1.00	1.00	47.90	42.80	1.00	1.00
Complex	Multi-Step Retrieval	31.90	25.80	3.60	7.58	56.54	47.00	5.53	9.38	58.85	55.40	4.17	7.37
	Adaptive Retrieval	15.80	8.00	0.50	0.55	32.22	25.00	0.50	0.55	39.44	34.20	0.50	0.55
Adaptive	Self-RAG	8.10	12.00	0.73	0.51	17.53	29.60	0.73	0.45	19.59	38.80	0.93	0.49
	Adaptive-RAG	31.80	26.00	3.22	6.61	53.82	44.40	3.55	5.99	49.75	46.40	2.63	4.68
	Our Method	**41.97**	**26.55**	**3.88**	**4.45**	**69.96**	**53.8**	**3.42**	**3.92**	**54.65**	**37.60**	**3.88**	**4.32**

Our proposed method demonstrates both high accuracy and efficiency, particularly on the HotpotQA dataset, where it surpasses the Adaptive-RAG method, improving accuracy by 9.4% and the F1 score by 16.14%. The step metric reveals that our method involves multiple retrievals. For complex queries, retrieving information just once may not provide sufficient data; hence, our approach conducts multiple queries to address this gap. In contrast, the non-retrieval, single-step retrieval, and multi-step retrieval methods apply a single strategy for any given query. Our adaptive retrieval method, however, enhances model accuracy while only marginally increasing the time spent on retrieval and generation.

We conducted experiments using our proposed research method on three single-hop datasets. Table 3 shows that our method outperforms others by more than 10% in both accuracy and F1 score. Upon analyzing the experimental results, we observed that some single-hop queries are relatively complex and can be decomposed into two sub-queries, allowing for multiple retrievals to gather more relevant information.

Table 3. The results obtained from the Llama3-8B model on the single-hop datasets are compared with those of other methods. Our results are highlighted in bold for emphasis.

Type	Method	SQuAD				Natural Questions				TriviaQA			
		F1	Acc	Step	Time	F1	Acc	Step	Time	F1	Acc	Step	Time
Simple	Non-Retrieval	10.50	5.00	0.00	0.11	19.00	15.60	0.00	0.13	31.80	27.00	0.00	0.13
	Single-Step Retrieval	39.30	34.00	1.00	1.00	47.30	44.60	1.00	1.00	62.40	60.20	1.00	1.00
Complex	Multi-Step Retrieval	35.60	29.60	4.52	9.03	47.80	44.20	5.04	10.18	62.40	60.20	5.28	9.22
	Adaptive Retrieval	23.10	17.60	0.50	0.55	36.00	33.00	0.50	0.56	46.90	42.60	0.50	0.56
Adaptive	Self-RAG	11.20	18.40	0.63	0.50	39.00	33.60	0.63	0.17	29.30	57.00	0.68	0.45
	Adaptive-RAG	38.30	33.00	1.37	2.02	47.30	44.60	1.00	1.00	60.70	58.20	1.23	1.54
	Our Method	**46.50**	**43.40**	**3.73**	**4.30**	**60.16**	**58.8**	**3.24**	**3.41**	**70.20**	**68.80**	**3.84**	**4.50**

For simple queries, apart from the multi-step retrieval method, all other methods classify them as simple and do not require multiple retrievals. Our proposed research method places greater emphasis on classifying the complexity of queries, demonstrating strong sensitivity to this factor.

Through a comparative analysis of three public datasets, we observed that the LQR method demonstrates superior performance. The following section focuses on examining the contribution of each component within the LQR method.

5.2. Ablation Study

This section analyzes the various modules of the proposed research method to evaluate their contributions to the overall architecture.

- The effect of the complexity classifier

In Tables 2 and 3, the single-step retrieval and multi-step retrieval methods do not assess the complexity of the queries, applying a single retrieval generation strategy for all cases. Our proposed method demonstrates significant improvements in both performance and efficiency, validating the effectiveness of the adaptive retrieval strategy.

- The effect of the first-stage retriever

We varied the top-k values while keeping other parameters constant. As shown in Table 4, increasing the top-k values results in a decrease in efficiency, accuracy, and F1 score. However, this impact is relatively small. It is hypothesized that as top-k increases, the volume of retrieved information rises, leading to the accumulation of irrelevant and redundant data, which can negatively affect the LLM's responses. Furthermore, when the LLM receives longer texts, the processing time increases, requiring more time to generate responses.

Table 4. Experimental results of the Llama3-8B on HotpotQA under different Top-k conditions with the first-stage retriever.

