ActiShade: Activating Overshadowed Knowledge to Guide Multi-Hop Reasoning in Large Language Models

Huipeng Ma[1,2*], Luan Zhang[1*], Dandan Song[1†], Linmei Hu[1†], Yuhang Tian[1], Jun Yang[1], Changzhi Zhou[1], Chenhao Li[1], Yizhou Jin[1], Xudong Li[1], Meng Lin[1], Mingxing Zhang[3], Shuhao Zhang[4]

[1]Beijing Institute of Technology, China
[2]QiYuan Lab
[3]Tsinghua University, China
[4]Huazhong University of Science and Technology, China
{mahuipeng, luan_zhang, sdd, hulinmei}@bit.edu.cn

Abstract

In multi-hop reasoning, multi-round retrieval-augmented generation (RAG) methods typically rely on LLM-generated content as the retrieval query. However, these approaches are inherently vulnerable to *knowledge overshadowing*—a phenomenon where critical information is overshadowed during generation. As a result, the LLM-generated content may be incomplete or inaccurate, leading to irrelevant retrieval and causing error accumulation during the iteration process. To address this challenge, we propose **ActiShade**, which detects and activates overshadowed knowledge to guide large language models (LLMs) in multi-hop reasoning. Specifically, ActiShade iteratively detects the overshadowed keyphrase in the given query, retrieves documents relevant to both the query and the overshadowed keyphrase, and generates a new query based on the retrieved documents to guide the next-round iteration. By supplementing the overshadowed knowledge during the formulation of next-round queries while minimizing the introduction of irrelevant noise, ActiShade reduces the error accumulation caused by *knowledge overshadowing*. Extensive experiments show that ActiShade outperforms existing methods across multiple datasets and LLMs.

Introduction

Large language models (LLMs) have demonstrated remarkable performance across various of NLP tasks, such as multi-hop reasoning (OpenAI 2023; Meta 2024a). However, LLMs have a risk of generating factually incorrect responses, also known as hallucinations (Bang et al. 2023; Ji et al. 2023; Huang et al. 2023). Retrieval-augmented generation (RAG) techniques have been widely adopted to enhance the factual correctness of LLM-generated responses by incorporating knowledge from external resources (Gao et al. 2023; Fan et al. 2024).

Early RAG methods often adopt one-round retrieval, *i.e.*, use the original question as the retrieval query (Guu et al. 2020; Borgeaud et al. 2022; Izacard et al. 2023; Zhang et al. 2023). Although these methods show satisfactory performance in answering single-hop questions (Joshi et al. 2017; Kwiatkowski et al. 2019), they fail in answering multi-hop questions, where more knowledge is needed beyond the one-round retrieved knowledge.

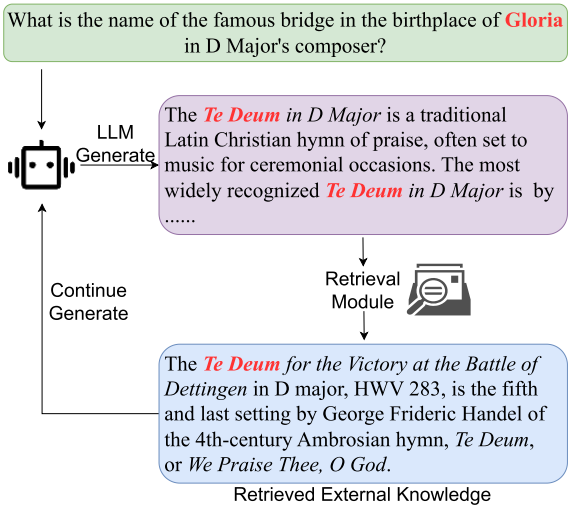

Figure 1: Illustration of error accumulation caused by *knowledge overshadowing*.The keyphrase Gloria in the query is overshadowed, leading the LLM to generate inaccurate content, such as Te Deum. This results in the retrieval of irrelevant documents, which in turn causes LLM to generate more inaccurate content in the next-round iteration.

