

# Hierarchical Retrieval-Augmented Generation Model with Rethink for Multi-hop Question Answering

Xiaoming Zhang\*  
Northeastern University  
Shenyang, China  
zxm2282588541@gmail.com

Daling Wang†  
Northeastern University  
Shenyang, China  
wangdaling@cse.neu.edu.cn

Ming Wang\*  
Northeastern University  
Shenyang, China  
sci.m.wang@gmail.com

Shi Feng  
Northeastern University  
Shenyang, China  
fengshi@cse.neu.edu.cng

Xiaocui Yang\*  
Northeastern University  
Shenyang, China  
neu.yangxiaocui@gmail.com

Yifei Zhang  
Northeastern University  
Shenyang, China  
zhangyifei@cse.neu.edu.cn

## Abstract

Multi-hop Question Answering (QA) necessitates complex reasoning by integrating multiple pieces of information to resolve intricate questions. However, existing QA systems encounter challenges such as outdated information, context window length limitations, and an accuracy-quantity trade-off. To address these issues, we propose a novel framework, the Hierarchical Retrieval-Augmented Generation Model with Rethink (HiRAG), comprising Decomposer, Definer, Retriever, Filter, and Summarizer five key modules. We introduce a new hierarchical retrieval strategy that incorporates both sparse retrieval at the document level and dense retrieval at the chunk level, effectively integrating their strengths. Additionally, we propose a single-candidate retrieval method to mitigate the limitations of multi-candidate retrieval. We also construct two new corpora, Indexed Wikicorpus and Profile Wikicorpus, to address the issues of outdated and insufficient knowledge. Our experimental results on four datasets demonstrate that HiRAG outperforms state-of-the-art models across most metrics, and our Indexed Wikicorpus is effective. The code for HiRAG is available at <https://github.com/2282588541a/HiRAG>.

## CCS Concepts

- Information systems → Combination, fusion and federated search.

## Keywords

Large Language Models, Hierarchical Retrieval-Augmented Generation, Indexed Wikicorpus Corpus, Profile Corpus

\*These authors contributed equally to this research.

†Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference acronym, 2024,

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM  
<https://doi.org/XXXXXXX.XXXXXXX>

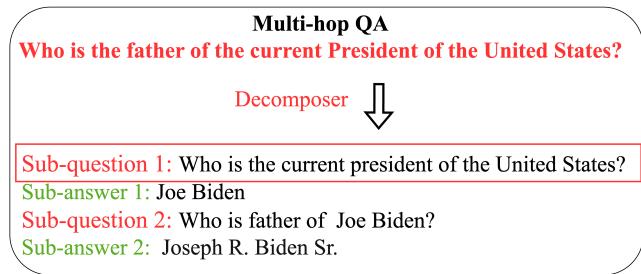
## ACM Reference Format:

Xiaoming Zhang, Ming Wang, Xiaocui Yang, Daling Wang, Shi Feng, and Yifei Zhang. 2018. Hierarchical Retrieval-Augmented Generation Model with Rethink for Multi-hop Question Answering. In *Proceedings of (Conference acronym)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/XXXXXXX.XXXXXXX>

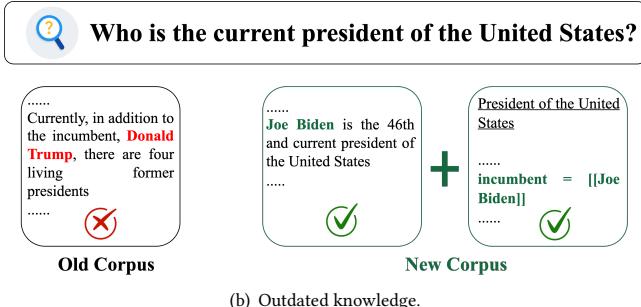
## 1 INTRODUCTION

Multi-hop Question Answering (QA) involves complex reasoning by integrating multiple pieces of information to resolve intricate questions [6, 13, 28]. Unlike single-hop QA, where answers are readily available, complex questions require decomposing the original query into a series of targeted sub-questions. The knowledge required to answer each sub-question varies, drawing from both internal knowledge encoded in Large Language Models (LLMs) and external knowledge retrieved from local knowledge bases, such as Wikipedia [9], or open search engines like Google [26, 27]. As internal knowledge is derived from large-scale data and extensive pre-training, making it challenging to modify, we focus primarily on updating, mining, and effectively leveraging the retrieved knowledge to enhance QA performance.

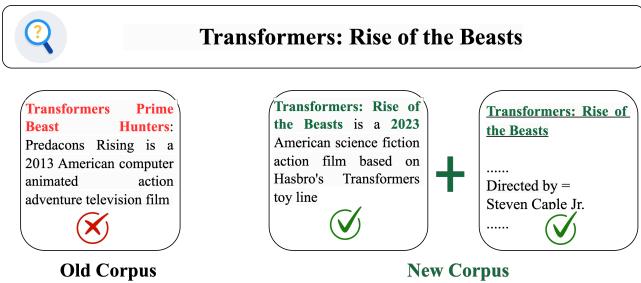
The selection and integration of knowledge retrieved from multiple sources pose various challenges. Several key issues, such as outdated information, context window length limitations, and accuracy-quantity trade-off issues, significantly impact the performance of multi-hop QA systems. Firstly, new multi-hop QA systems often rely on outdated and insufficient knowledge within localized knowledge bases. As illustrated in Figure 1(b), a QA system searches an outdated corpus [9] to answer the query *Who is the current president of the United States?*, resulting in the incorrect answer ‘Donald Trump’. Similarly, the outdated corpus cannot provide real-time updates for newly released news, movies, books, etc. For instance, it does not include the 2023 movie *Transformers: Rise of the Beasts*, as shown in Figure 1(c). Although outdated information can be mitigated through online retrieval, excessive retrieval introduces efficiency and cost issues. To address this, we first update the older corpus and construct an entity-centric, more comprehensive corpus, **Indexed Wikicorpus**, based on Wikipedia, to reduce the reliance on over-searching. Additionally, to quickly understand the basic information of entities, we create a corpus containing some basic information of entities, **Profile Wikicorpus**.



(a) Multi-hop QA.



(b) Outdated knowledge.



(c) Insufficient knowledge.