Top-k	F1	Acc	Step	Time
3	**51.64**	**34.80**	**3.43**	**4.30**
4	47.06	31.05	4.78	5.78
5	46.46	30.26	6.78	8.56

- The effect of different LLMs

We conducted experiments using the GPT-4o model on three multi-hop datasets. Table 5 demonstrates that, compared to the Llama3-8B model, the GPT-4o model improves the F1 score and accuracy by an average of over 10% across the multi-hop datasets. The use of a more powerful model in our proposed method leads to better performance. This also indirectly suggests that our strategy effectively retrieves relevant information, as there may be instances where the LLM alone does not identify relevant clues.

Table 5. Results of Llama3-8B and GPT-4o on the MusiQue, HotpotQA, and 2Wiki multi-hop datasets.

Model	MuSiQue		HotpotQA		2Wiki	
	F1	Acc	F1	Acc	F1	Acc
Llama-8B	41.97	26.55	69.96	53.80	54.65	37.60
GPT4o	**59.94**	**42.80**	**79.37**	**68.60**	**59.55**	**42.40**

- The effect of the relevance classifier

Table 4 shows that as the top-k value increases, more information is retrieved, leading to an increase in noisy data. This raises the question: can the relevance classifier help reduce irrelevant information? Table 6 indicates that when the relevance classifier is removed, the F1 score and accuracy decrease by an average of 2.38% and 1.67%, respectively. This highlights the necessity of incorporating the relevance classifier. Enhancing the performance of the relevance classifier will also be one of our key research directions in the future.

Table 6. Experimental results of Llama3-8B on three datasets without the relevance classifier. "rcls" represents the relevance classifier.

Model	MuSiQue		HotpotQA		2Wiki	
	F1	Acc	F1	Acc	F1	Acc
Ours	**41.97**	**26.55**	**69.96**	**53.80**	**54.65**	**37.60**
w/o rcls	37.58	23.13	67.63	52.60	54.22	37.20

5.3. Case Study

We conducted a case study to observe the results of our proposed method across various datasets. We selected three cases from both single-hop and multi-hop datasets to examine whether the simple queries defined in the data require decomposition and to assess the output performance of the large model. Additionally, we evaluate the effectiveness of the retrieval strategy.

Table 7 presents the inference results using Llama3-8B and GPT-4o on three single-hop datasets. The data in Table 7 indicate that some queries in the single-hop datasets also need to be broken down into multiple simpler query statements. The examples in Table A1 in Appendix A demonstrate that after splitting the user queries, more relevant information can be retrieved.

Table 7. Examples of responses from our proposed method using Llama3-8B and GPT-4o on three single-hop datasets.

Dataset	Query and Answer	Llama3-8B	GPT-4o	Complexity Classification
SQuAD	**Query:** Who was the CEO of GE when it announced its Ecoimagination program? **Answer:** Jeff Immelt	Jeff Immelt	Jeff Immelt	Complex Q1: When did GE announce its Ecoimagination program? Q2: Who was the CEO of GE at that time?
Natural Questions	**Query:** who wrote just you and me by chicago **Answer:** James Pankow	James Pankow	James Pankow	Simple
TriviaQA	**Query:** Which studio owns the rights to Bugs Bunny? **Answer:** Warner Bros	Warner Bros	Warner Bros	Simple

Table 8 presents the inference results using Llama3-8B and GPT-4o on three multi-hop datasets. From Table 8, it can be observed that for complex queries, retrieval can be per-

formed directly without splitting, which helps reduce the impact of irrelevant information. By combining Tables 7, 8, A1, and A2, it is evident that GPT-4o is more adept at extracting key information from long contexts. The experiments revealed that contextual information can assist the LLM in leveraging stored world knowledge, thereby enhancing the accuracy of responses.

Table 8. Examples of responses from our proposed method using Llama3-8B and GPT-4o on three multi-hop datasets.

Dataset	Query and Answer	Llama3-8B	GPT-4o	Complexity Classification
MusiQue	**Query:** What field does the author of Parallel Worlds work in? **Answer:** physics	Physics	Science	Complex Q1: Who is the author of Parallel Worlds? Q2: What field does the author of Parallel Worlds work in?
HotpotQA	**Query:** Gerd Neggo trained under the founder of which type of dance analysis? **Answer:** Laban Movement Analysis	Laban	Laban Movement Analysis	Simple
2Wiki	**Query:** Which country the founder of Team Red Bull (Nascar Team) is from? **Answer:** Austria	Austrian	Austria	Complex Q1: Who founded the Team Red Bull NASCAR team? Q2: What country is the founder of the Team Red Bull NASCAR team from?

