

Retrieval-Augmented Generation by Evidence Retroactivity in LLMs

Liang Xiao^{1*}, Wen Dai², Shuai Chen², Bin Qin², Chongyang Shi^{1†}, Haopeng Jing¹, Tianyu Guo¹

¹Beijing Institute of Technology, ²Xiaomi Corporation

{patrickxiao, cy_shi, jinghp, guotianyu}@bit.edu.cn, {daiwen, chenshuai3, qinbin}@xiaomi.com

Abstract

Retrieval-augmented generation has gained significant attention due to its ability to integrate relevant external knowledge, enhancing the accuracy and reliability of the LLMs' responses. Most of the existing methods apply a dynamic multiple retrieval-generating process, to address multi-hop complex questions by decomposing them into sub-problems. However, these methods rely on an unidirectional forward reasoning paradigm, where errors from insufficient reasoning steps or inherent flaws in current retrieval systems are irreversible, potentially derailing the entire reasoning chain. For the first time, this work introduces **Retroactive Retrieval-Augmented Generation** (RetroRAG), a novel framework to build a retroactive reasoning paradigm. RetroRAG revises and updates the evidence, redirecting the reasoning chain to the correct direction. RetroRAG constructs an evidence-collation-discovery framework to search, generate, and refine credible evidence. It synthesizes inferential evidence related to the key entities in the question from the existing source knowledge and formulates search queries to uncover additional information. As new evidence is found, RetroRAG continually updates and organizes this information, enhancing its ability to locate further necessary evidence. Paired with an Answerer to generate and evaluate outputs, RetroRAG is capable of refining its reasoning process iteratively until a reliable answer is obtained. Empirical evaluations show that RetroRAG significantly outperforms existing methods.

Introduction

Large language models (LLMs), such as ChatGPT(OpenAI 2022) and ChatGLM(THUDM 2024), have demonstrated outstanding performance across a wide range of natural language processing tasks. However, despite the vast amount of knowledge stored during training, these models still exhibit a tendency to generate hallucinatory content, resulting in unverified or factually incorrect answers(Huang et al. 2023a; Wu, Wu, and Zou 2024). To address this issue, the Retrieval-Augmented Generation (RAG) framework is leveraged to acquire and subsequently inject relevant external source knowledge into the LLM's prompt, significantly enhancing the accuracy and reliability of LLM's responses(Lewis et al. 2020; Mao et al. 2021; Chen et al. 2024).

*This work was completed during an internship at Xiaomi Corporation.

†Corresponding authors

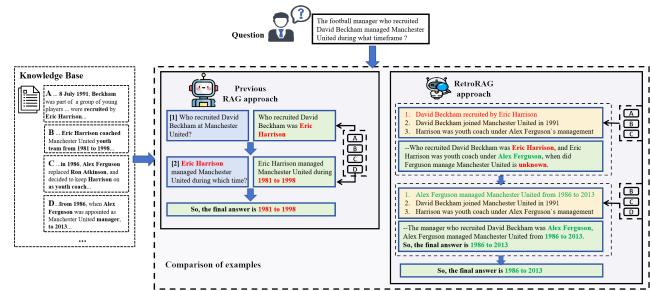


Figure 1: An example of previous RAG approaches causes hallucinatory content due to their unidirectional forwards reasoning paradigm, and how RetroRAG address this issue.

Traditional retrieval-augmented models typically retrieve and extract knowledge documents once based on the initial query, these approaches struggle with addressing multi-hop complex questions due to insufficient knowledge. To tackle this issue, recent studies have transformed the single retrieval-generating into a dynamic multiple retrieval-generating process. These approaches decompose the complex question into several sub-questions, and obtain the final output by answering all these sub-questions(Shao et al. 2023; Trivedi et al. 2023; Press et al. 2023; Yao et al. 2023). Even though the latest approaches have been proposed to improve the effectiveness of knowledge documents retrieval(Asai et al. 2024; Xu et al. 2024), current RAG frameworks are susceptible to the threat of insufficient reasoning from the documents which are factual, and related, but irrelevant due to the inherent flaws of current retrieval systems(Wu et al. 2024), and cause external hallucination. As illustrated in Figure 1, the excessive focus on the local sub-question of *who recruited Beckham?* obtaining the incorrect answer of *Eric Harrison*, the youth coach of Manchester United who *recruited Beckham* in the youth team but never as the *manager*, rather than the manager at time *Alex Ferguson*, and mislead the following reasoning steps.

We argue that this flaw originates from the **Unidirectional Forwards** reasoning paradigm inherent in traditional RAG methods. In this paradigm, any errors produced during reasoning steps are irreversible for the whole reasoning chain. Although the paradigm can be altered by enabling LLMs to continuously reason from scratch through the cumulative retrieval of documents(Trivedi et al. 2023; Zhou et al. 2024), due to the limited useful information in

documents, excessive retrieval would introduce much more noise information which distracts LLMs to engage in over-reasoning and build erroneous correlation, thereby generates hallucinatory content(Yu et al. 2023; Chiang and Lee 2024; Wu et al. 2024; Liu et al. 2024).

To address these issues, we refer the investigative process of the detective, that iteratively collates evidence to gather all related factual information, and through evidence discovery process to validate the relevance of evidence then update them, while uncovering unresolved issues along the way, to ultimately reach a valid conclusion. By continuously revising and reconstructing the reasoning chain through the evidence collation and discovery process, a comprehensive and definitive conclusion can be obtained(Findley 2011; Fahsing 2022). This evidence-collation-and-discovery structure allows LLMs to rethink and revise the reasoning chain through a **Retroactive** paradigm, correcting previous errors resulting from insufficient information by utilizing newly discovered evidence. As illustrated in Figure 1, LLMs can collate the evidence that *Eric Harrison* was the youth coach of the manager *Alex Ferguson* since 1986, and *Ferguson* was actually the *manager*, then update the reasoning process and answer correctly. Inspired by this detective-like approach, we introduce the **Retroactive Retrieval-Augmented Generation** framework (RetroRAG).

To effectively generate and utilize evidence, there are two aspects need to be considered: (1)*The Effectiveness*: the evidence should align with external inherent knowledge(Flores and Woodard 2023), while being attributable(Gao et al. 2023) and relevant(Liu et al. 2024), to avoid the irrelevant noise; (2)*Dynamic Updating*: the evidence should be continuously updated based on newly discovered information and overarching question, rather than being confined to local sub-questions. Hence, RetroRAG constructs Evidence-collation-and-discovery framework (**ELLERY**) to retrieve, generate, and update the evidence, which involves two major components: (1)**Evidence Collation** retrieves relevant documents from the retrieval corpus as the **source evidence**, which would be utilized as doubtless material to generate **inferential evidence**, besides serving as the primary reference for the answering. The source evidence will be continuously updated as the question-solving process progresses, to mitigate the influence of retrieved irrelevant information; (2)**Evidence Discovery** first generates as much inferential evidence related to the key entities in question as possible from the source evidence. Then, inferential evidence will be filtered from the perspectives of relevance to the question and its attribution to the source evidence, to ensure the effectiveness. Inferential evidence would also be updated to only remain the most relevant parts. While both evidence would be used to help generate answers, the gap between the stored evidence and the initial question would be simultaneously analyzed, and the search-query would be proposed for retrieving more information in need.