Recent research proposes multi-round retrieval, which typically relies on the LLM-generated content to guide subsequent-round retrieval. A possible approach uses the response generated by LLMs for retrieval and, in turn, uses the newly retrieved knowledge for generation. By iteratively alternating between retrieval-augmented generation and generation-augmented retrieval, the retrieval and generation are improved (Shao et al. 2023; Trivedi et al. 2023; Jiang et al. 2023; Su et al. 2024). Another approach prompts LLMs to decompose the complex question into a sequence of sub-questions, using the sub-question as the retrieval

*These authors contributed equally.
†Corresponding author.

query to obtain more precise knowledge (Press et al. 2023; Zhou et al. 2023; Cao et al. 2023; Chu et al. 2024).

However, these methods suffer from *knowledge overshadowing* (Zhang et al. 2025) — a phenomenon where dominant conditions can overshadow others, causing the LLM to overlook essential information during generation. As illustrated in Figure 1, given the original query, the dominant condition *Te Deum in D Major* overshadows the condition *Gloria in D Major*, causing the LLM to generate next-round query related to *Te Deum in D Major*. As a result, it retrieves irrelevant documents, which mislead the LLM in generating the subsequent query, ultimately leading to error accumulation over multi-round iterations. This phenomenon is particularly problematic in multi-hop reasoning, where the reasoning process relies on multiple interrelated conditions within the query.

Motivated by this, we propose **ActiShade**, a novel framework designed to detect and subsequently leverage overshadowed knowledge, thereby reducing error accumulation in multi-hop reasoning. ActiShade first detects the overshadowed knowledge within the query, then retrieves documents relevant to it, enabling LLMs to focus on critical but overlooked information during reasoning. Specifically, ActiShade consists of three modules. (i) *Knowledge Overshadowing Detection*: We design a new Gaussian perturbation-based method (**GaP**), to detect overshadowed keyphrases by perturbing the embeddings of candidate keyphrases with Gaussian noise and assessing the changes in the LLM's output distribution. (ii) *Retrieval based on Overshadowed Keyphrase*: We train a dense retriever using our constructed contrastive learning loss. This loss enables the retriever to effectively discriminate among positive, semi-positive, and negative samples, which are categorized based on their relevance to the query and the overshadowed keyphrase. As a result, the retriever achieves improved retrieval of documents relevant to the overshadowed keyphrase while avoiding query-irrelevant noise. (iii) *Query Formulation*: Given the retrieved documents, we prompt the LLM to select the most relevant one and generate a new query that articulates the next reasoning step. In summary, ActiShade supplements the overshadowed knowledge when generating the query for the next-round retrieval, while minimizing the introduction of query-irrelevant noise, thereby reducing the error accumulation caused by *knowledge overshadowing*.

We evaluate ActiShade on three widely used multi-hop reasoning datasets: HotpotQA (Yang et al. 2018), 2WikiMQA (Ho et al. 2020), and MuSiQue (Trivedi et al. 2022). Experimental results show that ActiShade significantly outperforms the state-of-the-art baselines on three datasets. Our contributions can be summarized as follows:

- We propose ActiShade, a novel framework designed to detect and leverage overshadowed knowledge for multi-hop reasoning.
- In ActiShade, we design a new Gaussian perturbation-based method, GaP, to detect the overshadowed knowledge.
- In ActiShade, we introduce a novel contrastive learning loss for retriever training and a query formulation strategy to leverage the overshadowed knowledge.
- We conduct comprehensive experiments on three datasets, demonstrating that ActiShade outperforms the state-of-the-art methods across multiple LLMs in terms of effectiveness.

Related Work

Knowledge Overshadowing

Zhang et al. (2025) observed when extracting knowledge from LLMs using queries involving multiple conditions, some conditions may overshadow others, causing them to be ignored and thus leading to hallucinated outputs-a phenomenon they refer to as *knowledge overshadowing*. This phenomenon is present in multi-hop QA scenarios, which limits the effectiveness of multi-round retrieval approaches. Specifically, *knowledge overshadowing* causes LLMs to generate factually incorrect outputs. As multi-round retrieval methods typically rely on LLM-generated output as the next-round retrieval query, such hallucinations lead to irrelevant retrieval and cause error accumulation during the iterative process. To overcome this limitation, we propose a novel multi-round retrieval framework, ActiShade, which reduces the error accumulation caused by *knowledge overshadowing*. Zhang et al. (2025) also proposed CoDA to detect overshadowed knowledge by removing tokens from the query and measuring changes in the output distribution. In contrast, our GaP preserves the reasoning chain by adding Gaussian noise instead of removing tokens, which enhances the detection of *knowledge overshadowing* in multi-hop reasoning.