6. Discussion

The experimental results on both single-hop and multi-hop datasets demonstrate that the adaptive retrieval approach shows promising results in retrieval-augmented generation. Compared to other adaptive retrieval methods, e.g., self-RAG [32], adaptive-RAG [23], and parametric and non-parametric memory-based methods [31], our proposed LQR not only identifies query complexity but also focuses more on finding documents relevant to the query. Specifically for complex queries, we employ multi-step retrieval, relevance recognition, and other techniques to locate pertinent information while simultaneously filtering out irrelevant content.

In the process of retrieval-augmented generation, it is crucial not only to focus on the performance of the generative model but also to retrieve documents that can effectively answer the query. Retrieval-augmented generation must balance the quality of the generated responses with the time consumed throughout the process. In multi-step methods, continuous retrieval and generation cycles eventually lead to the final answer. However, this approach is time-consuming and resource-intensive. On the other hand, single-step retrieval is faster, but the quality of the generated output is often suboptimal. The challenge becomes even more pronounced when dealing with complex queries—how can we ensure that all relevant documents are retrieved?

To further evaluate the model, we conducted a case study using data related to sheep. For instance, when presented with the questions "What is the scientific name of a domestic sheep?" and "Which breed of sheep is native to China and known for producing fine wool?", the LLM demonstrated its ability to not only retrieve relevant information from the context but also leverage its pre-trained knowledge to provide more comprehensive responses. This highlights the role of the LLM's pre-trained world knowledge in enhancing answer quality. However, the case study also revealed challenges. Irrelevant information occasionally emerged during the retrieval process, which could affect the quality of the model's responses. Despite these issues, LQR demonstrated greater flexibility compared to single-step and multi-step retrieval methods. Nonetheless, further optimization of LQR's individual components remains necessary to fully realize its potential.

This study reveals that for particularly complex queries, the proposed method struggles to retrieve effective documents. How to optimize retrieval and improve document relevance recognition will be key research directions in the future. Our experiments indicate that using larger models yields higher accuracy. There are two reasons for this. First, larger models possess stronger comprehension capabilities, allowing them to find clues in the documents that lead to the correct answer. Interestingly, we observed that certain queries received correct answers even without retrieving all relevant documents. This can be attributed to the second reason: larger models, guided by context and the user's query, leverage their pre-trained world knowledge to provide answers. An important future research direction will be to explore how to utilize existing knowledge to guide models in responding to queries effectively.

7. Conclusions

This study presents a novel retrieval-augmented generation framework LQR that is designed to accommodate varying query complexities. Specifically, we employ tailored retrieval strategies based on an assessment of query complexity. Our proposed method emphasizes complexity classification, query retrieval strategies, and relevance analysis, utilizing a custom-built training dataset to train smaller models that assist the LLM in efficiently retrieving relevant information. We tested our approach across various single-hop and multi-hop datasets, showing substantial improvements in both accuracy and efficiency compared to existing single-step, multi-step, and adaptive methods. The method we propose achieves notable improvements in both accuracy and efficiency, especially on the HotpotQA dataset, where it outperforms the adaptive-RAG approach with a 9.4% increase in accuracy and a 16.14% boost in the F1 score. The case study revealed that irrelevant information can arise during retrieval, impacting the model's responses. Still, LQR proves more flexible than single- or multi-step retrieval, though its components require further refinement. In general, the LQR achieves an effective balance between retrieval speed and the precision of the LLM's generated responses.

Author Contributions: Conceptualization, J.H., M.W., Y.C. and J.L.; Methodology, J.H., M.W., Y.C. and J.L.; Validation, J.H., L.C. and T.W.; Formal analysis, L.C., T.W. and H.L.; Investigation, J.H., M.W. and J.W.; Data curation, J.H. and J.W.; Writing—original draft, J.H. and M.W.; Writing—review and editing, M.W., Y.C. and J.L.; Visualization, J.H., M.W. and L.C.; Supervision, T.W., H.L. and J.W.; funding acquisition, Y.C. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program Project, grant number 2023YFD1600305; the Beijing Smart Agriculture Innovation Consortium Project, grant number BAIC10-2024-E18; the Project of Northern Agriculture and Livestock Husbandry Technical Innovation Center, Chinese Academy of Agricultural Sciences, grant number BFGJ2022006; and the Chinese Academy of Agricultural Sciences Science and Technology Innovation Project Grant, grant number CAAS-ASTIP-2024-AII.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Examples of retrieval results from the single-step retrieval method and our proposed method on three single-hop datasets. The red font indicates answer clues.