**Figure 1: Outdated knowledge and insufficient knowledge existing in the old corpus.**

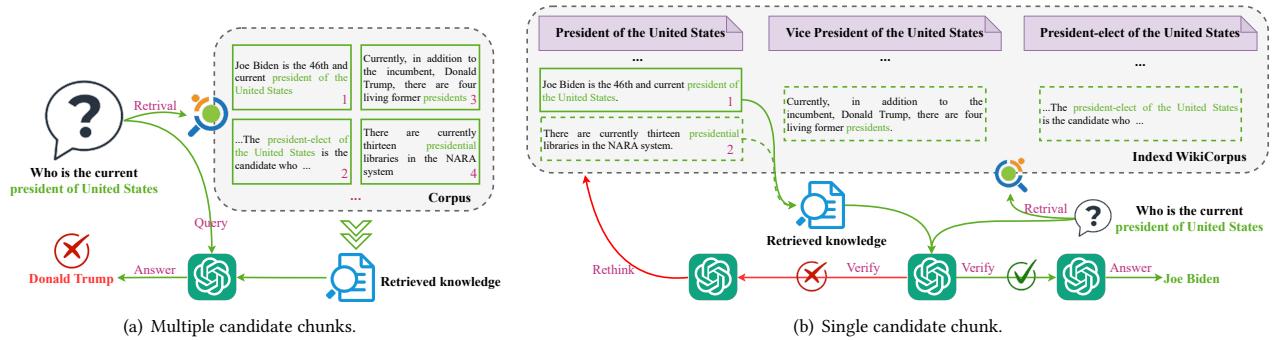
Building on previous studies [15, 19, 29], we decompose the original question into multi-hop questions, as illustrated in Figure 1(a). Each sub-question is answered based on retrieved knowledge, and the answers are then integrated using the Chain-of-Thought (CoT) approach to derive the final answer. In local retrieval, most studies, such as FLARE [8] and MetaRAG [34], employ a retrieval method that involves searching the corpus in chunks and returning the top  $n$  most similar candidate chunks as retrieved knowledge. We refer to this method as *multi-candidate retrieval*, as depicted in Figure 2(a). Intuitively, increasing the value of  $n$  acquires more knowledge, thereby improving the probability of obtaining the correct answer to the query. However, these approaches face two significant challenges. Firstly, the **context window limit** imposes an upper limit on the value of  $n$ , preventing the unlimited expansion of the retrieved content due to the model's context window size constraints. Secondly, the **accuracy-number tradeoff** arises, where increasing  $n$  also increases the amount of potentially irrelevant or redundant

information in the retrieved content. This noisy data may not only mislead the LLM and exacerbate its hallucination problem but also distract its attention, causing it to miss relevant information.

To address the challenges, we propose a novel framework called the **Hierarchical Retrieval-Augmented Generation Model with Rethink (HiRAG)**, as illustrated in Figure 4. HiRAG comprises five key modules, i.e., Decomposer, Definer, Retriever, Filter, and Summarizer. The Filter module further incorporates two submodules, including Verify and Rethink. Previous methods typically rely on either sparse retrieval [18], which focuses on lexical matching, or dense retrieval [31], which leverages latent embeddings for semantic matching. While some approaches combine the results of these two retrieval methods at the output level [1, 9, 11], they do not deeply integrate the strengths of each retrieval method within the retrieval process itself. We propose the Retrieval module which incorporates a new hierarchical retrieval strategy, performing multi-level retrieval from the document level to the chunk level. Initially, we employ sparse retrieval at the document level to extract key information, such as entity names, through lexical matching. Subsequently, dense retrieval is used to retrieve specific information at the chunk level, leveraging semantic matching. We also introduce a novel retrieval method, **single-candidate retrieval**, as shown in Figure 2(b). This method returns only the most similar chunk as the retrieved knowledge for each decomposed sub-question. However, selecting just one candidate does not guarantee the correctness of the answer. To mitigate this, we apply the Filter module to assess the answer based on the retrieved knowledge. If the answer to a sub-question is found to be incorrect, we utilize the Rethink module, in conjunction with the Retrieval module, to select another chunk as the retrieved knowledge to answer the question. The rethinking process is iterated until the Filter module confirms the answer as correct, thereby yielding the solution to the current sub-question. This cycle of decomposition, retrieval, and answering is repeated for subsequent sub-questions until the Definer module determines that the current question, accompanied by its sub-answers, is answerable. At this point, the Summarizer module is invoked to generate the final answer to the current question.

We conduct experiments on four datasets, including HotPotQA [28], 2WikiMultihopQA [6], MuSiQue [20], and Bamboogle [15], for multi-hop question answering tasks. The experimental results demonstrate that HiRAG outperforms state-of-the-art models across all metrics on three datasets while showing advancements in several metrics on the remaining dataset, and our Indexed Wikicorpus is effective. Our main contributions are summarized as follows:

- To address the limitations of outdated and insufficient knowledge in existing corpora, we construct the Indexed Wikicorpus, which is organized by entity names. Subsequent experiments demonstrate that this corpus is more comprehensive and logically structured. Additionally, we propose the Profile Wikicorpus, which extracts auxiliary information on key entities to further provide effective knowledge.
- We propose a novel retrieval-augmented generation framework, **Hierarchical Retrieval-Augmented Generation Model with Rethink (HiRAG)**, which not only introduces a hierarchical retrieval method to integrate the advantages of different retrieval technologies but also uses single-candidate



**Figure 2: Multiple candidate chunks vs. single candidate chunk. The knowledge within the solid box represents the retrieved knowledge. In subfigure 2(b), the knowledge within the dotted box will be considered as newly retrieved knowledge only if the answer to the sub-question is incorrect.**

retrieval to solve the problems currently encountered in multi-candidate retrieval.

- We conduct comprehensive experiments on four datasets and verify the effectiveness of different components in HiRAG. The experimental results demonstrate that HiRAG outperforms state-of-the-art models on most metrics and confirms the effectiveness of the Indexed Wikicorpus.

## 2 RELATED WORK

Multi-hop QA is the task of answering natural language questions that involve extracting and combining multiple pieces of information and doing multiple steps of reasoning [12]. It can be divided into two steps, a retrieval (IR) step to extract all relevant context from the corpus, and a reading comprehension (MRC) step to find the answer from the reading result context. In the era of LLM, research [5, 14] finds that on the one hand, LLM can be used to complete the decomposition of the problem and achieve more fine-grained retrieval in the IR step, and on the other hand, Retrieval-Augmented Generation technology can be used to achieve the integration of retrieval information in the MRC step.

### 2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a current research hotspot for Multi-hop QA tasks [10]. It can integrate the internal knowledge of the model with the external knowledge retrieved. LLM can retrieve external content through RAG to expand their knowledge base, thereby improving their ability to generate accurate and contextually relevant responses. Historically, various studies attempt to adapt the use of generative models to improve their performance. For instance, REPLUG [17] uses different retrieved content to generate corresponding answers and then combine them. Self-Rag [2] fine-tunes a generation model to simultaneously produce answers along with relevance, support, and usefulness scores. Concurrently, several methods for multi-hop QA emphasize the content and timing of retrieval. Self-Ask [15] lets the model generate sub-questions and queries, and continuously alternate between retrieval and generation. PROMPTAGATOR [4], Take a step back [32] focus on abstracting high-level concepts and utilizing LLMs for

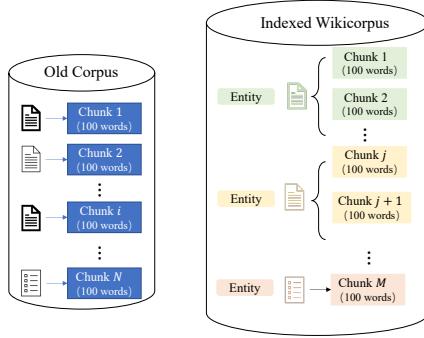
**Table 1: Old Corpus vs. Indexed Wikicorpus.**

Corpus	Number of entities	Number of words
Old corpus (from DPR)	3232908	2101532400
Indexed Wikicorpus	6416724	2642615682

prompt-based query generation. Additionally, the confidence-based method, FLARE [8], generates queries using low-confidence tokens. However, most studies directly feed the retrieval content into the generation model, ignoring the evaluation and processing of the retrieval content. Unlike them, HiRAG highlights the importance of verifying retrieved content and adjusts the retriever to enhance the relevance of results when the quality of the retrieved information is subpar.