Along with the **Answerer** to generate and evaluate the answer, RetroRAG provides an approach for refining effective reasoning chains through credible evidence. The experimental results on two multi-hop question answering (QA) datasets verify the effectiveness and state-of-the-art perfor-

mance of RetroRAG, while also demonstrating the explainability in the reasoning process.

The contributions of this paper are summarized as:

- We introduce RetroRAG approach, an innovative retrieval-augmented generation framework. Unlike existing RAG approaches uses an unidirectional forwards reasoning paradigm that cannot reverse the error in preceding reasoning steps, RetroRAG uses a retroactive reasoning paradigm that can revise and reconstruct the reasoning chain through two types of evidence, provides effective answers with less hallucination.
- To the best of our knowledge, this is the first time an evidence-collation-and-discovery framework has been proposed and used in a retrieval-augmented framework, which generates and updates the evidence to support the reasoning process, significantly enhances knowledge retrieval performance on question-answering tasks.

RELATED WORK

Hallucination in Large Language Model

Currently, hallucination is referred as generated content that either does not align with real-world facts or deviates from the source material and self-consistency(Huang et al. 2023b; Ye et al. 2023). In the context of question-answering tasks, hallucination specifically manifests as the generation of arbitrary, and incorrect answers. This phenomenon occurs because, in cases of hallucination, the internal consistency of the generation process in LLMs is unstable. (Manakul, Liusie, and Gales 2023; Mündler et al. 2024; Farquhar et al. 2024). Some studies consider addressing the hallucination problem based on the tendencies of generation from LLMs, they generate multiple outputs and then employ a majority voting strategy to obtain relatively reliable answers(Wang et al. 2022; Huang et al. 2022). More studies consider that the inconsistency generation of LLMs stems from the lack of knowledge, therefore, they introduce reliable external knowledge through the Retrieval-Augmented Generation (RAG) framework, to enhance the factual or specific domain knowledge of LLMs(He, Zhang, and Roth 2022; Gao et al. 2023; Siriwardhana et al. 2023; Ram et al. 2023). Besides, some methods enhance the LLMs to better perceive factual information by fine-tuning the model with the external knowledge (Lee et al. 2022; Tian et al. 2023).

Retrieval Augmented Language Model

Many studies have demonstrated the impressive performance of the retrieval-augmented language model (RALM) in various natural language tasks, which is enhanced by the provision of detailed and specific external knowledge to supplement LLMs(Lewis et al. 2020; Guu et al. 2020; Shi et al. 2024). These models typically employ a retriever to obtain a set of relevant documents from a knowledge corpus, such as Wikipedia, to enhance the effectiveness of the answers. While initial RALM performs the single-time retrieval strategy, which extracts knowledge once based on the user’s initial query(He, Neubig, and Berg-Kirkpatrick 2021; Izacard and Grave 2021; Ram et al. 2023), recent studies have focused on multi-time retrieval models to overcome the issues

of insufficient knowledge, due to retriever may focus only on parts of the query when addressing multi-hop complex questions. Some models decompose the initial query into multiple sub-questions, then iteratively retrieve knowledge and answer these sub-questions, until the original query can be finally resolved(Yao et al. 2023; Press et al. 2023; Shao et al. 2023; Xu et al. 2024); while the others construct an iterative process of holistic thinking, continuously increasing the amount of knowledge retrieved based on unsolved questions, until effective reasoning can be achieved(Trivedi et al. 2023; Zhou et al. 2024), and more recent studies have explored fine-tuning certain components to enhance the reliability of retrieval process(Yan et al. 2024; Liu et al. 2024).

All these approaches are proven to be effective. However, due to factually related irrelevant documents from the inherent flaws of the current retrieval system, they address issues of insufficient reasoning and over-reasoning, respectively. This paper addresses the aforementioned issues by exploring the application of evidence, to construct a retroactive reasoning process. Through continuously generating and updating credible evidence, our work builds an effective RAG framework to address the hallucination in question answering task without fine-tuning or pre-training of LLMs.

Methodology

Existing retrieval augmented methods, due to their unidirectional forward reasoning paradigm, are prone to the risk of external hallucination from factually related irrelevant documents. Since the process of answering decomposed sub-questions can be equivalently regarded as the process of obtaining sub-evidence to answer the initial question, as illustrated in Figure 2, previous approaches have employed a linear paradigm of progressive sub-evidence generation, where the generation of each node is highly depends on the previous nodes. Although the verification of knowledge can prevent the emergence of subsequent unreliable node \textcircled{B} , it is incapable of correcting erroneous validation node \textcircled{C} caused from the inherent flaws of current retrieval systems, like *Eric Harrison* in Figure 1 when it has reached node \textcircled{D} . The LLMs will propagate the erroneous information as definitive knowledge, leading to inaccuracies in the following output.

By generating and updating of evidence, RetroRAG enables LLMs to refine their knowledge by integrating newly evidence \textcircled{D} , \textcircled{J} , \textcircled{K} , with the previous evidence \textcircled{A} , \textcircled{B} , \textcircled{C} based on the relevance to the question, retaining only the most pertinent evidence. This process allows for a more comprehensive understanding as new evidence is integrated(for instance, *Eric Harrison is the youth coach of Alex Ferguson*), while erroneous nodes at any stage are discarded (for instance, *Harrison recruited Beckham as manager* will be correct). The correct nodes \textcircled{A} , \textcircled{J} , \textcircled{K} will be preserved to the next stage. Essentially, it can be considered that RetroRAG constructs a graph-based thinking structure.

Overview

We propose Retroactive Retrieval-Augmented Generation (RetroRAG) framework to tackle the issue of insufficient reasoning and over-reasoning, which applies Answerer to

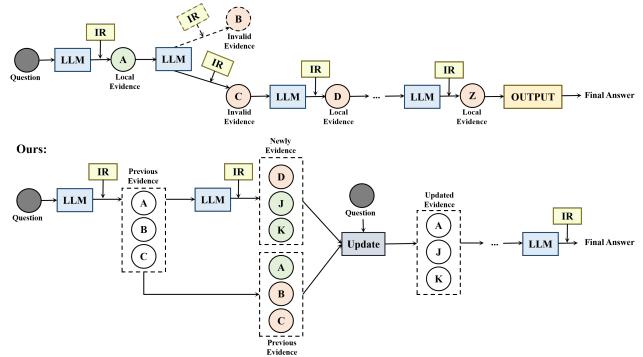


Figure 2: Comparison between previous methods and RetroRAG in mechanism.

generate and evaluate credible answers, and Evidence-collation-and-discovery (ELLERY) framework retrieves, generates, and updates the evidence.