Retrieval-augmented LLM

LLMs have a risk of generating hallucinated responses, thus necessitating external retrieval for retrieval-augmented generation. Previous methods typically adopt one-round retrieval, *i.e.*, retrieve knowledge using only the original question once (Guu et al. 2020; Borgeaud et al. 2022; Zhang et al. 2023; Izacard et al. 2023; Shi et al. 2024). This line of work, however, struggles to gather all the necessary knowledge to answer multi-hop questions. Recently, another line of work arose, which adopts multi-round retrieval to meet multi-hop knowledge needs. SelfASK (Press et al. 2023) prompts LLMs to decompose a complex question into sub-questions and answer them through a search engine. IterRetGen (Shao et al. 2023) leverages the output from the previous round concatenated with the question as the query for next-round retrieval. IRCoT (Trivedi et al. 2023) uses each CoT sentence as a query for retrieval until the final answer is obtained. FLARE (Jiang et al. 2023) determine when to retrieve based on reasoning confidence. BeamAggR (Chu et al. 2024) decomposes complex questions, then performs bottom-up multi-source reasoning via post-order traversal, and uses beam aggregation to obtain the final answer. As it relies on multi-source knowledge, which differs from our setting, we do not include it as a baseline for fair comparison. EfficientRAG (Zhuang et al. 2024) iteratively generates new retrieval queries and filters out irrelevant information using small models. DRAGIN (Su et al. 2024) decides when