### 2.2 Chain-of-Thought (CoT)

In the task of Multi-hop QA in addition to using retrieval-enhanced generation to obtain knowledge, CoT is also needed to improve the logic. CoT can significantly enhance the reasoning capabilities of models [3]. For instance, iterative Context-Aware Prompter [21] employs an iterative approach to knowledge acquisition from the model to accomplish reasoning tasks. Similarly, LEAST-TO-MOST [33] decomposes complex problems into a series of sub-problems, addressing each step methodically. Inspired by the concept of self-consistency [23], MCR [30] extends the application of self-consistency beyond final results to include intermediate steps, thereby enhancing the overall accuracy of reasoning. Following [33], we design a comprehensive prompt to let the model decompose the question into multiple sub-questions and finally integrate the multiple sub-answers into the final answer through CoT. The main difference between our work and previous work is that we do not decompose the problem all at once, but proceed in a loop, generating only one sub-problem in each round; at the same time, we design a matching Definer to realize the judgment of whether the problem can be solved and exit the loop.



**Figure 3: Old corpus vs. Indexed Wikicorpus.** In the new corpus, we specialize in constructing a document for each entity and then dividing it into chunks.

### 3 INDEXED WIKICORPUS AND PROFILE WIKICORPUS

The importance of corpus cannot be overstated in the realm of external knowledge acquisition. The Wikipedia corpus released by DPR [9] is widely recognized as a standard resource. However, as time progresses, this corpus faces challenges such as knowledge gaps and outdated information. Moreover, to accommodate dense retrieval requirements, the corpus is designed to fragment entity information into multiple segments, often resulting in incoherent data representation. To address these limitations, we develop a novel Wikipedia corpus called **Indexed Wikicorpus** along with a corresponding key entity profile corpus called **Profile Wikicorpus**. Our approach differs significantly from its predecessor in prioritizing the coherence of entity information. In Indexed Wikicorpus, each entry represents a complete entity as shown in Figure 3, ensuring a more comprehensive and cohesive representation of information. Building upon this foundation, we draw inspiration from the Web version of Wikipedia to extract entity profiles. These profiles, curated by Wikipedia, encapsulate the essential information about each entity, which is saved to Profile Wikicorpus. Data analysis about the number of words and entities as shown in Table 1 reveals that our new corpus contains a higher volume of entity information compared to its predecessor. Furthermore, subsequent experiments demonstrate the superior performance of our corpus over the older version.

### 4 HIERARCHICAL RETRIEVAL-AUGMENTED MODEL

To address the challenges in retrieval-augmented generation, including the lack of deep integration between different retrieval methods and the potential introduction of noise from multi-candidate retrieval, we propose a novel framework called **HiRAG**. This framework consists of five key components: Decomposer, Summarizer, Retriever, Definer, and Filter, as illustrated in Figure 4. The detailed algorithm is presented in Algorithm 1 in Appendix A. Specifically, the Decomposer is designed to tackle complex questions by decomposing them into smaller, more manageable sub-questions that can be easily answered. The Definer then determines whether the question can be solved. If it can, the Summarizer leverages the

sub-answers to generate a response to the original question. Otherwise, the Retriever extracts relevant information related to the sub-question through a hierarchical retrieval process at both the document and chunk levels. Finally, the Filter verifies the validity of the retrieved content, generates a sub-answer, and re-evaluates the results if they are found to be inaccurate.

#### 4.1 Notations and Definitions

For a multi-hop question,  $x$ , we can decompose it into a series of sub-questions  $Q$ , where answers to previous sub-questions can inform the generation of subsequent sub-questions. The objective is to iteratively obtain the set of sub-questions  $Q = \{q_i\}_{i=1}^{|Q|}$  and their corresponding set of sub-answers  $\mathcal{A} = \{a_i\}_{i=1}^{|Q|}$ , where  $q_i$  represents the  $i$ -th decomposed sub-question and  $a_i$  its corresponding sub-answer. Ultimately, we combine all sub-answers with the original question,  $x$ , to get the final answer  $o$ . To augment the ability of the model to answer questions, we leverage external knowledge from the Wikipedia corpus, denoted as  $\mathcal{D}$ , which comprises numerous sub-documents. Each sub-document,  $d$ , contains relevant content, including a title,  $t$ , and body text. For a given sub-question, we first employ a retriever to identify the most relevant sub-document,  $d = \text{Retrieval}(q, \mathcal{D})$ . We then pinpoint the most relevant chunk of text,  $c$ , in  $d$ . Finally, we utilize the most relevant chunk to answer the sub-question using a LLM, yielding the sub-answer  $a = \text{LLM}(q, c)$ .

#### 4.2 Decomposer and Summarizer

Building upon the previous work [33], we use the Decomposer module to break down complex problems into manageable sub-questions leveraging the prompt engineering [22]. We develop a comprehensive prompt for the Decomposer, which includes its background, goals, constraints, workflow, examples, and initialization. The prompt of the Decomposer module is formulated as follows:

Prompt of Decomposer:

**Background:**

- You are an expert at analyzing problems...

**Goal:**

Helping the user decompose the question and tell the user at the right time that the problem can be solved.

**Constraint:**

- You can only decompose the question, do not answer it directly...

**Workflow:**

1. Analyse the original complex question...

**Example:**

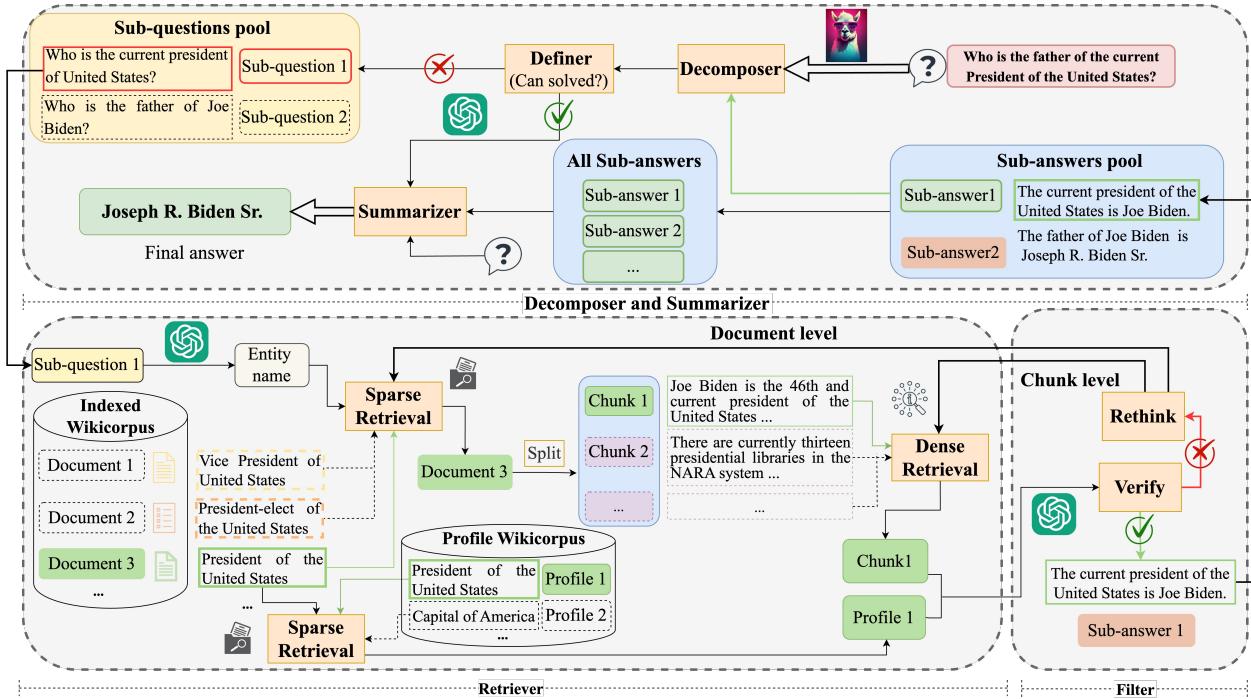
....