In the QA scenario, the target of RetroRAG is to generate an answer a to the given question q with key entities k . As illustrated in Figure 3, RetroRAG utilizes the iteration application of two processes: (1) *Answerer* generates an answer based on the current knowledge context, and determines if a consistent response can be generated within the current knowledge context. (2) *ELLERY* obtains documents D_Q from the retrieval corpus $D = \{d_i\}_{i=1}^{|D|}$ (with Wikipedia dumps served as the primary data source in this study) related to the question, as the source evidence, and by which generates credible evidence E , and re-queries q_r based on q , k , D_Q , and the last reasoning chain r . ELLERY continues this process of collating and discovering evidence until a definitive answer is obtained by Answerer. The details of the prompts we designed and used will be introduced in Section A of the Appendix.

Answerer

The main target of Answerer is to generate a reliable answer to the question. To achieve this, the Answerer first generates pseudo-answer a and corresponding reason r with the current quantities of information E through an answer-generator, which utilizes COT-prompt M_C with a low-temperature parameter to obtain a more fixed output.

To assess whether the current knowledge context is sufficient, we refer Self-Consistency (SC) concept that outputs of LLMs should converge towards the correct answer under a strong knowledge context, with which LLMs are capable of constructing reliable reasoning and answers. To this end, we employ an SC-generator with a direct-answering-prompt M_D and high-temperature parameter to obtain monitoring answer a_{sc} with more divergent thinking pattern and the same knowledge content of a , and the similarity between these two outputs can assess the self-consistency score and the degree of hallucination. We designed an LLM-based evaluator S to convert scores, e.g., similarity or relevance, into the probability of generating indicative tokens (e.g., 'yes' or 'no'). We calculate the similarity scores s_{sc} through LLM-based evaluator S_{sc} as the following formulation:

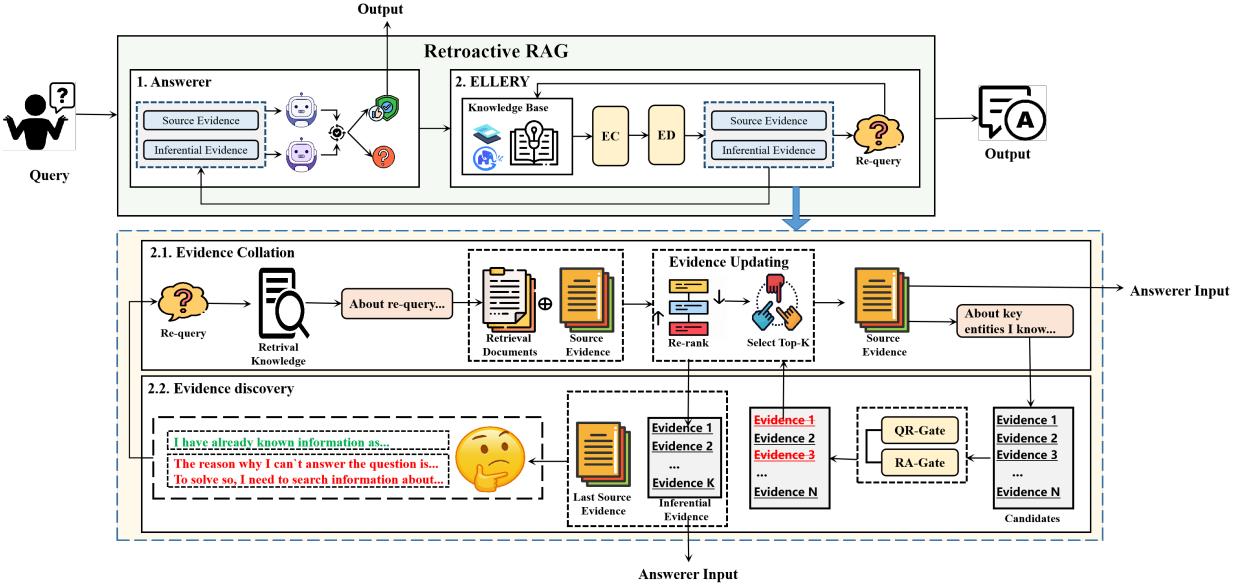


Figure 3: Overview of our RetroRAG structure.

$$x = LLM([a, a_{sc}, q], M_{sc})$$

$$S_{sc} = \frac{P(x = 'yes' | ([a, a_{sc}, q], M_{sc}))}{\sum_{i \in ['yes', 'no']} P(x = i | ([a, a_{sc}, q], M_{sc}))} \quad (1)$$

$$s_{sc} = S_{sc}(([a, a_{sc}, q], M_{sc}))$$

Where M_{sc} is a customized prompt, and a threshold t is utilized to govern the model's output, only when $s_{sc} > t$, the iteration process is stop, and output the answer as the final answer. A higher value of t implies that a more stringent requirement of knowledge context. And follow the research from (Zhou et al. 2024), a declarative assessor is implemented to ensure the standardization of the answer.

Evidence Collation and Discovery

In each round of iteration, the initial query q and its key entities k will be fixed as constant input for ELLERY structure. The ELLERY structure obtains and updates source evidence through Evidence Collation, while based on which generating inferential evidence and proposing re-querie to obtain the required knowledge through Evidence Discovery.

Evidence Collation In the L -th iteration, we use the search query generated from end of the last round $q_s^{(L-1)}$, which is designed to specifically target missing information, to retrieve relevant passages $D_C^{(L)}$. And concatenate $q_s^{(L-1)}$ with q to obtain the matching query $q_m^{(L)} = [q, q_s^{(L-1)}]$, which is designed to match the most relevant evidence within the current knowledge context, while avoiding the issue of deviating from the initial question by focusing too much on the generated search query. Specifically, we set $q_s^{(0)} = q_m^{(0)} = q$ as the initialization. After obtaining $D_C^{(L)}$, we merge them with the last stored source evidence $E_s^{(L-1)}$, and apply the customized prompt M_e with LLM to individually score each source evidence candidate, ranging from 0 to 1, to assess the

contribution of source evidence to answering the question. Since source evidence should be updated to ensure the progression of current answering process, the scoring process of evaluators S_{E_s} can be formulated as:

$$s_{E_s} = S_{E_s}([E_s^{(L-1)} \cup D_C^{(L)}, q_m^{(L)}], M_e)) \quad (2)$$

Based on s_{E_s} , we can extract the top-N source evidence as the current source evidence $E_s^{(L)}$, which would be sent to Answerer to help the answering process in the L -th round, and be used to generate the inferential evidence and the re-query in the failure answering case.