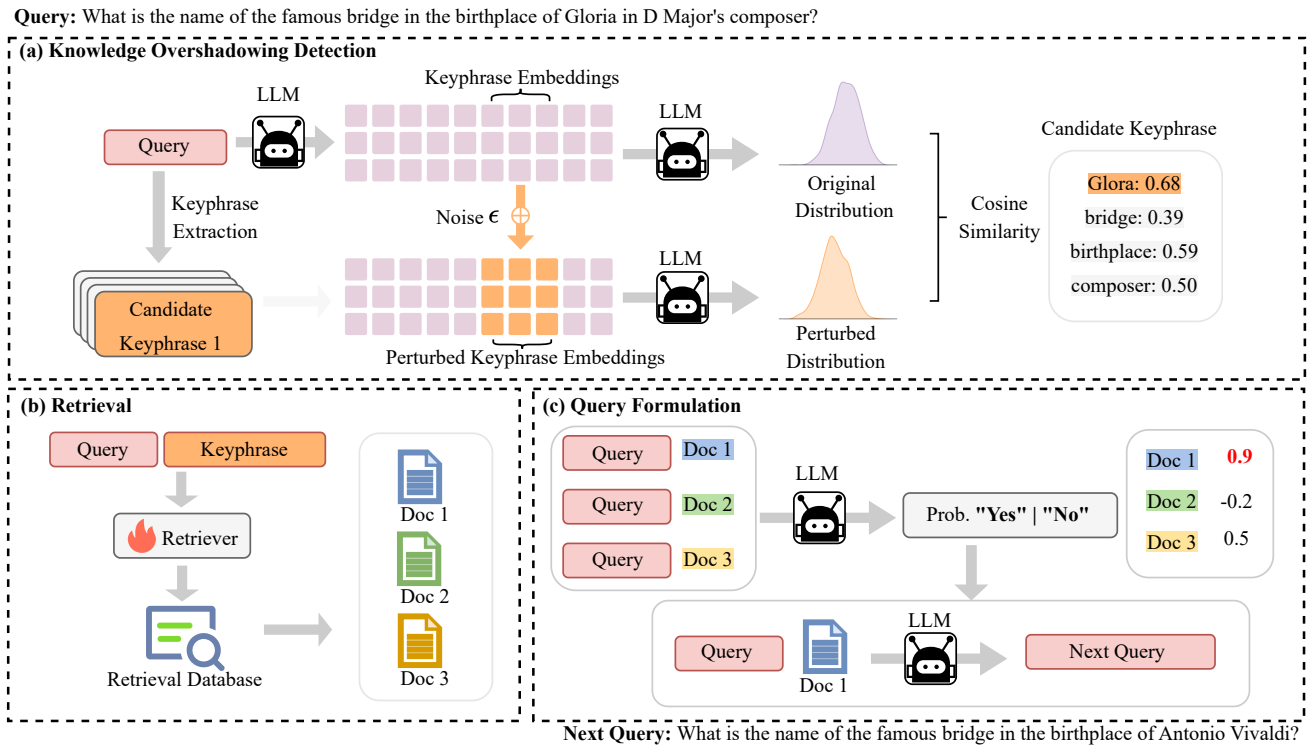

Figure 2: Overview of ActiShade. ActiShade first detects the overshadowed keyphrase in the query, then retrieves relevant documents based on it, and finally formulates a new query for the next-round retrieval.

and what to retrieve based on the LLM's information needs during the generation process.

Compared to them, our method is designed to reduce the error accumulation caused by *knowledge overshadowing* and shows superior performance. Besides, we propose a novel method to detect overshadowed keyphrases through noise perturbation.

ActiShade Framework

In this section, we introduce ActiShade, a novel multi-round retrieval framework that aims to reduce error accumulation caused by *knowledge overshadowing*. ActiShade consists of three modules: (1) *Knowledge Overshadowing Detection* for detecting the overshadowed keyphrase; (2) *Retrieval based on Overshadowed Keyphrase* for relevant document retrieval; and (3) *Query Formulation* for next-round retrieval query generation. An overview of the framework is illustrated in Figure 2.

Knowledge Overshadowing Detection

In this module, we propose a novel method, GaP, to detect *knowledge overshadowing* in the query. The method consists of three steps: keyphrase extraction, keyphrase perturbation, and knowledge overshadowing measuring.

Step 1. Keyphrase Extraction. Given a query Q, we extract a set of candidate keyphrases $P = \{p_1, p_2, ..., p_n\}$, where each p_i is a span within Q. Specifically, we utilize SpaCy (Honnibal 2017) to extract named entities and meaningful tokens with POS tags in the set {`NOUN`, `ADJ`, `VERB`, `PROPN`, `NUM`, `ADV`}, and remove stopwords to reduce noise.

Step 2. Keyphrase Perturbation. For each keyphrase $p_i \in P$, we inject Gaussian noise into its token embeddings while keeping other tokens unchanged. The perturbed input embeddings are computed as:

$$\tilde{\mathbf{H}}_{p_i} = \mathbf{H} + \boldsymbol{m}_{p_i} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $\mathbf{H}$ denotes the original input embeddings of the query, $\boldsymbol{m}_{p_i}$ is a binary mask that takes the value 1 at token positions corresponding to the keyphrase p_i and 0 elsewhere, and $\boldsymbol{\epsilon}$ is Gaussian noise with zero mean and standard deviation σ.

We then input $\tilde{\mathbf{H}}_{p_i}$ into the LLM to generate the perturbed output distribution:

$$\tilde{\mathbf{O}}_{p_i} = \mathbb{P}(y \mid \tilde{\mathbf{H}}_{p_i}), \quad i = 1, 2, ..., n \tag{2}$$

For comparison, the unperturbed output distribution is given by:

$$\mathbf{O} = \mathbb{P}(y \mid \mathbf{H}), \tag{3}$$

where y denotes the LLM's output.

Step 3. Knowledge Overshadowing Measuring. To assess the influence of each keyphrase $p_i \in P$ on the LLM output, we first apply average pooling to the original and perturbed output distributions, $\mathbf{O}$ and $\tilde{\mathbf{O}}_{p_i}$, along the temporal dimension, resulting in pooled representations $\mathbf{r}$ and $\tilde{\mathbf{r}}_{p_i}$.