Dataset	Query and Answer	Single-Step	Our Method
SQuAD	**Query:** Who was the CEO of GE when it announced its Ecoimagination program? **Answer:** Jeff Immelt	Doc 1: GE Digital across field service processes. This acquisition closed in January 2017. In November 2016, GE Digital also launched its collaborative app development program. In January, GE completed its acquisition of ServiceMax, a cloud-based field service management company. In December, GE announced its intent to establish an independent IIoT company, taking over large parts of GE Digital. At the same time, GE announced its intended sale of a majority stake of ServiceMax. In February 2016, GE announced the GE Digital Alliance Program, a program dedicated to building the digital industrial ecosystem. This new program encompassed existing agreements with Accenture, AT&T, Cisco, SoftBank, Doc 2: GE Healthcare in the Technology Research for Independent Living Centre (TRIL). The Irish centre seeks to enhance independence for elderly people through technological innovation. In July 2015, GE Healthcare partnered with the 2015 CrossFit Games to provide athletes with mobile imaging equipment. In January 2016, it was announced GE Healthcare's global headquarters will move to Chicago effective early 2016. In June 2017, GE announced Kieran Murphy as the new CEO of GE Healthcare, with former CEO John Flannery's appointment as CEO of GE. In April 2018, GE announced the sale of several healthcare information technology assets for $1.05 billion to Veritas Capital. Doc 3: GE Healthcare of filtration products and technologies at 270p per share in cash for each Whatman share, valuing Whatman at approximately £363 million (approximately $713 million.) In July 2008, Joseph Hogan announced his intent to leave his post as CEO of GE Healthcare to take the role of CEO at ABB. On 17 July 2008, GE Healthcare announced John Dineen had been chosen to replace outgoing CEO Joseph Hogan. Mr. Dineen had been head of GE's Transportation division since 2005. On 24 March 2010, GE Healthcare announced acquisition of MedPlexus. In late April 2010, GE Healthcare announced it was investing €3 million	Doc 1: Wiregrass Electric Cooperative first manager of the cooperative. He served until 1958. Subsequent leadership was as follows: • B.L Woodham, general manager, 1953–1968 • Bruce Dyess, general manager, 1969–1985 • Jerry Mosley, general manager, 1985–2005 • Michael McWaters, CEO, 2005–2013 • Les Moreland, CEO, 2014–present At Wiregrass Electric Cooperative, the consumer/member is the basis of the organization. Each person receiving electric service must be a member of the cooperative. Annual membership meetings are held where members receive reports on the operation of the system and elect trustees to handle the affairs of the Doc 2: James McNerney Information Services. McNerney competed with Bob Nardelli and Jeff Immelt to succeed the retiring Jack Welch as chairman and CEO of General Electric. When Immelt won the three-way race, McNerney and Nardelli left GE (as was Welch's plan); McNerney was hired by 3M in 2001. From 2001 to 2005, McNerney held the position as chairman of the board and CEO of 3M, a $20 billion global industrial company with leading positions in electronics, telecommunications, industrial, consumer and office products, health care, safety and other businesses. On 30 June 2005 The Boeing Company hired McNerney as the Chairman, President and CEO. Doc 3: United States withdrawal from the Paris Agreement with the United Nations to submit a plan for limiting American climate-change emissions in accord with the Paris Agreement guidelines. Goldman Sachs CEO Lloyd Blankfein described Trump's decision as "a setback for the environment and for the U.S.'s leadership position in the world". General Electric CEO Jeff Immelt stated that "climate change is real". Multiple tech company executives—including Google CEO Sundar Pichai, Microsoft President and Chief Legal Officer Brad Smith, Apple CEO Tim Cook, Facebook CEO Mark Zuckerberg, and General Electric CEO Jeff Immelt—condemned the decision. Microsoft's Satya Nadella said Microsoft believes that "climate change is an urgent issue that', Doc 4: Robert Nardelli often known as "Little Jack", after his mentor Jack Welch, whom Nardelli had ambitions to succeed as CEO of GE. When Jack Welch retired as chairman and CEO of GE, a lengthy and well-publicized succession planning saga ensued. Nardelli competed with James McNerney and Jeffrey R. Immelt to succeed Welch. With Immelt winning the three-way race, Nardelli and McNerney left GE (as was Welch's plan). About 10 min after Welch let him go, Nardelli received a job offer from Kenneth Langone, who at the time was on the boards of both GE and The Home Depot. Nardelli became CEO of Doc 5: Jeff Immelt when Jack Welch asked Gary Wendt, the head of GE Financial Services, Inc, then a $41 billion empire and GE's most profitable division, to resign, on the theory that he would overshadow Welch's other candidates. Immelt surpassed Robert Nardelli, who was so close to Welch that his nickname was "Little Jack", and James McNerney; Nardelli would leave GE to head Home Depot and then Chrysler, and McNerney left GE to lead 3M (before going on to lead Boeing). Immelt was selected by the GE Board in November 2000, to succeed Welch as CEO of GE; he served as GE's president