Evidence Discovery Follow the actual detective, we uses **Deductive** method to generates inferential evidence, which contains two steps: (1)*Deductive Reasoning*: we utilize an LLM-based evidence generating prompt M_{IE} to generate inferential evidence candidates $e_{ic}^{(L)}$ related to k from $E_s^{(L)}$, to obtain as much **deductive inference** to the initial question as possible, from the current retrieval documents. (2)*Hypothesis Testing*: we design two specified LLM-based gated prompt M_{qr} and M_{ra} to calculate the Question-Relevance (QR) score and the Reference-Attribution (RA) score with their LLM-based evaluator S_{qr} and S_{ra} , to ensure the effectiveness of each inference. The specific functions of these two evaluation scores are as follow:

- *Question-Relevance (QR)*: On knowing the last inferential evidence $E_i^{(L-1)}$, if $e_{ic}^{(L)}$ could be directly related to answering the matching query $q_m^{(L)}$:
- *Reference-Attribution (RA)*: If the claim of $e_{ic}^{(L)}$ can be directly found in any claims of $E_s^{(L)}$.

Through these step, we only reserve the useful and confirmed inference as evidence, and this process can be for-

mulated as:

$$\begin{aligned} s_{qr} &= S_{qr}([e_{ic}^{(L)}, E_i^{(L-1)}, q_m^{(L)}], M_{qr}) \\ s_{ra} &= S_{ra}([e_{ic}^{(L)}, E_s^{(L)}], M_{ra}) \\ e_{ic}^{(L)} &= ((s_{qr} > 0.5) \cap (s_{ra} > 0.5))e_{ic}^{(L)} \end{aligned} \quad (3)$$

Next, we merge $e_{ic}^{(L)}$ and last inferential evidence $E_i^{(L-1)}$. Following the updating process of source evidence, we apply the same evaluator S_e score each inferential evidence candidate. Since inferential evidence should align with the initial question for a long term memory, the scoring process of evaluators S_{E_i} can be formulated as:

$$s_{E_i} = S_{E_i}([E_i^{(L-1)} \cup e_{ic}^{(L)}, q], M_e) \quad (4)$$

We select the top-K inferential evidence as current inferential evidence $E_i^{(L)}$ in the same way, through which achieving the revising and reconstructing of reasoning nodes. And, $E_i^{(L)}$ would be sent to Answerer as referenced evidence in the $(L+1)$ -th round.

It is important to note that since $E_i^{(L)} \in E_s^{(L)} \cup E_i^{(L-1)}$, and current quantities of information $E^{(L)} = E_s^{(L)} \cup E_i^{(L-1)}$, no new information is brought in after the updating of E_i . If Answerer fails to provide an effective answer with the knowledge context $E^{(L)}$, it is necessary to retrieve more source evidence to fill the knowledge gap. To achieve this, we should first know the information $E^{(L)}$ LLM already have right now, and the reason r why LLM can't (or wrongly) answer the question based on these information, then generate a new query to further retrieve information from the corpus to answer the initial question q . Hence, we construct a LLM-based re-query generator G_R with customized prompt M_R , to generate search query $q_s^{(L)}$ to deduce what information LLM needs to answer the question:

$$q_s^{(L)} = LLM([E_s^{(L)}, E_i^{(L)}, r, q], M_R) \quad (5)$$

Experiments

In this section, we evaluate the effectiveness of our proposed model on two multi-hop question answering (QA) datasets.

Experimental Setup

Datasets and Evaluation Metrics We conduct experiments on two multi-hop question answering datasets: HotpotQA(Yang et al. 2018) and 2WikiMQA(Ho et al. 2020). Since both of the datasets are constructed based on Wikipedia documents, we use the same document corpus and retrievers to provide external references for LLMs. Due to the constraints of experimental costs, following(Zhou et al. 2024), we sub-sample 500 questions from the validation set of each dataset for experiments.

For evaluation metrics, we use exact match (EM) as our standard metrics at answer-level, to measure whether the predicted answer is completely consistent with the standard answer. And we use token-level F1, precision (Pre) and recall (Rec) for comprehensive evaluation at token-level, to evaluate the proportion of correct answer tokens in the overall tokens.

Baselines We compare our RetroRAG to recent baseline approaches: Standard Prompting(Brown et al. 2020), Chain-of-Thought(Wei et al. 2023), Standard RAG(Lewis et al. 2020), ReAct(Yao et al. 2023), Self-Ask(Press et al. 2023), IR-COT(Trivedi et al. 2023), SearChain(Xu et al. 2024), and MetaRAG(Zhou et al. 2024). We comprehensively describe each baseline in Appendix B.1 and explain the rationale behind selecting these specific baselines.

Settings We choose GLM4-9B-chat(THUDM 2024) LLM as the base LLM for all baseline and our RetroRAG approach with the temperature setting of 0.01, except the SC-generator of our RetroRAG whose temperature is set to 1.00. We utilize the Wikipedia dump(Karpukhin et al. 2020) as the document corpus for both datasets, where articles are segmented into passages of 100 tokens. We employ the BM25 algorithm(Robertson, Zaragoza et al. 2009) and SimLM retriever(Wang et al. 2023) to retrieve the top 5 relevant passages to be the external knowledge for all approaches. And we set a default judgment threshold for our answering evaluation mechanism at 0.7 to ensure consistency of answers. The maximum number of both iterations and the size of the evidence repository are set to 5.

Main Results

Performance on multi-hop question answering datasets is shown in Table 1. It can be observed that:

(1) Our proposed RetroRAG consistently surpasses all baseline methods across two datasets. At answer-level, the performance improvement on EM is **+8.8** on HotpotQA and **+5.2** on 2WikiMQA compared to the best baseline results; At token-level, the performance improvement on F1 is **+10.6** on HotpotQA and **+4.0** on 2WikiMQA compared to the best baseline results. This suggests that when directly employing LLM, without additional pre-training or fine-tuning, our approach exhibits optimal performance.

(2) When compared to SearChain, which adapts the unidirectional forward reasoning paradigm but verifies each node in COT and outperforms other methods using the same paradigm, such as COT, ReACT, Self-Ask, etc., RetroRAG shows an improvement of **+11.6** on HotpotQA and **+5.2** on 2WikiMQA. This reflects that the retroactive reasoning paradigm RetroRAG uses can solve the issue of local insufficient reasoning, and provides more comprehensive reasoning, thereby improving the performance markedly.

(3) When compared to IR-COT and MetaRAG, which also do not adhere to the paradigm of linear reasoning, but increase the quantity of retrieved documents to re-generate the answer, RetroRAG shows an improvement of **+8.8** on HotpotQA and **+9.2** on 2WikiMQA. This reflects that the evidence-collation-and-discovery framework RetroRAG uses can address the issue of over-reasoning, and mitigate the irrelevant and noisy information from knowledge documents, thereby improving the performance significantly.