We then compute the cosine similarity between $\mathbf{r}$ and $\tilde{\mathbf{r}}_{p_i}$, and consider the keyphrase with the highest similarity to be overshadowed:

$$p_{ko} = \arg\max_{p_i \in P} \cos(\mathbf{r}, \tilde{\mathbf{r}}_{p_i}) \tag{4}$$

A high similarity score indicates that the perturbation had minimal influence on the LLM's output, suggesting that the keyphrase is underutilized, i.e., overshadowed.

Our method applies perturbations rather than removing tokens from the query, thereby preserving the structure of the query, which enhances the detection of knowledge overshadowing.

Retrieval based on Overshadowed Keyphrase

Given the detected overshadowed keyphrase p_{ko}, this module retrieves the documents that are relevant to both the query and the overshadowed keyphrase. To enhance the retriever's ability to focus on the overshadowed keyphrase and avoid introducing query-irrelevant noise, we propose a novel contrastive learning loss and train a dense retriever to discriminate three types of documents: positive (relevant to both the query and the keyphrase), semi-positive (relevant to the query but not the keyphrase), and negative (irrelevant to both).

Data Preparation. We construct our training dataset based on the MuSiQue (Trivedi et al. 2022) benchmark. Each data in the MuSiQue dataset is formulated in a dictionary format with the keys `question_decomposition`, `question`, and `paragraphs`. The `paragraphs` field contains a set of documents that are either relevant or irrelevant to the question. The `question_decomposition` field provides a list of sub-questions derived from the original question, each annotated with the supporting document required to answer it, which can be found in the `paragraphs` set.

We first identify the subject entity of the first sub-question and define it as the keyphrase. The supporting document associated with the first sub-question is labeled as the **positive document** (D^+). The supporting documents for other sub-questions are labeled as **semi-positive documents** (D^*), as they are necessary for answering the original question but are not directly related to the keyphrase. All remaining documents are labeled as **negative document** (D^-), which are irrelevant to the question. Due to space limitations, annotation examples can be found in the Appendix.

Loss Function Construction. We extend the contrastive loss proposed by (Izacard et al. 2021) to improve the capability of the retriever to prioritize documents relevant to both the query and a specified phrase within it. The loss function $\mathcal{L}$ is defined as follow:

$$\mathcal{L}_1 = -\log \frac{S(Q,D^+)}{S(Q,D^+)+\sum S(Q,D^*)+\sum S(Q,D^-)}, \tag{5}$$

$$\mathcal{L}_2 = -\log \frac{S(Q,D^+)+\sum S(Q,D^*)}{S(Q,D^+)+\sum S(Q,D^*)+\sum S(Q,D^-)}, \tag{6}$$

$$\mathcal{L} = \alpha\mathcal{L}_1 + (1-\alpha)\mathcal{L}_2 \tag{7}$$

For brevity, we denote $S(Q,D) = e^{\text{sim}(Q,D)}$, where sim indicates the cosine similarity, and introduce hyperparameter α to balance the loss terms. The loss $\mathcal{L}_1$ encourages positive pairs to have higher scores over both semi-positive and negative pairs. Although semi-positive documents D^* are not directly relevant to any phrase in the question, they are required to answer the question and thus are more important than negative documents D^-. We introduce loss $\mathcal{L}_2$ to further distinguish semi-positive documents from negative ones. The combined loss $\mathcal{L}$ ensures the retriever ranks documents in the desired order: $D^+ > D^* > D^-$.

Retrieval. We concatenate the query Q with its corresponding overshadowed keyphrase p_{ko} as input to retrieve a set of relevant documents $RD = \{rd_1, rd_2, ..., rd_n\}$. The trained retriever is capable of retrieving documents relevant to both the query and the overshadowed keyphrase, ensuring that the retrieved documents are not only query-relevant but also supplement the overshadowed knowledge, thereby enhancing LLMs' reasoning.

Query Formulation

The previous module returns a set of retrieved documents RD, which is relevant to both the query and the overshadowed keyphrase within it. This module then formulate a new query based on the retrieved documents for the subsequent retrieval round. The query formulation process consists of three steps: relevant document selection, query generation, and subsequent-round retrieval decision.