Table A1. *Cont.*

Dataset	Query and Answer	Single-Step	Our Method
Natural Questions	**Query:** who wrote just you and me by chicago **Answer:** James Pankow	Doc 1: "Just You 'n' Me 'n' Me" was written after a fight between Pankow and his future wife Karen: "Just You 'n' Me" was the very last song played by Chicago radio station WLS before switching to a talk radio format in 1989. Just You 'n' Me "Just You 'n' Me" is a song written by James Pankow for the group Chicago and recorded for their fifth studio album "Chicago VI" (1973). The lead vocals are sung by bassist Peter Cetera. The second single released from that album, it was more successful than the first single, "Feelin' Stronger Every Day", reaching #4 on the U.S. Doc 2: 'Just You 'n' Me Just You 'n' Me "Just You 'n' Me" is a song written by James Pankow for the group Chicago and recorded for their fifth studio album "Chicago VI" (1973). The lead vocals are sung by bassist Peter Cetera. The second single released from that album, it was more successful than the first single, "Feelin' Stronger Every Day", reaching #4 on the U.S. "Billboard" Hot 100 and #1 on the "Cash Box" Top 100. Walter Parazaider plays a soprano saxophone solo during the instrumental section while guitarist Terry Kath uses a wah-wah pedal and phase shifter on his guitar. "Just You" Doc 3: Chicago VI would remain their recording base for the next four years. Robert Lamm authored half of the album's tracks, including his response to some of Chicago's negative reviewers in "Critics' Choice". James Pankow wrote the album's two hits, "Just You 'n' Me" (#4) and "Feelin' Stronger Every Day" (#10). The latter was co-composed with Peter Cetera, who also wrote "In Terms of Two", and sang lead vocal on all three songs. Released in June 1973, "Chicago VI" was another commercial success, spending five weeks at #1 in the US. The band would not chart in the UK at all until 1976's	Doc 1: "Just You 'n' Me 'n' Me" was written after a fight between Pankow and his future wife Karen: "Just You 'n' Me" was the very last song played by Chicago radio station WLS before switching to a talk radio format in 1989. Just You 'n' Me "Just You 'n' Me" is a song written by James Pankow for the group Chicago and recorded for their fifth studio album "Chicago VI" (1973). The lead vocals are sung by bassist Peter Cetera. The second single released from that album, it was more successful than the first single, "Feelin' Stronger Every Day", reaching #4 on the U.S. Doc 2: 'Just You 'n' Me Just You 'n' Me "Just You 'n' Me" is a song written by James Pankow for the group Chicago and recorded for their fifth studio album "Chicago VI" (1973). The lead vocals are sung by bassist Peter Cetera. The second single released from that album, it was more successful than the first single, "Feelin' Stronger Every Day", reaching #4 on the U.S. "Billboard" Hot 100 and #1 on the "Cash Box" Top 100. Walter Parazaider plays a soprano saxophone solo during the instrumental section while guitarist Terry Kath uses a wah-wah pedal and phase shifter on his guitar. "Just You" Doc 3: Chicago VI would remain their recording base for the next four years. Robert Lamm authored half of the album's tracks, including his response to some of Chicago's negative reviewers in "Critics' Choice". James Pankow wrote the album's two hits, "Just You 'n' Me" (#4) and "Feelin' Stronger Every Day" (#10). The latter was co-composed with Peter Cetera, who also wrote "In Terms of Two", and sang lead vocal on all three songs. Released in June 1973, "Chicago VI" was another commercial success, spending five weeks at #1 in the US. The band would not chart in the UK at all until 1976's