**Initialization:**

Now, a first simple question.

We initialize the model turn,  $mt$ , to 0 which indicates the round of the current decomposition problem. For the  $mt$ -th iteration, the Decomposer makes a new sub-question with the previous sub-answers,  $\{a_1, a_2, \dots, a_{mt-1}\}$ , and original question,  $x$ . Then the **Definer** determines whether the original question can be solved with all known sub-answers based on the output of **Decomposer**. Therefore, this process can be formulated as follows:

$$\begin{aligned} q_{mt} &= \text{Decomposer}(\{a_1, a_2, \dots, a_{mt-1}\}, x), \\ \text{Judge} &= \text{Definer}(q_{mt}), \\ mt &= mt + 1. \end{aligned} \tag{1}$$



**Figure 4: The framework of the Hierarchical Retrieval-Augmented Generation Model with Rethink (HiRAG), including Decomposer module, Definer module, Retriever module, Filter module, and Summarizer module. The process begins with the Decomposer module, which breaks down the original question into several sub-questions. Each sub-question is then forwarded to the Retriever module to retrieve pertinent knowledge. For illustration, we consider the handling of the first sub-question as an example. The knowledge obtained is passed to the Verify module, where each sub-answer is evaluated for accuracy. If a sub-answer is verified as correct, it is stored in the sub-answer pool. Should it be incorrect, a rethinking process is initiated. Subsequent sub-questions undergo a similar sequence of decomposition, retrieval, and verification. Ultimately, all sub-answers, along with the original question, are compiled by the Summarizer module to output the final answer. Solid lines denote items that are currently being processed, while dotted lines indicate items that are pending processing or are about to be addressed. The llama icon represents LLaMa-3-70B and the ChatGPT icon is GPT-3.5-turbo in this paper.**

If the *Judge* returns a result indicating that the original question can be solved, we send the original question,  $x$ , with all sub-answers before the current decomposition round to the Summarizer. The *Summarizer* tries to generate the output  $o$  which can be formulated as follows:

$$o = \text{Summarizer}(\{a_1, a_2 \dots, a_{mt}\}, x). \quad (2)$$

In Appendix A.1, we present experimental results investigating the relationship between model performance, as measured by the EM metric, and the number of model turns  $mt$ . The results are summarized in Table 7.

### 4.3 Hierarchical Retriever

In this section, we describe the design of a novel hierarchical retrieval mechanism that enhances retrieval accuracy and provides semantic clarification in cases of ambiguous semantics. Our approach employs a two-layer hierarchical retrieval process. The first layer identifies the most relevant documents within the corpus at the document level, while the second layer locates the most relevant chunks within those documents at the chunk level, resulting in a

more accurate and refined retrieval process. We begin by obtaining a decomposed question,  $q$ , through the Decomposer module. Next, we utilize the LLM to extract the entity name,  $e$ , from the question. We then leverage this entity name to identify the document with the title,  $t_c$ , that most closely matches the entity name among all available documents in the corpus. As this step primarily involves matching entity names with indexes, lexical matching takes precedence. Therefore, we employ sparse retrieval [18] to implement this step, capitalizing on its strengths in lexical matching.

After we get the document with the title  $t_c$ , we look for the corresponding information,  $p^*$ , in the Profile WikiCorpus and add it to the retrieval result if it exists.

$$\begin{aligned} e &= \text{LLM}(q), \\ t_c &= \text{SparseRetrieval}(e, \{t_1, t_2 \dots, t_n\}), \\ p^* &= \text{SparseRetrieval}(t_c, \{p_1, p_2 \dots, p_m\}). \end{aligned} \quad (3)$$

where  $n$  represents the number of documents in the Indexd WikiCorpus and  $m$  represents the number of profiles in the Profile WikiCorpus.

During the process of question decomposition and entity extraction, a notable challenge emerges. The use of LLMs for both sub-question generation and entity information extraction can lead to information loss. Specifically, when semantic loss occurs, it is often characterized by a high similarity between the extracted entity name and multiple candidate titles. To mitigate this issue, the retriever enhances the semantics by incorporating contextual information. In the context of Multi-hop QA tasks, this semantic enrichment draws upon two primary sources: the original question and the preceding answers. By leveraging this contextual information, we first select the most relevant details and then generate a new question that better captures the intended meaning. Specifically, we first select an answer,  $a^*$ , containing the entity,  $e$ , from the set of sub-answers,  $\mathcal{A}' = \{a_1, a_2 \dots, a_{mt-1}\}$ , which includes all sub-answers from previous  $mt - 1$  rounds and utilize it as a supplement to our query. If this approach yields no results, we instead use the original question  $x$  as the supplement to guide our investigation. We then instruct the **LLM** to generate a new question  $q^*$ , which is subsequently retrieved using other retrieval engines.

$$\begin{aligned} a^* &= \text{Select}(\{a_1, a_2 \dots, a_{mt-1}\}, e), \\ q^* &= \text{LLM}(q, s). \end{aligned} \quad (4)$$

where **Select** is the process of finding the contextual information.

Upon obtaining the relevant sub-document  $d_c$  associated with the title  $t_c$ , we proceed to split it into uniform chunks,  $C = \{c_1, c_2, \dots, c_n\}$ . In this phase, we prioritize semantic matching and employ a dense retriever, Contriever [7], to facilitate the process. The similarity score is calculated by computing the dot product of the vector representations of the question and each chunk. The chunk with the highest similarity score is then selected as the most relevant chunk, denoted as,  $c_s$ .

$$\begin{aligned} \{c_1, c_2 \dots, c_n\} &= \text{Spilt}(d_c), \\ \text{Sim}(q, c_i) &= \langle E(q), E(c_i) \rangle, \quad \forall 1 \leq i \leq n, \\ c_s &= \arg \max_{c_i \in C} \text{Sim}(q, c_i), \quad \forall 1 \leq i \leq n, \end{aligned} \quad (5)$$

where **Spilt** is the process of dividing the documents into many chunks and  $n$  is the number of chunks.

#### 4.4 Filter

Even the most advanced retrieval engines can struggle to consistently retrieve the most accurate content. To mitigate the impact of potential retrieval engine errors, we utilize a specialized filter. This filter is designed to evaluate retrieval results and refine the retrieval engine through a series of iterative adjustments when inaccuracies are detected. Upon receiving the results from the retrieval engine  $c_s$ ,  $p^*$ , and the sub-question,  $q$ , generated by the Decomposer, we first attempt to leverage a **LLM** to generate a response,  $r$ . The model then assesses whether the current sub-question,  $q$  can be resolved based on the response,  $r$ .

$$r = \text{LLM}(q, c_s, p^*). \quad (6)$$

If the question can be solved, the Filter passes  $r$  as the answer,  $a$ , to the sub-question to the Decomposer. However, when faced with an unresolvable question, the Filter triggers a **two-tiered rethinking**

**retrieval process**, encompassing both chunk-level retrieval and document-level retrieval to facilitate a more comprehensive search.