Step 1. Relevant Document Selection. Given a collection of retrieved documents RD, we first prompt the LLM to select the most relevant one rd_m. Specifically, each retrieved document and the query are jointly input into the LLM. The LLM is required to determine whether the retrieved document is relevant to the query. If it is relevant, output "Yes"; otherwise, output "No". A higher probability assigned to "Yes" suggests a higher degree of relevance. The most relevant retrieved document is then selected based on the probability of outputting "Yes". The prompt template used for this step is detailed in the Appendix.

Step 2. Query Generation. In the second step, we prompt the LLM to generate a new query Q_{next} based on the most relevant retrieved document rd_m. The newly generated query is used for the subsequent retrieval round, aiming to retrieve more information beyond the scope of the initial query. The prompt template used for this step is detailed in Appendix. Figure 2 presents examples of query generation.

Since the retrieved document serves to supplement the overshadowed knowledge, it enable the generation of a more accurate query. Moreover, the newly generated query explicitly presents implicit reasoning results. These allow the new query to lead to more accurate and relevant retrieval in the next round, thereby reducing the error accumulation caused by *knowledge overshadowing*.

Model	Method	MuSiQue		HotpotQA		2WikiMQA	
		ACC	F1	ACC	F1	ACC	F1
Llama-3-8B-Instruct	**Direct**	5.60	9.96	22.40	25.34	26.60	31.25
	CoT	11.65	16.29	29.00	34.09	27.60	34.19
	Direct-R$^{\heartsuit}$	11.42	16.06	37.7	44.89	28.37	35.56
	Iter-RetGen$^{\clubsuit}$	18.24	20.59	48.23	49.41	38.71	44.56
	IRCoT$^{\clubsuit}$	15.57	18.32	40.10	47.03	34.20	41.01
	SelfASK$^{\clubsuit}$	20.60	21.41	47.10	48.70	39.50	43.87
	FLARE$^{\clubsuit}$	19.74	20.50	48.45	50.40	41.35	42.24
	DRAGIN$^{\clubsuit}$	21.11	22.61	50.87	52.52	40.78	42.31
	ActiShade (Ours)	**25.25**	**26.94**	**54.60**	**56.33**	**45.80**	**46.02**
Qwen2.5-7B-Instruct	**Direct**	3.80	11.09	19.40	19.52	26.80	29.95
	CoT	6.00	13.93	22.00	27.61	29.00	32.24
	Direct-R$^{\heartsuit}$	11.60	17.24	43.00	47.99	38.60	41.17
	Iter-RetGen$^{\clubsuit}$	15.40	18.07	44.40	48.24	41.20	42.98
	IRCoT$^{\clubsuit}$	14.90	18.01	43.79	48.13	40.21	40.84
	SelfASK$^{\clubsuit}$	17.35	20.60	42.18	46.10	43.17	44.30
	FLARE$^{\clubsuit}$	10.69	14.89	40.19	42.03	39.21	40.80
	DRAGIN$^{\clubsuit}$	19.80	22.01	46.10	50.30	45.90	45.87
	ActiShade (Ours)	**22.80**	**26.11**	**48.20**	**55.45**	**52.80**	**50.47**
Qwen2.5-14B-Instruct	**Direct**	6.20	13.48	27.20	32.94	29.00	32.39
	CoT	10.40	17.74	29.40	34.92	33.40	35.76
	Direct-R$^{\heartsuit}$	14.80	19.16	46.20	48.11	39.00	43.14
	Iter-RetGen$^{\clubsuit}$	17.20	21.54	49.03	51.04	43.20	45.19
	IRCoT$^{\clubsuit}$	15.89	19.90	47.98	50.50	44.60	45.71
	SelfASK$^{\clubsuit}$	18.48	21.75	45.97	47.19	47.49	49.13
	FLARE$^{\clubsuit}$	13.10	18.00	42.14	44.87	40.01	40.74
	DRAGIN$^{\clubsuit}$	22.70	24.11	51.21	54.30	48.10	49.87
	ActiShade (Ours)	**25.59**	**27.47**	**53.97**	**57.45**	**51.13**	**53.29**

Table 1: The overall experimental results of ActiShade and other baselines on three benchmarks. The best results are in bold. $\heartsuit$ donates single-round retrieval. $\clubsuit$ indicates multi-round retrieval.