TriviaQA	Query: Which studio owns the rights to Bugs Bunny? Answer: Warner bros	Doc 1: Bugs Bunny's Easter Special\nin an Easter bunny suit (finally getting it right), which neither surprises nor disappoints Bugs or Granny, having known all along it was Daffy. This special was released on DVD in 2011, in time for the Easter holiday. Bugs Bunny's Easter Special Bugs Bunny's Easter Special (also known as "The Bugs Bunny Easter Special" and "Bugs Bunny's Easter Funnies") is a "Looney Tunes" television special featuring a number of Warner Bros. cartoons. It originally debuted on the CBS network on 7 April 1977. The Easter bunny is ill, Granny needs to find a replacement for him and suggests Bugs Bunny. Doc 2: Bugs Bunny & Lola Bunny: Operation Carrot Patch\nBugs Bunny & Lola Bunny: Operation Carrot Patch Bugs Bunny & Lola Bunny: Operation Carrot Patch, known in North America as Looney Tunes: Carrot Crazy, is a 1998 Game Boy Color video game starring Bugs Bunny and Lola Bunny. Yosemite Sam, Daffy Duck, The Tasmanian Devil, Marvin the Martian and Elmer Fudd have stolen Bugs and Lola's carrots and hid them in different sets in the nearby studio. Bugs and Lola head through five different "sets" in order to reclaim them. In the game, Bugs and Lola (the player can switch between them at any time) can traverse two levels Doc 3: Bugs Bunny He would not appear in new material on-screen again until "Bugs and Daffy's Carnival of the Animals" aired in 1976. From the late 1970s through the early 1990s, Bugs was featured in various animated specials for network television, such as "Bugs Bunny's Thanksgiving Diet", "Bugs Bunny's Easter Special", "Bugs Bunny's Looney Christmas Tales", and "Bugs Bunny's Bustin' Out All Over". Bugs also starred in several theatrical compilation features during this time, including the United Artists distributed documentary (1975) and Warner Bros.' own releases: "The Bugs Bunny/Road Runner Movie" (1979), "The Looney Looney Looney Bugs Bunny Movie" (1981), (1982)	Doc 1: Bugs Bunny's Easter Special in an Easter bunny suit (finally getting it right), which neither surprises nor disappoints Bugs or Granny, having known all along it was Daffy. This special was released on DVD in 2011, in time for the Easter holiday. Bugs Bunny's Easter Special Bugs Bunny's Easter Special (also known as "The Bugs Bunny Easter Special" and "Bugs Bunny's Easter Funnies") is a "Looney Tunes" television special featuring a number of Warner Bros. cartoons. It originally debuted on the CBS network on 7 April 1977. The Easter bunny is ill, Granny needs to find a replacement for him and suggests Bugs Bunny. Doc 2: Bugs Bunny & Lola Bunny: Operation Carrot Patch Bugs Bunny & Lola Bunny: Operation Carrot Patch Bugs Bunny & Lola Bunny: Operation Carrot Patch, known in North America as Looney Tunes: Carrot Crazy, is a 1998 Game Boy Color video game starring Bugs Bunny and Lola Bunny. Yosemite Sam, Daffy Duck, The Tasmanian Devil, Marvin the Martian and Elmer Fudd have stolen Bugs and Lola's carrots and hid them in different sets in the nearby studio. Bugs and Lola head through five different "sets" in order to reclaim them. In the game, Bugs and Lola (the player can switch between them at any time) can traverse two levels Doc 3: Bugs Bunny He would not appear in new material on-screen again until "Bugs and Daffy's Carnival of the Animals" aired in 1976. From the late 1970s through the early 1990s, Bugs was featured in various animated specials for network television, such as "Bugs Bunny's Thanksgiving Diet", "Bugs Bunny's Easter Special", "Bugs Bunny's Looney Christmas Tales", and "Bugs Bunny's Bustin' Out All Over". Bugs also starred in several theatrical compilation features during this time, including the United Artists distributed documentary (1975) and Warner Bros.' own releases: "The Bugs Bunny/Road Runner Movie" (1979), "The Looney Looney Looney Bugs Bunny Movie" (1981), (1982)