(4) Compared with Standard Prompting and COT approaches, both the idea of decomposing the initial query into multiple sub-questions and iteratively increasing the amount of knowledge retrieved can doubtlessly improve the ability of reasoning of LLMs. When multi-hop questions

Table 1: Evaluation results on two multi-hop question answering datasets. '*' denotes the result outperforms baseline models in t-test at $p < 0.05$ level. The best results are in **bold**, and the second best results are underlined.

Methods	HotpotQA				2WikiMQA			
	EM	F1	Pre	Rec	EM	F1	Pre	Recall
Standard Prompting(Brown et al. 2020)	14.2	21.7	23.2	21.2	21.4	27.3	28.8	26.9
Chain-of-Thought(Wei et al. 2023)	16.8	24.9	26.1	24.9	23.6	30.4	31.1	30.5
Standard RAG(Lewis et al. 2020)	23.4	35.6	36.6	36.6	22.8	26.8	27.2	28.0
ReAct(Yao et al. 2023)	20.6	29.4	29.6	32.1	21.2	28.5	28.2	30.3
Self-Ask(Press et al. 2023)	24.8	35.1	36.5	36.4	29.4	36.7	36.4	38.2
IR-COT(Trivedi et al. 2023)	30.4	40.1	41.6	41.0	25.6	30.9	31.0	32.1
SearChain(Xu et al. 2024)	29.6	41.2	41.5	43.4	33.4	42.6	42.5	44.8
MetaRAG(Zhou et al. 2024)	<u>32.4</u>	<u>44.3</u>	<u>45.5</u>	<u>45.6</u>	28.8	36.0	35.7	38.4
RetroRAG	41.2*	54.9*	56.2*	58.3*	38.6*	46.6*	46.9*	49.5*

present a clearer progressive structure, such as samples in 2WikiMQA dataset, the approaches of decomposing sub-problems will perform better, so Self-Ask and SearChain demonstrate superior performance compared to IR-COT and MetaRAG in 2WikiMQA dataset since the latter may introduce excessive noise; On the contrary, the idea of collecting enough knowledge would help more in answering since the decomposing-answering mode could mislead the reasoning process, causes IR-COT and MetaRAG perform better on HotpotQA dataset. Due to our work’s design of generation of inferential evidence and updating of evidence, which suppresses the introduction of noisy information and achieves a retroactive-progressive reasoning structure, achieving the best performance above all baselines on both datasets.

Table 2: Ablation Studies on RetroRAG.

	HotpotQA		2WikiMQA	
	EM	F1	EM	F1
RetroRAG	41.2	54.9	38.6	46.6
Ablation of structure				
-w/o AE	38.6	51.7	32.6	40.1
-w/o ELLERY	19.0	25.5	26.0	30.7
Ablation of evidence utilization				
-w/o SE	34.8	46.1	29.2	35.6
-w/o IE	38.4	51.9	36.6	44.5
-w/o UoE	39.4	53.0	38.2	45.9
-w/o EoE	40.5	54.8	38.0	45.8

Ablation Study

To verify the effectiveness of different components in RetroRAG, we conduct a comparative analysis respectively focus on frameworks, including Answerer (solely on the Answer Evaluation (AE) mechanism since Answerer needs to generate answer output) and ELLERY; Besides, we also focus on the designs of evidence utilization, including Source Evidence (SE), Inferential Evidence (IE), Updating of Evidence (UoE) and Evaluation of Evidence (EoE). The results are shown in Table 2.

Ablation of structure As shown in Table 2, it is evident that removing either of the frameworks adversely affects the performance of both datasets. Specifically, the removal of AE makes the LLM consistently set to a state of knowledge deficiency, thereby posing the risk of over-reasoning

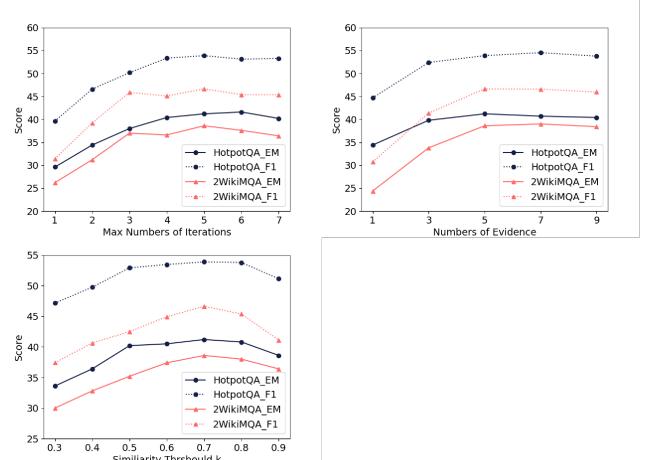


Figure 4: Comparison with different maximum numbers of iterations, numbers of evidence, and similarity thresholds.

that turns the correct reasoning path built, into the incorrect one. Meanwhile, the removal of ELLERY would result in a complete absence of external knowledge, thereby leading to a significant decline in performance. This emphasizes the notion that many hallucinations stem from an insufficiency in external knowledge.

Ablation of evidence utilization To conduct a more detailed analysis of the mechanisms of ELLERY, we performed ablation studies based on the characteristics of the evidence and the methods of evidence processing. Experimental results show that due to the complementary design, both source evidence and inferential evidence are crucial. Inferential evidence provides a summary of past effective information, thereby helping enhance the performance of LLMs. However, since it contains much less information than source evidence, source evidence has a greater impact on the quality of responses. Besides, the updating and evaluating of evidence can alleviate the introduction of irrelevant information, ensuring the progression of the current answering process while aligning with the initial question, thereby enhancing performance.

Qualitative Analysis

To better delve into the impact of the numbers of iterations and evidence, as well as similarity thresholds, we have em-

barked on a series of qualitative experiments. The results are shown in Figure 4. It can be observed that:

Different maximum numbers of iterations Although the Answer Evaluation mechanism plays a significant role in the performance of RetroRAG, the maximum number of iterations also has a substantial impact on the final results. As depicted in Figure 4, we can find that on both datasets, the accuracy of RetroRAG improves progressively before the iterations reach 3, and then grows relatively stable, peaking when the iterations reach 5 or 6. After the peak, we observed a risk of decline in the performance with the increase of iterations. We think these results suggest that, by increasing the number of iterations, RetroRAG can extract more effective evidence, thereby improving performance. However, excessively increasing may cause over-reasoning and introduce noise information, which in turn to a decline in performance.

Different numbers of evidence For the convenience of the experiment, we set the numbers of both source evidence and inferential evidence to the same. Based on this setting, we design a series of experiments to investigate the impact of the number of evidence. As depicted in Figure 4, we find that on both datasets, increasing the numbers of evidence from 1 to 3 can significantly improve the performance of RetroRAG, which reflects that under the condition of insufficient information, it is difficult for LLMs to perform effective reasoning. The accuracy of RetroRAG reaches the peak at the number of 5 pieces of evidence, and then slowly decreases as the increase of the evidence. This suggests that although the filtering and updating mechanism can suppress irrelevant information to some extent, an excessively large evidence window can still introduce noise and impair performance.