Step 3. Subsequent-Round Retrieval Decision. To decide whether to terminate the iterative process, we prompt the LLM to assess whether more information is needed to answer the initial query. Specifically, the new query Q_{next} is input into the LLM, which is required to determine whether it is a single-hop query. If it is, we perform an additional retrieval round and then terminate; otherwise, the retrieval continues iteratively. The iterative process also terminates if the maximum number of iterations is reached. The prompt template used for this step is detailed in the Appendix.

After iteration, we input the initial query and the relevant documents retrieved during the iterative process into the LLM to obtain the final response.

Experimental Setup

Datasets

We evaluate ActiShade on three multi-hop reasoning datasets. HotpotQA (Yang et al. 2018), 2WikiMQA (Ho et al. 2020) consist of two-hop questions, and MusiQue (Trivedi et al. 2022) contains questions with 2 to 4 hops. For HotpotQA, 2WikiMQA, and MuSiQue, we use the same test set provided by IRCoT (Trivedi et al. 2023), which contains 500 randomly sampled instances from the original development set.

Implementation Details

We use Llama-3-8B-Instruct (Meta 2024b) and Qwen2.5-Instruct (7B and 14B) (Yang et al. 2024) as the backbone language models for generation and reasoning. For retriever training, we fine-tune `contriever-msmarco` (Izacard et al. 2021) on a subset of the MuSiQue training set. We manually select 5,000 high-quality examples, of which 3,500 are used for training, 750 for validation, and 750 for testing. The retriever is trained using the AdamW optimizer (learning rate 5e-5, batch size 32) for up to 20 epochs with early stopping based on validation loss. The contrastive loss combines two objectives with a weighting coefficient $\alpha = 0.7$. All experiments are conducted on two NVIDIA A6000 GPUs.

Baselines

We choose the following methods as baselines. **Standard Prompting (Brown et al. 2020)** directly generates the final answer. **CoT Prompting (Wei et al. 2022)** generates reasoning steps before the final answer. **One-time Retrieval** retrieves relevant documents from external resources and incorporates them to generate the final answer. **IRCoT (Trivedi et al. 2023)** alternates between retrieval-augmented reasoning and reasoning-augmented retrieval until enough information is obtained to answer the given ques-

Model / Dataset	MuSiQue	HotpotQA	2WikiMQA
Direct-R	16.06	44.89	35.56
-*w* CoDA	15.86	45.98	35.49
-*w* GaP	**17.83**	**46.81**	**38.67**
ActiShade-NoKOD	22.83	51.23	45.18
-*w* CoDA	21.23	52.45	41.29
-*w* GaP	**26.94**	**56.33**	**46.02**

Table 2: Performance comparison between GaP and CoDA in single-round and multi-round retrieval settings. ActiShade-NoKOD denotes ActiShade without the Knowledge Overshadowing Detection module. All results are reported in F1 score.

tion. **Iter-RetGen (Shao et al. 2023)** synergizes retrieval and generation in an iterative manner: the LLM's response serves as a query for retrieving more relevant knowledge, which in turn helps generate a better response in another iteration. **Self-Ask (Press et al. 2023)** iteratively decomposes complex questions into sub-questions, retrieves and answers them to reach the final answer. **FLARE (Jiang et al. 2023)** dynamically adjusts retrieval timing based on reasoning confidence and retrieves guided by the upcoming reasoning sentences. **DRAGIN (Su et al. 2024)** detects information needs in real time and uses self-attention over context to form retrieval queries during the generation process.

Evaluation Metrics

For evaluation metrics, we utilize Accuracy (Acc) and F1 score metrics for evaluation. The ACC checks if the ground-truth answer is in the LLM-generated answer, which is also named Cover Exact Match. The F1 score is used to measure the overlap between the LLM-generated answer and the ground truth answer.