Initially, if the Filter determines that the problem cannot be resolved, it initiates a rethink at the chunk level. During this phase, the result,  $d_c$ , from the first-tier retrieval by the retriever remains unchanged. Instead, modifications are made to the second tier of the retriever, sequentially selecting chunks with higher similarity scores from the divided chunks. This approach is particularly effective when the correct entity information and corresponding references are identified, but the specific chunk required to resolve the problem is not found. If the chunk-level rethink proves unsuccessful, the process escalates to the document level. At this level, adjustments are made to the first level of the retriever by selecting titles with higher similarity to the entity,  $e$ , from the candidate titles. Notably, this two-tiered rethink process operates as a nested loop, where each higher-level rethink informs and drives the lower-level rethinks, ensuring a comprehensive and systematic approach to problem-solving. Concurrently, we also address the issue of knowledge balance during the retrieval process. The rapid advancement of LLMs has led to a significant increase in the quantity and quality of their internal knowledge. As a result, a critical challenge in RAG emerges: striking a balance between the external knowledge retrieved and the internal knowledge of the model. While previous research [25] has focused on developing classifiers to tackle this challenge, we propose a novel and straightforward approach. By passing the number of filter rethinks as a parameter to the classifier, we observe that as the number of rethinks increases, the semantic similarity between the model and the retrieved content generally decreases.

In the context of the sub-question,  $q$ , if the retrieved result after the  $t$ -th round of rethink fails to yield an answer, we propose incorporating a probability  $y$  of leveraging the internal knowledge of the model to address the question.

$$y = \left( \frac{t}{m} \right)^2. \quad (7)$$

where  $m$  is a hyperparameter and  $t$  is the round of rethink. In our experiment, the maximum number of retrievals is set to 4, where the value of  $m$  we employ is 5. We recommend that  $m$  be greater than or equal to the maximum number of retrievals. This choice is the result of a trade-off. On one hand, the model should not abandon retrieval and resort to its internal knowledge too hastily, as the reliability of internal knowledge cannot be guaranteed. On the other hand, it should also ensure that when external knowledge is suboptimal and the retrieval method fails, it can attempt to utilize uncertain internal knowledge to provide an answer.

## 5 EXPERIMENTS SETUP

### 5.1 Datasets

In our investigation, we conduct experiments on four datasets specifically designed for the Multi-hop question answering task, namely **HotPotQA** [28], **2WikiMultiHopQA** [6], **MuSiQue** [20], and **Bamboogle** [15]. We only use questions and answers in the dataset, with all external knowledge retrieved.

**Table 2: Experimental results on four datasets. Without retrieval means only using the internal knowledge of LLM while With retrieval means using the external knowledge from the retriever. HiRAG is divided into HiRAG (online) and HiRAG (local) according to whether it uses Google for retrieval after semantic supplementation. The best results are in bold.**

Settings	Models	HotpotQA				2WikiMultihopQA				MuSiQue				Bamboogle			
		EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall
W/O retrieval	Direct	24.00	31.18	35.51	29.58	24.20	29.21	31.72	28.48	2.20	7.15	9.87	6.20	15.20	18.53	19.47	18.00
	CoT	31.80	43.16	44.14	45.34	26.80	35.26	34.14	39.32	8.20	19.07	19.91	20.84	52.80	64.97	65.32	67.47
	CoT-SC	33.40	44.95	45.61	47.65	29.00	36.62	35.62	40.21	10.00	20.33	21.10	21.89	52.80	63.55	63.83	64.30
	Self-Ask w/o Ret	26.40	37.82	38.54	41.07	28.80	37.50	36.77	40.50	9.20	19.29	19.36	22.22	51.20	59.05	58.92	60.80
W/ retrieval	HiRAG w/o Ret	36.40	47.46	49.39	48.07	37.20	46.24	45.08	49.23	13.20	25.22	26.82	25.78	<b>60.80</b>	69.95	69.66	<b>71.57</b>
	Direct	24.40	35.00	41.48	33.00	29.40	38.50	37.57	41.90	5.20	10.30	13.91	9.14	19.20	27.23	31.60	25.53
	ReAct	25.20	32.68	33.85	33.17	25.20	30.98	30.81	32.25	5.80	8.17	8.11	8.67	20.00	24.09	25.63	23.63
	Self-Ask	25.60	36.04	37.15	38.96	35.00	46.42	45.26	50.99	6.20	12.95	12.58	15.51	40.80	49.62	49.34	52.43
	Flare	41.60	54.37	56.32	55.54	40.60	52.05	50.34	57.07	13.39	<b>26.22</b>	<b>27.72</b>	27.51	46.34	57.71	57.50	58.27
	MetaRAG	36.80	48.43	50.41	49.16	23.15	30.38	29.79	32.52	8.40	17.23	18.64	18.25	23.20	30.36	31.20	30.33
	HiRAG (online)	40.48	52.51	53.98	53.70	50.38	<b>65.30</b>	<b>63.22</b>	63.22	<b>14.65</b>	26.03	26.58	<b>28.15</b>	60.00	<b>70.34</b>	<b>70.68</b>	71.50
	HiRAG (local)	42.52	<b>54.98</b>	<b>57.16</b>	<b>57.16</b>	<b>52.29</b>	63.95	61.99	<b>70.27</b>	11.11	21.99	21.89	25.72	53.17	64.79	65.81	66.27

**Table 3: Result of substitution of different LLMs as the backbone, i.e. the icon of ChatGPT in Figure 4. The best results are in bold.**

Online or local	LLMs	HotpotQA				2WikiMultihopQA				MuSiQue				Bamgoole			
		EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall
Local	Qwen2-7B	41.50	51.65	52.97	53.53	52.50	63.64	61.29	70.12	9.80	21.87	21.32	25.16	41.60	52.87	54.92	53.57
	LLaMa-3-8B	37.84	47.06	49.13	47.20	43.60	54.33	52.64	58.87	6.80	16.45	16.81	17.76	42.86	51.74	51.29	54.05
	LLaMa-3-70B	<b>44.50</b>	<b>56.10</b>	<b>58.15</b>	56.52	57.00	68.50	<b>66.69</b>	74.24	16.60	<b>26.27</b>	<b>28.18</b>	48.80	62.79	63.19	63.90	
	HiRAG (GPT-3.5-turbo)	42.52	54.98	57.16	<b>57.16</b>	52.29	63.95	61.99	70.27	11.11	21.99	21.89	25.72	53.17	64.79	65.81	66.27
Online	Qwen2-7B	32.56	41.51	40.89	44.45	45.59	58.45	55.53	68.01	9.46	20.57	20.72	22.29	36.54	50.25	52.47	52.56
	LLaMa-3-8B	37.66	46.28	48.20	46.86	43.68	54.28	52.43	59.74	7.03	14.86	15.48	15.77	40.00	49.95	50.72	50.83
	LLaMa3-70-B	41.12	52.78	55.64	52.88	<b>57.68</b>	<b>70.61</b>	68.48	<b>77.04</b>	<b>15.37</b>	25.50	26.24	26.78	54.76	66.54	66.94	67.33
	HiRAG (GPT-3.5-turbo)	40.48	52.51	53.98	53.70	50.38	65.30	63.22	63.22	14.65	26.03	26.58	28.15	<b>60.00</b>	<b>70.34</b>	<b>70.68</b>	71.50

## 5.2 Baselines

Our baselines can be divided into two categories according to whether external knowledge is retrieved: Without retrieval baselines and With retrieval baselines. The Without retrieval (**W/o retrieval**) setting includes four baselines. **Direct** Prompting directly uses the question as a prompt for LLM to respond. For retrieval, directly use the question as the query to search. **CoT** Prompting [24] adds “Let’s think this question step by step” after the initial question. **CoT-SC** [23] consolidates its responses by answering the same question multiple times. **Self-Ask** [15] enables the model to decompose the problem into a series of sub-questions, allowing it to answer or retrieve relevant information for sub-questions. Direct and Self-Ask models belong to two working modes, so the With retrieval (**W/ retrieval**) setting includes five baselines. **ReAct** [29] is similar to self-ask, the difference is that react will extract the query to be retrieved from the sub-questions after breaking down the problem, while Self-ask will directly use the sub-questions as the query. **Flare** [8] uses a confidence-based strategy when generating questions. It generates questions about the less-confident parts of the generated content and retrieves them. **MetaRAG** [34] gains insights from metacognition, which can identify logical errors in reasoning and use the three-step metacognitive regulation pipeline to identify and repair deficiencies in initial cognitive responses.