Different similarity thresholds Although threshold $t > 0.5$ signifies that the LLM evaluator considers the result to be valid, a higher t represents the LLM has more confidence in the determination. This could lead to a more reliable monitor but might also result in overly stringent requirements for the results, causing misjudgments and excessive iterations. As illustrated in Figure 4, we find that for HotpotQA dataset, $t = 0.5$ can provide a good discriminative effect, but for 2WikiMQA, LLMs need a higher threshold which is about 0.7 to ensure the effectiveness. And for both datasets, excessively high thresholds lead to a decline in performance. This is also related to the over-reasoning issue.

Case Study

Due to factually related irrelevant documents from the inherent flaws of the current retrieval system, previous approaches cause issues of insufficient reasoning and over-reasoning, respectively. To address these issues, we introduce inferential evidence as a form of knowledge caching. By summarizing and updating past relevant information, this approach enables retroactive awareness when new knowledge is introduced, while simultaneously reducing interference from irrelevant information. The case in Figure 5 provides an intuitive demonstration of the differences between our RetroRAG and previous approaches, traditional

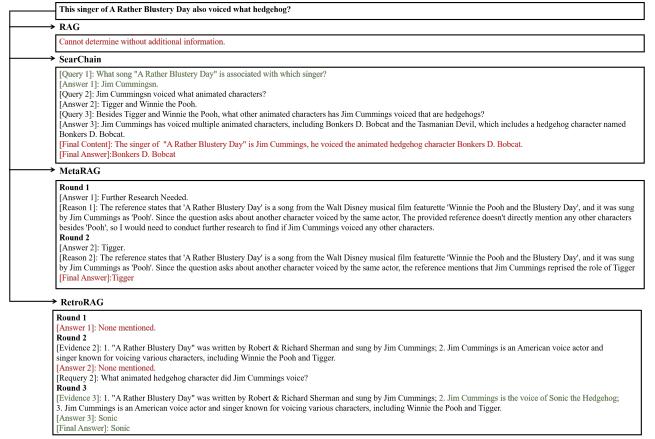


Figure 5: Case study of RetroRAG and previous approaches.

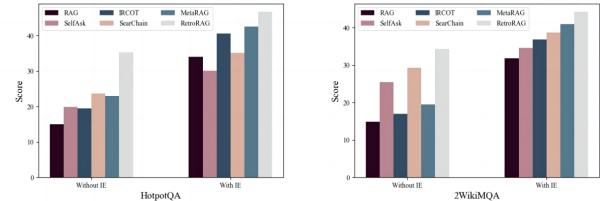


Figure 6: Comparison of different evidence cases of RetroRAG and previous approaches.

RAG stops reasoning due to insufficient retrieval information; SearChain makes the correct decomposing process but answers wrongly due to the related irrelevant knowledge; MetaRAG is interfered with excessive retrieval documents and makes the wrong reasoning process. Being aware of the evidence, RetroRAG makes the correct reasoning process that leads to an accurate answer. More cases would be given in Section B.2 of the Appendix.

Effectiveness of evidence

To better explore the effectiveness of Inferential Evidence, we divided the data into two categories based on the presence of evidence generated through Question-Relevance and Reference-Attribution evaluations during the iterative process. We then analyzed the differences in performance on the main baselines. As shown in Figure 6, we find in scenarios where Inferential Evidence is generated, iteratively increasing the amount of knowledge performs better compared to decomposing sub-questions. We believe this because the retrieval knowledge is extensively relevant and mutually corroborative, enhancing the LLM’s performance. Furthermore, our work enhances the LLM’s understanding of knowledge by generating Inferential Evidence, thereby further improving the reliability of its responses. Additionally, the absence of Inferential Evidence generation may be due to the sparse distribution of required knowledge within individual documents. In such cases, decomposing sub-questions allows the LLM to better understand and answer the questions. Moreover, our work ensures the accumulation of effective information through Source Evidence updating, thereby also enhancing the LLM’s performance.

Conclusion

In this paper, we point out the threat from the unidirectional forward reasoning paradigm inherent in traditional RAG methods, within which any errors produced during reasoning steps are irreversible and affect the whole reasoning chain. We then introduce RetroRAG, a novel framework that uses a detective-like retroactive reasoning paradigm that can revise and reconstruct the reasoning chain, ensuring it on the correct direction. Through the evidence-collation-discovery framework, RetroRAG can search, generate, and update credible evidence, empower the model to perceive existing information, and seek out more necessary evidence to complete the reasoning process. Experimental results on two multi-hop QA datasets demonstrated that RetroRAG performs better than all baselines. In the future, we aspire to explore the possibility of allowing LLMs to independently learn the aforementioned evidence-collation-discovery process through methods such as fine-tuning or pre-training.

References

- Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165*.
- Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Cheng, Q.; Sun, T.; Zhang, W.; Wang, S.; Liu, X.; Zhang, M.; He, J.; Huang, M.; Yin, Z.; Chen, K.; et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.
- Chiang, C.-H.; and Lee, H.-y. 2024. Over-Reasoning and Redundant Calculation of Large Language Models. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 161–169. St. Julian’s, Malta: Association for Computational Linguistics.
- Fahsing, I. 2022. Beyond reasonable doubt: how to think like an expert detective. In *Police psychology*, 267–295. Elsevier.
- Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.
- Findley, K. A. 2011. Adversarial inquisitions: Rethinking the search for the truth. *NYL Sch. L. Rev.*, 56: 911.
- Flores, C.; and Woodard, E. 2023. Epistemic norms on evidence-gathering. *Philosophical Studies*, 180(9): 2547–2571.
- Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A. T.; Fan, Y.; Zhao, V.; Lao, N.; Lee, H.; Juan, D.-C.; and Guu, K. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16477–16508. Toronto, Canada: Association for Computational Linguistics.
- Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org.
- He, H.; Zhang, H.; and Roth, D. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.
- He, J.; Neubig, G.; and Berg-Kirkpatrick, T. 2021. Efficient Nearest Neighbor Language Models. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5703–5714. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Ho, X.; Duong Nguyen, A.-K.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 6609–6625. Barcelona, Spain (Online): International Committee on Computational Linguistics.
- Huang, J.; Gu, S. S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. Online: Association for Computational Linguistics.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.