Table A2. Examples of retrieval results from the single-step retrieval method and our proposed method on three multi-hop datasets. The red font indicates answer clues.

Dataset	Query and Answer	Single-Step	Our Method
MusiQue	**Query:** What field does the author of Parallel Worlds work in? **Answer:** physics	Doc 1: Declaration of war by the United States For the United States, Article One, Section Eight of the Constitution says Congress shall have power to... declare War. "However, that passage provides no specific format for what form legislation must have in order to be considered a "declaration of war" nor does the Constitution itself use this term. In the courts, the United States Court of Appeals for the First Circuit, in Doe v. Bush, said: The text of the October Resolution itself spells out justifications for a war and frames itself as an 'authorization' of such a war. in effect saying an authorization suffices for declaration and what some may view as a formal Congressional "Declaration of War" was not required by the Constitution. Doc 2: Timoshenko Medal The Timoshenko Medal, widely regarded as the highest international award in the field of applied mechanics, was established in 1957 in honor of Stephen Timoshenko, world-renowned authority in the field. The Medal commemorates his contributions as author and teacher.	Doc 1: Parallel Worlds (Dave Douglas album) Parallel Worlds is the debut album by trumpeter Dave Douglas released on the Italian Soul Note label in 1993. It features six of Douglas' compositions and compositions by Anton Webern, Kurt Weill, Duke Ellington and Igor Stravinsky performed by Douglas, Mark Feldman, Erik Friedlander, Mark Dresser and Michael Sarin, Doc 2: Parallel Worlds (book) Parallel Worlds: A Journey Through Creation, Higher Dimensions, and the Future of the Cosmos is a popular science book by Michio Kaku first published in 2004. Doc 3: José Antunes Sobrinho Brasília has also been the focus of modern-day literature. Published in 2008, The World In Grey: Dom Bosco's Prophecy, by author Ryan J. Lucero, tells an apocalypticle story based on the famous prophecy from the late 19th century by the Italian saint Don Bosco. According to Don Bosco's prophecy: "Between parallels 15 and 20, around a lake which shall be formed; A great civilization will thrive, and that will be the Promised Land". Brasília lies between the parallels 15° S and 20° S, where an artificial lake (Paranoá Lake) was formed. Don Bosco is Brasília's patron saint. Doc 4: Don Bosco (author) Don Bosco (born in June 1971) is a writer and publisher of fiction books from Singapore. In 2011, he founded Super Cool Books, a publishing company with interest in fantasy and mystery stories for children and young adults. As an active advocate of self-publishing and digital publishing, he has published ebooks, paperbacks and an iPad app.

Table A2. *Cont.*

Dataset	Query and Answer	Single-Step	Our Method
HotpotQA	**Query:** Gerd Neggo trained under the founder of which type of dance analysis? **Answer:** Laban Movement Analysis	Doc 1: Person Analysis Person Analysis is a phase of training needs analysis directed at identifying which individuals within an organization should receive training. Doc 2: Gerd Neggo Gerd Neggo (9 November 1891–1 September 1974) was an Estonian dancer, dance teacher and choreographer. She studied the musical response methods of É. Jaques-Dalcroze, trained under Rudolf von Laban in Hamburg, Germany, and in 1924 established her own dance studio at Tallinn, Estonia, and promoted modern dance and mime based on classical ballet. During the Soviet occupation of Estonia, she and her husband Paul Olak migrated to Sweden. Her contributions to the cultural heritage of Estonia, as the founder of modern dance and mime in her country, is recognised via a scholarship, awarded annually since 2011. Doc 3: Joyce K. Paul Joyce K. Paul is an Indian classical dancer and exercise physiologist. Classically trained in Bharat Natyam, she has also trained briefly in Mohiniattam, the classical dance of Kerala. Paul had her initial dance training under the Bharatanatyam guru, Leela Samson. She is the founder and the Artistic Director of Arpan set up in 2003 in Redmond, WA	Doc 1: Person Analysis Person Analysis is a phase of training needs analysis directed at identifying which individuals within an organization should receive training. Doc 2: Gerd Neggo Gerd Neggo (9 November 1891–1 September 1974) was an Estonian dancer, dance teacher and choreographer. She studied the musical response methods of É. Jaques-Dalcroze, trained under Rudolf von Laban in Hamburg, Germany, and in 1924 established her own dance studio at Tallinn, Estonia, and promoted modern dance and mime based on classical ballet. During the Soviet occupation of Estonia, she and her husband Paul Olak migrated to Sweden. Her contributions to the cultural heritage of Estonia, as the founder of modern dance and mime in her country, is recognised via a scholarship, awarded annually since 2011. Doc 3: Joyce K. Paul Joyce K. Paul is an Indian classical dancer and exercise physiologist. Classically trained in Bharat Natyam, she has also trained briefly in Mohiniattam, the classical dance of Kerala. Paul had her initial dance training under the Bharatanatyam guru, Leela Samson. She is the founder and the Artistic Director of Arpan set up in 2003 in Redmond, WA
2Wiki	**Query:** Which country the founder of Team Red Bull (Nascar Team) is from? **Answer:** Austria	Doc 1: Team Red Bull (NASCAR team) Red Bull Racing Team, also known as Team Red Bull, was a NASCAR team owned by Red Bull founders Dietrich Mateschitz and Chaleo Yoovidhya. The team was based in Mooresville, North Carolina in the United States and was managed by Jay Frye. The team suspended operations on 8 December 2011 and their cars were sold to BK Racing. Doc 2: Helmut Marko Helmut Marko LL.D., (born 27 April 1943) is an Austrian former professional racing driver and current advisor to the Red Bull GmbH Formula One Teams and head of Red Bull's driver development program. Doc 3: Red Bull GmbH Red Bull GmbH is an Austrian company which produces Red Bull energy drink. The company is also known for its sponsorship of a range of sporting events and teams. In 2018, a total of 6.8 billion cans of Red Bull were sold worldwide in over 171 countries. The headquarters of Red Bull GmbH are located in Fuschl am See, Austria.	Doc 1: Team Red Bull (NASCAR team) Red Bull Racing Team, also known as Team Red Bull, was a NASCAR team owned by Red Bull founders Dietrich Mateschitz and Chaleo Yoovidhya. The team was based in Mooresville, North Carolina in the United States and was managed by Jay Frye. The team suspended operations on 8 December 2011 and their cars were sold to BK Racing. Doc 2: Red Bull GmbH Red Bull GmbH is an Austrian company which produces Red Bull energy drink. The company is also known for its sponsorship of a range of sporting events and teams. In 2018, a total of 6.8 billion cans of Red Bull were sold worldwide in over 171 countries. The headquarters of Red Bull GmbH are located in Fuschl am See, Austria. Doc 3: Helmut Marko Helmut Marko LL.D., (born 27 April 1943) is an Austrian former professional racing driver and current advisor to the Red Bull GmbH Formula One Teams and head of Red Bull's driver development program.