## 5.3 Implementation Details

We use GPT-3.5-turbo as the backend LLM for the most part while using LLaMa-3-70B for the Decomposer. For baselines, we use the family of GPT-3.5 as the backend LLM. For online retrieval, we use Wikipedia and Google through the Wikipedia API and Serper API.

For local retrieval, we use Contriever-MSMARCO [7] and elastic with the BM25 algorithm [16] as dense retriever and sparse retriever. The default maximum number of retrievals is 4. For HiRAG, we use Indexed Wikicorpus and Profile Wikicorpus for retrieval while we use the DPR corpus [9] for retrieval in baselines. Following [34], we employ four evaluation metrics to assess the performance of our model. At the answer level, we use Exact Match (EM) to determine whether the predicted answer aligns perfectly with the reference answer. Additionally, at the token level, we adopt a more fine-grained approach, which includes token-level F1 score, precision, and recall. Following MetaRAG [34], for cost considerations, we sub-sample 500 questions from the validation set of HotPotQA, 2WikiMultiHopQA, and MuSiQue for experiments while using the full set of Bamboogle for experiments as the quantity of Bamboogle is small.

## 6 RESULTS AND ANALYSIS

### 6.1 Main Results

The main results are shown in Table 2 and reveal the following notable findings. (1) Our proposed framework, HiRAG, exhibits superior performance across multiple evaluation metrics, outperforming state-of-the-art methods on three out of four datasets. Additionally, it demonstrates improvements in several metrics on the remaining dataset. The key advantage of our approach is its emphasis on the retrieval process, which consistently produces high-quality results. This highlights the crucial role of the retrieval component in achieving exceptional outcomes. (2) In the Without Retrieval setting, HiRAG significantly outperforms the baselines, showcasing its effectiveness in question decomposition and answer

**Table 4: Ablation experimental results for the Retriever module.**

Online or local	HotpotQA				2WikiMultihopQA				MuSiQue				Bamgoole			
	EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall
HiRAG	42.52	54.98	57.16	57.16	52.29	63.95	61.99	70.27	11.11	21.99	21.89	25.72	53.17	64.79	65.81	66.27
HiRAG (w/o chunk)	41.85	53.01	55.16	54.17	49.04	62.69	61.00	67.85	6.02	15.92	15.23	18.95	38.40	58.58	57.46	62.50
HiRAG (w/o document)	39.67	52.30	54.70	52.92	48.68	61.52	59.63	67.17	9.45	18.37	21.85	20.37	48.36	58.72	59.56	59.70
HiRAG (w/o chunk + document)	39.57	50.36	50.52	51.04	47.50	53.87	52.62	56.70	5.51	13.97	13.95	15.35	42.53	53.37	54.79	54.02
HiRAG (w/o Profile WikiCorpus)	40.98	52.84	54.51	53.96	50.67	65.04	63.09	71.09	10.81	21.54	20.81	25.68	51.61	63.39	65.59	64.78

**Table 5: Corpus experiments based on Flare, such as the corpus generated by DPR and Indexed Wikicorpus.**

Corpus	HotpotQA				2WikiMultihopQA				MuSiQue				Bamboogle			
	EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall	EM	F1	Precision	Recall
Old corpus (DPR)	41.60	54.37	56.32	55.54	40.60	52.05	50.34	57.07	13.39	26.22	27.72	27.51	46.34	57.71	57.50	58.27
Indexed Wikicorpus	<b>43.00</b>	<b>55.06</b>	<b>56.95</b>	<b>55.65</b>	<b>50.00</b>	<b>62.27</b>	<b>60.30</b>	<b>68.15</b>	<b>13.60</b>	<b>25.62</b>	<b>26.55</b>	<b>26.58</b>	<b>49.60</b>	<b>60.20</b>	<b>60.28</b>	<b>61.13</b>

summarization. This is due to the framework’s design, which enables autonomous sub-question generation and termination of the loop when the original problem is resolvable. Our approach also differs from self-ask in its segregation of subtasks and implementation of each subtask through a separate prompt, resulting in superior performance. (3) A comparative analysis of results across datasets reveals that HiRAG achieves the most significant breakthrough in 2WikiMultiHopQA, with a notable improvement of over 12% in the EM index compared to the state-of-the-art method. This is primarily attributed to the fact that most external knowledge required by 2WikiMultiHopQA can be retrieved from Wikipedia, and the question format in this dataset is relatively standardized. However, the MuSiQue dataset poses a greater challenge due to the complexity of the questions and the inability to directly retrieve required knowledge from Wikipedia. This complexity affects our framework’s ability to evaluate retrieval results, diminishing the effectiveness of the response.

## 6.2 Generalization

To assess the generalizability of HiRAG, we conduct a series of experiments where we substitute different base models, with the results presented in Table 3. Specifically, we evaluate the performance of HiRAG when paired with several prominent base models, including LLaMA-3-70B, LLaMA-3-8B, and Qwen2-7B. The results demonstrate that our method exhibits robustness and effectiveness across different base models, showcasing its adaptability to various model architectures. When paired with LLaMA-3-70B, our approach achieves state-of-the-art performance on most datasets, underscoring its ability to enhance the performance of strong base models and highlighting its potential for widespread applicability. *Notably, to ensure a fair comparison with existing SOTA models, like Meta-RAG, which is built on GPT-3.5-turbo, we use the same base model in our experiments, even though LLaMA-70B has shown better performance.* This allows for a more direct and meaningful comparison of our approach with prior work.

## 6.3 Ablation Experiments

We conduct ablation experiments to evaluate the contributions of our Retriever module, focusing on the hierarchical retrieval approach. By removing the processing during rethinking at the chunk

level and document level, we investigate the impact of these components on the overall performance. The results, presented in Table 4, demonstrate the effectiveness of our approach. Our method is designed to improve the accuracy of retrieved results through chunk level and document level rethink. To isolate the impact of each level, we perform ablation experiments by eliminating one level of rethink at a time. Specifically, we remove chunk level rethink by considering only the highest-scoring chunk within a document and remove document level rethink by retrieving only the highest-scoring document. The results show that eliminating either level of rethink leads to a decrease in performance, confirming the importance of both chunk level and document level rethink in our hierarchical retrieval approach. Furthermore, we investigate the influence of knowledge incorporated in the Profile WikiCorpus. Our findings indicate that removing this knowledge component leads to a decrease in the performance of HiRAG.

## 6.4 Indexed Wikicorpus

We conduct experiments under the FLARE framework to assess the effectiveness of our corpus in improving model performance. As shown in Table 5, replacing the external knowledge source with our corpus leads to notable advancements in Exact Match (EM) scores across four datasets, with three datasets experiencing improvements in all evaluated metrics. The comprehensive coverage and meticulous entity name segmentation in our corpus are key factors contributing to these improvements, enabling more efficient retrieval and utilization of relevant information.