- Lee, N.; Ping, W.; Xu, P.; Patwary, M.; Fung, P. N.; Shoeybi, M.; and Catanzaro, B. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35: 34586–34599.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktaschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.
- Liu, Y.; Peng, X.; Zhang, X.; Liu, W.; Yin, J.; Cao, J.; and Du, T. 2024. RA-ISF: Learning to Answer and Understand from Retrieval Augmentation via Iterative Self-Feedback. *arXiv:2403.06840*.
- Manakul, P.; Liusie, A.; and Gales, M. 2023. SelfCheck-GPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017. Singapore: Association for Computational Linguistics.
- Mao, Y.; He, P.; Liu, X.; Shen, Y.; Gao, J.; Han, J.; and Chen, W. 2021. Generation-Augmented Retrieval for Open-Domain Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.
- Mündler, N.; He, J.; Jenko, S.; and Vechev, M. 2024. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. *arXiv:2305.15852*.
- OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt>.
- Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics.
- Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.
- Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.
- Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; and Chen, W. 2023. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics.
- Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2024. REPLUG: Retrieval-Augmented Black-Box Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8371–8384. Mexico City, Mexico: Association for Computational Linguistics.
- Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; and Nanayakkara, S. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 11: 1–17.
- THUDM. 2024. GLM-4. <https://github.com/THUDM/GLM-4>.
- Tian, K.; Mitchell, E.; Yao, H.; Manning, C. D.; and Finn, C. 2023. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.
- Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics.
- Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2023. SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2244–2258. Toronto, Canada: Association for Computational Linguistics.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903*.
- Wu, K.; Wu, E.; and Zou, J. 2024. ClashEval: Quantifying the tug-of-war between an LLM’s internal prior and external evidence. *arXiv:2404.10198*.
- Wu, S.; Xie, J.; Chen, J.; Zhu, T.; Zhang, K.; and Xiao, Y. 2024. How Easily do Irrelevant Inputs Skew the Responses of Large Language Models? *arXiv preprint arXiv:2404.03302*.
- Xu, S.; Pang, L.; Shen, H.; Cheng, X.; and Chua, T.-S. 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, WWW ’24*, 1362–1373. New York, NY, USA: Association for Computing Machinery.
- Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; and Ling, Z.-H. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In

Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsuji, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.

Ye, H.; Liu, T.; Zhang, A.; Hua, W.; and Jia, W. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Yu, W.; Zhang, H.; Pan, X.; Ma, K.; Wang, H.; and Yu, D. 2023. Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. *arXiv:2311.09210*.

Zhou, Y.; Liu, Z.; Jin, J.; Nie, J.-Y.; and Dou, Z. 2024. Metacognitive Retrieval-Augmented Large Language Models. In *Proceedings of the ACM on Web Conference 2024*, 1453–1463. New York, NY, USA: Association for Computing Machinery.

A. Prompt Detail

We show the prompt used in experiment on both datasets in Figure 7 to Figure 13. In this work, we constructed the few-shot CoT prompt, and the declarative assessor prompt by referencing (Zhou et al. 2024), and design different prompt for the corresponding functions. For all calculations as detailed in the Methodology section, we quantified the results by generating probability distributions of the 'yes' and 'no' tokens.

Answer following Question in one or few words.
question: #####
Here are some evidence you can use: #####
Answer:

Figure 7: Prompt M_D for generating the monitoring answer.

Carefully determine whether answer1 and answer2 have same meaning in answering the question. You just need to answer yes or no.

question: #####
answer1: #####
answer2: #####
output:

Figure 8: Prompt M_{sc} for calculating the Self-Consistency score to determine if the answer is reliable.

I need you to help me determine if the provided evidence is RELATED to the query
You just need to answer yes or no.
evidence: #####
query: #####
You just need to answer yes or no.
Answer:

Figure 9: Prompt M_e for calculating the score of relevance between evidence and query for evidence updating.

Please CAREFULLY extract DIRECTLY RELEVANT factual information related to the entity from the reference in SHORT sentences.
Please ONLY extract DIRECTLY RELEVANT factual information from the reference. Extract only the factual details relevant to the question, and ignore any meta information about the lack of knowledge.
Please AVOID using ANY PRONOUNS and you MUST use CLEAR references instead, and provide a full description of the event in information.

If you extract multiple relevant pieces of information, separate them with '#'

Please ONLY extract DIRECTLY RELEVANT factual information from the reference, WITHOUT ADDING ANY INFERS
entity: #####
reference: #####
output:

Figure 10: Prompt M_{IE} for generating inferential evidence.

I need you to help me determine if the provided evidence is RELATED to the query.
You just need to answer yes or no.
evidence: #####
query: #####
You just need to answer yes or no.
Answer:

Figure 11: Prompt M_{qr} for calculating the Question-Relevance score.

I need you to CAREFULLY determine if the evidence can directly comes from the provided reference.
Please note:
1. If the evidenceONLY include statements explaining why the question cannot be answered based on the given information, the answer is "no".
2. If the references contains description against the evidence, the answer should be "no".
reference: #####
evidence: #####
You just need to answer yes or no
answer:.

Figure 12: Prompt M_{ra} for calculating the Reference-Attribution score.

Question: #####
Reference: #####
Evidence: #####
Reason: #####
A person answered a question based on given evidence and reference information. However, due to issues with lacking of information, more references need to be retrieved now. Please carefully consider his reason and questions step by step, and provide a query for searching based on the information he needs to answer the question in SHORT sentences.

Figure 13: Prompt M_R for generating re-query.

B. Experimental Details

Baselines

We compare our proposed model with several state-of-theart baselines listed as follows:

- Standard Prompting(Brown et al. 2020): Standard Prompting directs LLM to answer the queries with a simple question-answer prompt.
- Chain-of-Thought(Wei et al. 2023): Chain-of-Thought provides a few-shot prompt to LLM, make it can answer with a reasoning process.
- Standard RAG(Lewis et al. 2020): Standard RAG first retrieves multiple documents by query, then inject these documents into prompts to LLM for answering.
- ReAct(Yao et al. 2023): ReAct introduces a reasoning and acting paradigm, alternately executing reasoning and task-specific actions to complete the QA task.
- Self-Ask(Press et al. 2023): Self-Ask introduces a paradigm that decomposes questions into sub-questions, continuously engaging in self-questioning until the final answer is obtained.
- IR-COT(Trivedi et al. 2023): IR-COT iteratively alternates between using COT reasoning to guide retrieval, and utilizing retrieval results to enhance CoT reasoning, continuing executing until the final answer is obtained.
- SearChain(Xu et al. 2024): SearChain introduces the concept of "search-in-chain," which corrects the reasoning process through the interaction between Information Retrieval (IR) and Chain-of-Query (COQ), let IR provides the knowledge that LLM really needs.
- MetaRAG(Zhou et al. 2024): MetaRAG combines the RAG process with metacognition, allowing LLM to execute different actions based on the reliability of internal and external knowledge, identify the sufficiency of knowledge and potential errors during reasoning.

Additional Case Studies

We present additional cases to further demonstrate the effectiveness of our proposed RetroRAG approach. As shown in Figure 14, we find that RetroRAG not only leverages the effective information from inferential evidence to make accurate reasoning, but can also compensate for insufficient inferential evidence by retrieving complementary source evidence through re-query, thereby enabling correct reasoning, which also highlights the necessity of leveraging both inferential evidence and source evidence.