References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
2. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
3. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
4. Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; Raffel, C. Large language models struggle to learn long-tail knowledge. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 15696–15707.
5. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
6. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv* **2023**, arXiv:2309.01219.
7. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
8. Li, H.; Su, Y.; Cai, D.; Wang, Y.; Liu, L. A survey on retrieval-augmented text generation. *arXiv* **2022**, arXiv:2202.01110.
9. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.* **2023**, *24*, 1–43.
10. Tang, Y.; Yang, Y. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv* **2024**, arXiv:2401.15391.
11. Lazaridou, A.; Gribovskaya, E.; Stokowiec, W.; Grigorev, N. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv* **2022**, arXiv:2203.05115.
12. Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; Shoham, Y. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1316–1331. [CrossRef]
13. Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.A.; Lewis, M. Measuring and narrowing the compositionality gap in language models. *arXiv* **2022**, arXiv:2210.03350.
14. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv* **2022**, arXiv:2212.10509.
15. Vu, T.; Iyyer, M.; Wang, X.; Constant, N.; Wei, J.; Wei, J.; Tar, C.; Sung, Y.-H.; Zhou, D.; Le, Q. Freshllms: Refreshing large language models with search engine augmentation. *arXiv* **2023**, arXiv:2310.03214.
16. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Cui, B. Retrieval-augmented generation for ai-generated content: A survey. *arXiv* **2024**, arXiv:2402.19473.
17. Liu, Y.; Peng, X.; Zhang, X.; Liu, W.; Yin, J.; Cao, J.; Du, T. RA-ISF: Learning to Answer and Understand from Retrieval Augmentation via Iterative Self-Feedback. *arXiv* **2024**, arXiv:2403.06840.
18. Yoran, O.; Wolfson, T.; Ram, O.; Berant, J. Making retrieval-augmented language models robust to irrelevant context. *arXiv* **2023**, arXiv:2310.01558.
19. Zhang, T.; Patil, S.G.; Jain, N.; Shen, S.; Zaharia, M.; Stoica, I.; Gonzalez, J.E. Raft: Adapting language model to domain specific rag. *arXiv* **2024**, arXiv:2403.10131.
20. Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; Chen, W. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv* **2023**, arXiv:2305.15294.
21. Huang, J.; Ping, W.; Xu, P.; Shoeybi, M.; Chang, K.C.-C.; Catanzaro, B. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv* **2023**, arXiv:2308.07922.
22. Yu, W.; Zhang, H.; Pan, X.; Ma, K.; Wang, H.; Yu, D. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv* **2023**, arXiv:2311.09210.
23. Jeong, S.; Baek, J.; Cho, S.; Hwang, S.J.; Park, J.C. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv* **2024**, arXiv:2403.14403.
24. Rajpurkar, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv* **2016**, arXiv:1606.05250.
25. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 453–466. [CrossRef]
26. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1601–1611.
27. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. ♫ MuSiQue: Multihop Questions via Single-hop Question Composition. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 539–554. [CrossRef]
28. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2369–2380.
29. Ho, X.; Nguyen, A.-K.D.; Sugawara, S.; Aizawa, A. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. *arXiv* **2020**, arXiv:2011.01060.
30. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.

31. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. 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, Toronto, ON, Canada, 9–14 July 2023; pp. 9802–9822.
32. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv* **2023**, arXiv:2310.11511.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.