## 6.5 Retriever Module as Plug-in

We extract retriever and filter modules HiRAG, pairing them with Indexed Wikicorpus and Profile Wikicorpus to create a novel intelligent retriever. This approach leverages LLMs to evaluate and refine retrieval results, differing from traditional retrievers. We conduct experiments on standardized RAG tasks for *single-hop questions using gold decomposed question-answer pairs* from 2WikiMultiHopQA and MuSiQue datasets. We compare the performance of five methods: (1) Direct Answering, (2) Sparse Retrieval with Elastic and BM25, (3) Dense Retrieval with Contriever-MSMARCO, (4) HiRAG (Online), and (5) HiRAG (Local). The results in Table 6 clearly demonstrate the superiority of our retrieval engine over traditional approaches, with significant gains observed even in local retrieval scenarios.

Specifically, we report a maximum improvement of over 9% in EM metric, outperforming current mainstream retrieval engines. Moreover, our engine exhibits robust and comprehensive improvements across all datasets and evaluation metrics, underscoring the effectiveness of our hierarchical retrieval method. The ability of our engine to serve as a plug-in, augmenting the performance of other methods, makes it a valuable tool for achieving state-of-the-art results in a variety of applications.

**Table 6: Comparison of HiRAG as a Plug-in with other retrieval engines.**

Dataset	Retrieval method	EM	F1	Precision	Recall
2WikiMultihopQA	Direct	10.93	16.59	17.04	17.42
	Sparse retrieval	49.28	54.91	55.69	55.99
	Dense retrieval	46.06	54.10	55.85	54.96
	HiRAG (online)	<b>57.46</b>	<b>76.33</b>	<b>75.75</b>	<b>79.66</b>
	HiRAG (local)	55.22	73.77	73.00	77.23
MuSiQue	Direct	9.61	12.29	12.62	12.85
	Sparse retrieval	26.60	33.77	34.92	34.67
	Dense retrieval	29.99	38.95	39.97	40.64
	HiRAG (online)	<b>52.10</b>	<b>73.01</b>	<b>74.09</b>	<b>76.82</b>
	HiRAG (local)	32.13	44.06	42.48	48.14

## 6.6 Case Study

We provide a case study in Figure 5 of Appendix A.3 to demonstrate the workings of our framework. We gradually generate sub-questions, retrieve sub-questions, evaluate the retrieval results, and answer the sub-question. When the retrieval results are not ideal, we use rethink to find a more satisfactory result. Finally, we use the answers to all sub-questions to get the answer to the original question.

## 7 CONCLUSION

We present a novel framework for multi-hop question answering, addressing the challenges of outdated and insufficient knowledge, context window limitations, and accuracy-quantity trade-offs. Our proposed framework, HiRAG, incorporates a hierarchical retrieval strategy, single-candidate retrieval, and a rethink mechanism to improve the efficiency and effectiveness of knowledge retrieval. The experimental results demonstrate that HiRAG outperforms state-of-the-art models on multiple datasets, confirming the effectiveness of our approach. Additionally, our newly constructed corpora, Indexed Wikicorpus which is shown to be more comprehensive and logically organized, and Profile Wikicorpus. Our contributions provide a significant step forward in improving the performance of multi-hop QA tasks, and we believe that our framework and corpora will be valuable resources for future research in this area. In the future, we hope to expand our research on the retrieval-verify-rethink pipeline to achieve more fine-grained and accurate retrieval.

## Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62272092, No.62172086, and No.62106039).

## References

- [1] Negar Arabzadeh, Xinyi Yan, and Charles LA Clarke. 2021. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2862–2866.
- [2] 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* (2023).
- [3] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402* (2023).
- [4] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755* (2022).
- [5] Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106* (2017).
- [6] 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* (2020).
- [7] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* (2021).
- [8] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983* (2023).
- [9] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).
- [10] 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 (2020), 9459–9474.
- [11] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics* 9 (2021), 329–345.
- [12] Vaibhav Mavi, Anubhav Jangra, Jatowt Adam, et al. 2024. Multi-hop Question Answering. *Foundations and Trends® in Information Retrieval* 17, 5 (2024), 457–586.
- [13] Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140* (2022).
- [14] Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. *arXiv preprint arXiv:1906.02900* (2019).
- [15] 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* (2022).
- [16] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.
- [17] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).
- [18] Krysta M Svore and Christopher JC Burges. 2009. A machine learning approach for improved BM25 retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 1811–1814.
- [19] Harsh Trivedi, Nirajan 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* (2022).
- [20] Harsh Trivedi, Nirajan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics* 10 (05 2022), 539–554. [https://doi.org/10.1162/tacl\\_a\\_00475](https://doi.org/10.1162/tacl_a_00475) [https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl\\_a\\_00475/2020694/tacl\\_a\\_00475.pdf](https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00475/2020694/tacl_a_00475.pdf)
- [21] Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively Prompt Pre-trained Language Models for Chain of Thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2714–2730. <https://doi.org/10.18653/v1/2022.emnlp-main.174>
- [22] Ming Wang, Yuanzhong Liu, Xiaoming Zhang, Songlian Li, Yijie Huang, Chi Zhang, Daling Wang, Shi Feng, and Jigang Li. 2024. LangGPT: Rethinking Structured Reusable Prompt Design Framework for LLMs from the Programming Language. *arXiv preprint arXiv:2402.16929* (2024).
- [23] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain

- of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
- [24] 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 (2022), 24824–24837.
  - [25] Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm’s internal prior and external evidence. *Preprint* (2024).
  - [26] Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are RAG models? Quantifying the tug-of-war between RAG and LLMs’ internal prior. *arXiv preprint arXiv:2404.10198* (2024).
  - [27] Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. In *Proceedings of the ACM on Web Conference 2024*. 1362–1373.
  - [28] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2369–2380. <https://doi.org/10.18653/v1/D18-1259>
  - [29] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).
  - [30] Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. *arXiv preprint arXiv:2304.13007* (2023).
  - [31] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems* 42, 4 (2024), 1–60.
  - [32] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117* (2023).
  - [33] Denny Zhou, Nathanael Schärlí, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).
  - [34] Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024. Metacognitive Retrieval-Augmented Large Language Models. *arXiv:2402.11626 [cs.CL]*

## A APPENDIX

### A.1 Experiment about Model Turn

We investigate the relationship between the number of model turn and the EM score, focusing exclusively on instances where the frequency exceeded 1%. As presented in Table 7, our results show that the model achieves relatively high EM when the number of jumps is 2 and 5. This can be attributed to the fact that most of our test datasets consist of two-hop problems. When the model successfully decomposes the problem into sub-problems that align with the dataset, it yields better results. Additionally, since we set the maximum number of decomposed sub-problems to 5, the model performs well when it breaks down the problem into more detailed and granular components.

**Table 7: Analyze the relationship between the number of model turn and EM.**

Models	1	2	3	4	5
<b>HotpotQA</b>	47.37	<b>49.76</b>	33.59	27.66	44.44
<b>2WikiMultihopQA</b>	14.27	45.25	33.93	72.26	<b>57.14</b>
<b>MuSiQue</b>	~	<b>24.84</b>	17.93	5.30	~
<b>Bamboogle</b>	~	64.04	50.00	42.86	<b>66.67</b>

### A.2 Algorithm of HiRAG

For clarity, we outline the workflow of the HiRAG framework in Algorithm 1, providing a step-by-step illustration of our method. This algorithm contains the Main function, Decompose function, RewriteAndAnswerQuestion function and Retrieval function. The Decompose function and the main function complete the decomposition and final summary of the problem, the Retrieval function describes the hierarchical Retrieval and the process of the Filter, and the RewriteAndAnswerQuestion function describes our solution when the semantics are found to be incomplete.