Alice David is the voice of Lara Croft in a video game developed by which company?
→ RAG
Cannot determine.
→ SearChain
[Query 1] Which video games feature the character Lara Croft? [Answer 1]: The character Lara Croft is primarily associated with the "Tomb Raider" series. [Query 2] Which companies have developed the "Tomb Raider" series? [Answer 2]: Tomb Raider was developed by Crystal Dynamics, Eidos Interactive, Crystal Dynamics, and currently published by Square Enix. [Query 3] Has Alice David provided voice acting for any of the "Tomb Raider" game? [Answer 3]: There seems to be a mix-up. The reference indicates Alice Bowie, not Alice David, is a fictional character. [Final Context]: Given the new understanding that Alice Bowie is a fictional character and not Alice David, we need to identify the correct Alice David and confirmed her work in Tomb Raider and which company developed the game where she voiced Lara Croft. [Final Answer]: Need to identify the correct Alice David
→ MetaRAG
Round 1 [Answer 1]: Unknown. [Reason 1]: There is no mention of 'Alice David' in the provided references. Therefore, I cannot determine which company developed the video game featuring Alice David as the voice of Lara Croft. Round 2 [Answer 2]: Unknown. [Reason 2]: There is no mention of 'Alice David' in the provided references. Therefore, I cannot determine which company developed the video game featuring Alice David as the voice of Lara Croft. [Final Answer]: Unknown
→ RetroRAG
Round 1 [Answer 1]: Unknown. [Reason 1]: Alice David voiced Lara Croft in the German dub of the 2013 video game "Tomb Raider". 2. Crystal Dynamics developed the "Tomb Raider" video game series, including the 2013 reboot. 3. Lara Croft's body was redesigned with more muscle tone and a portrait in the 2013 reboot, replacing the previous braided hairstyle. 4. Crystal Dynamics made significant graphical improvements and added gameplay features to the "Tomb Raider" series following the poor reception of the 2003 sequel. 5. Keeyah flaws initially voiced Lara Croft before being succeeded by Camilla Luddington in 2012 for future games in the series. [Answer 2]: Crystal Dynamics [Final Answer]: Crystal Dynamics
Which Eminem album included vocals from a singer who had an album titled "Unapologetic"?
→ RAG
Unknown.
→ SearChain
[Query 1] Who released an album titled "Unapologetic"? [Answer 1]: Rihanna released an album titled "Unapologetic". [Query 2] Yes, Rihanna collaborated with Eminem on any songs or albums? [Answer 2]: Yes, Rihanna collaborated with Eminem on the song "Numb", which is on her album "Unapologetic". [Final Context]: The Eminem album that included vocals from a singer who had an album titled "Unapologetic" is the album where they collaborated together, which is "Unapologetic" itself. [Final Answer]: Unapologetic
→ MetaRAG
Round 1 [Answer 1]: Not specified in the given references. [Reason 1]: Since we're looking for an Eminem album that includes vocals from a singer whose album is called 'Unapologetic', the relevant information is that Eminem collaborated with Rihanna on the song 'Numb' from her album 'Unapologetic'. However, the specific Eminem album containing these vocals isn't mentioned in the provided references. Round 2 [Answer 2]: None. [Reason 2]: The reference states that 'Numb' is a song by Rihanna from her album 'Unapologetic'. Since Eminem provided guest vocals on this song, we know that he collaborated with Rihanna on 'Unapologetic'. However, none of the references mention an Eminem album specifically named after another artist's album title. Round 3 [Answer 3]: Unknown without further information. [Reason 3]: From the provided references, we know that 'Numb' is a song by Rihanna featuring Eminem, and this song is part of Rihanna's album 'Unapologetic'. Since Eminem collaborated with Rihanna on this song, we must look into Eminem's discography around the same time period. However, none of the given references mention any specific Eminem album title directly related to the collaboration. We would need more detailed information about Eminem's releases during that time frame. [Final Answer]: Unknown without further information.
→ RetroRAG
Round 1 [Answer 1]: Cannot Determine. [Reason 1]: [Evidence 1]: 1. Eminem and Rihanna had previously collaborated on "I Love the Way You Lie." 2. Eminem and Rihanna had previously collaborated on three other songs before "Numb," all of which were also successful collaborations. 3. Eminem and Rihanna had previously collaborated on three other songs before this one. 4. American rapper Eminem sang vocals on the track "Numb," featured on Rihanna's seventh studio album "Unapologetic." [Answer 2]: Unknown without further information. [Reason 2]: What other Eminem albums feature collaborations with artists whose work includes the title "Unapologetic" specifically focusing on those involving Rihanna. Round 3 [Answer 3]: The Marshall Mathers LP². [Final Answer]: The Marshall Mathers LP²

Figure 14: Additional Case Studies.

C. Assessment using LLM-judge across Various Scenarios

To validate the transferability of our approach across different linguistic contexts, and its reliability in simple question-answering scenarios, we conducted performance evaluations on the knowledge question-answering segments of a Chinese hallucination evaluation dataset HalluQA(Cheng et al. 2023), which contain single-hop question build from Baidu Baike. We utilize the Baidu Baike dump as the document corpus for HalluQA datasets, where articles are segmented into passages of 100 tokens, and we employ the BM25 algorithm to retrieve the top 5 relevant documents to be the external knowledge. Given that the golden truth in HalluQA contains detailed descriptions, making it challenging to quantify performance using token-level metrics, we employed the

GPT4 as the LLM-Judge, with the same setting of (Cheng et al. 2023), to assess the answer semantic accuracy.

Additionally, we simultaneously applied the same LLM-Judge to evaluate the performance on HotpotQA and 2WikiMQA dataset from the semantic perspective.

Table 3: LLM-Judge on three datasets.

	HotpotQA	2WikiMQA	HalluQA
LLM-Judge	LLM-Judge	LLM-Judge	LLM-Judge
Standard	29.4	29.2	27.7
COT	32.0	32.4	37.0
RAG	47.4	38.6	45.2
ReAct	42.3	35.3	36.6
Self-Ask	52.0	39.7	30.4
IR-COT	53.4	42.5	56.8
SearChain	56.8	47.9	48.3
MetaRAG	56.2	40.9	57.8
RetroRAG	68.2	51.2	64.3

As shown in Table 3, our RetroRAG outperforms all baselines in semantic perspective on all three datasets, This indicates that our approach demonstrates more stable and reliable performance across various scenarios. Also, we observed that some baselines with unidirectional forward reasoning paradigm, exhibit performance degradation when encountering simple questions with single-hop. For instance, Self-Ask may degrade to standard prompting, ReAct may degrade to COT, and SearChain may degrade to RAG. This highlights once again the necessities of constructing a retroactive reasoning process.