### A.3 Case Study

We illustrate the effectiveness of the HiRAG framework through a case study on a multi-hop question, as depicted in Figure 5. Specifically, the original question presented in the figure is decomposed into two sub-questions, which are then retrieved and verified in a layered manner. Sub-questions that fail verification are re-evaluated through a rethink process. Notably, the second sub-question in the figure yields the correct result after multiple chunk-level rethinks. Ultimately, we combine the sub-answers to form the final answer to the original question.

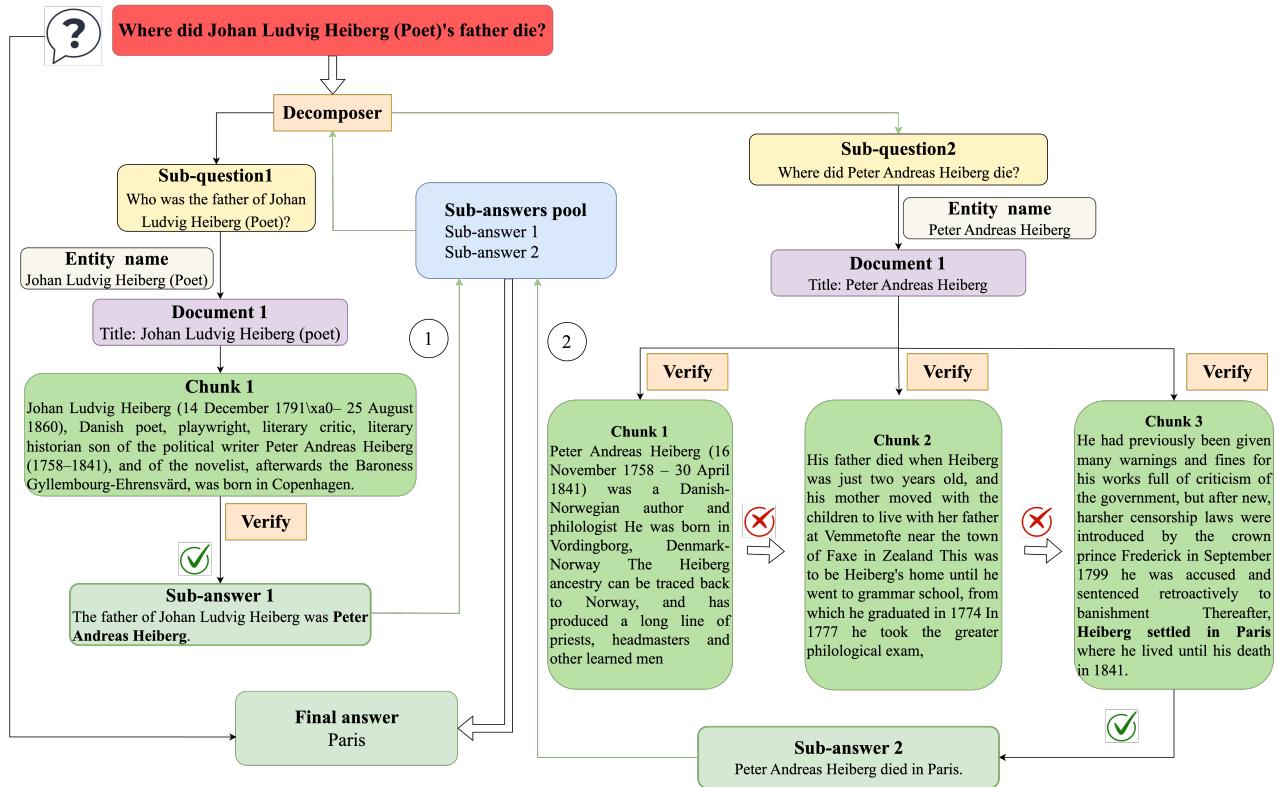


Figure 5: Case study of HiRAG process.

**Algorithm 1:** Description of the process of HiRAG.

---

**Input** :Multi-hop question:  $x$ ; Large Language Model: LLM;

**Initialize**:Decompose rounds:  $w$ ; Corpus:  $D = \{d_1, d_2 \dots, d_m\}$ ; Previous answers:  $A = \{a_1, a_2 \dots, a_{w-1}\}$ ; Previous questions:  $Q = \{q_1, q_2 \dots, q_{w-1}\}$ ; Rethink rounds:  $t$ ; Threshold:  $th_1, th_2, th_3$ .

**Output** :Final answer:  $o$ .

```

1 Function Decompose( $x, w, A$ ):
2    $q = \text{LLM}(x, w, A)$ ; //Use the original question and previous sub-answers to get new sub-question
3   return  $q$ ;

4 Function RewriteAndAnswerQuestion( $q, A, x, D, e$ ):
5   //There is the case of semantic incompleteness
6    $IncFlag = \text{False}$ ;
7   foreach  $a_i$  in  $A$  do
8     if  $e$  in  $a_i$  then
9        $q^* = \text{LLM}(q, a_i)$ ; //rewrite the question
10       $IncFlag = \text{True}$ ;
11   if  $IncFlag == \text{False}$  then
12     //previous sub-answers don't contain question's entity
13      $q^* = \text{LLM}(q, x)$ ; //rewrite the question
14    $c_s = \text{Google}(q^*)$ ;  $a_w = \text{LLM}(c_s, q)$ ;
15   return  $a_w$ ;

16 Function Retrieval( $q, D, x, A$ ):
17    $e = \text{LLM}(q)$ ;  $t = 0$ ; //get the entity name
18   while True do
19      $t = t + 1$ ;
20      $d_s = \text{SparseRetrieval}(e, D, t)$ ; //get the document
21     if  $\text{len}(d_s) > th_1$  then
22       //Multiple entities have high similarity and there is semantic incompleteness
23       return RewriteAndAnswerQuestion( $q, A, x, D, e$ );
24      $c_s = \text{DenseRetrieval}(d_s, q, t)$ ; //get the chunk
25      $\text{CanSolved} = \text{LLM}(c_s, q)$ ; //with the retrieval chunk can LLM solve the question
26     if  $\text{CanSolved} == \text{"yes"}$  then
27       //LLM can solve the question
28        $a_w = \text{LLM}(c_s, q)$ ; //get the answer for the question?
29       return  $a_w$ ;
30     else
31       if  $t >= th_2$  and ( $t \bmod th_2 == 0$ ) then
32         //the chunk level rethink fail, go to the document level rethink
33          $d_s = \text{SparseRetrieval}(e, D, t)$ ; //get other document
34       if  $t >= th_3$  then
35         //Too many attempts, empty result returned
36         return "";
37       //go to the chunk level rethink

38 Function Main( $x, w, D, A, Q$ ):
39    $q = \text{Decompose}(x, w, A)$ ;
40   while not ( $q == \text{"That's enough"}$  or  $q$  in  $Q$ ) do
41      $a_w = \text{Retrieval}(q, D, x, A)$ ; //Get the answer.
42      $w = w + 1$ ;
43      $q = \text{Decompose}(x, w, A)$ ;
44    $Output = \text{LLM}(x, A)$ ; //Get the output for original question using all sub-answers.
45   return  $Output$ ;

```

---