Experimental Results

Main Results

The experimental results on three multi-hop reasoning datasets are presented in Table 1. We can obtain the following observations:

Achieving Significant Performance Improvement across all datasets and LLMs. ActiShade outperforms the previous state-of-the-art, DRAGIN, across all datasets and LLMs, highlighting its effectiveness in multi-hop reasoning. This performance improvement can be attributed to ActiShade's ability to reduce error accumulation caused by *knowledge overshadowing* by iteratively detecting overshadowed keyphrases in the query, retrieving documents relevant to both the query and the overshadowed keyphrase, and generating a new query based on the retrieved documents. Notably, ActiShade surpasses SelfASK (Press et al. 2023), which decomposes a complex question into sub-questions and answers them via retrieval. We believe this suggests that our query formulation process makes implicit reasoning explicit, enabling more accurate and relevant retrieval compared to question decomposition.

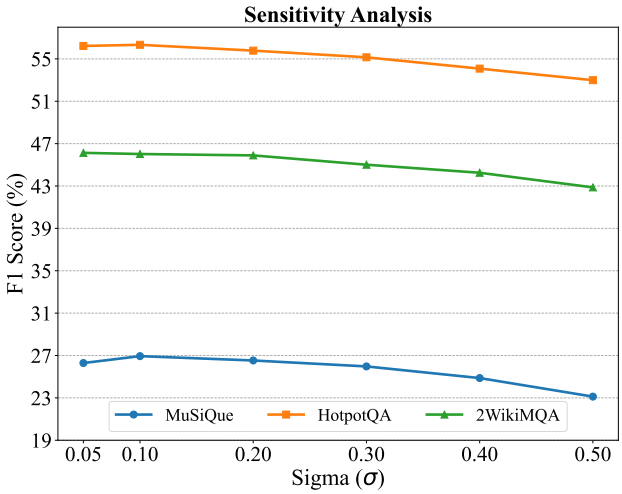

Figure 3: Sensitivity analysis of the Gaussian noise standard deviation σ.

Maintaining Generalization Ability. We train our retriever based on the MuSiQue dataset, as detailed in the ActiShade Framework section. However, ActiShade, on HotpotQA and 2WikiMQA, still outperforms all baselines across all LLMs, further demonstrating its effectiveness and generalization. This indicates that the retriever effectively learns to align retrieval not only with the query but also with the overshadowed keyphrase, allowing it to generalize well across various multi-hop reasoning benchmarks.

Effectiveness for larger models. To evaluate how effective ActiShade is at different model sizes, we conduct experiments on Qwen2.5-Instruct (7B and 14B). As shown in Table 1, the ActiShade's performance generally improves with the model size, demonstrating its scalability to larger models. Due to hardware resource constraints, we are unable to implement ActiShade on larger models.

Analysis of Knowledge Overshadowing Detection

To systematically evaluate the effectiveness of GaP, we conduct a series of analyses focusing on performance comparison, interpretability, and parameter sensitivity.

Comparative Performance of Detection Methods. To investigate the impact of the knowledge overshadowing detection module, we compare our proposed GaP against the CoDA (Zhang et al. 2025) method under both single-round and multi-round retrieval settings. In the single-round setting, we evaluate three variants: (1) Direct retrieval without overshadowing detection; (2) Direct retrieval with the CoDA integrated; (3) Direct retrieval with our GaP integrated. In the multi-round setting, we compare three corresponding setups: (1) ActiShade without the GaP method; (3) ActiShade replacing GaP with the CoDA method; (3) ActiShade (the full pipeline). The experimental results are shown in Table 2. We observe that, in both single-round and multi-round retrieval settings, models incorporating our GaP method consistently outperform those without such integration, high-

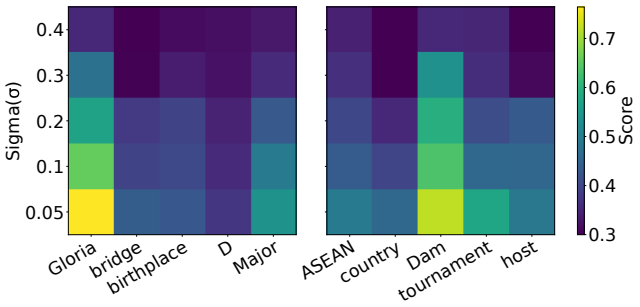

Figure 4: Visualization analysis of the Gaussian noise standard deviation σ.

lighting the effectiveness of GaP. In addition, we observe that, on the MuSiQue and 2WikiMQA datasets, models using CoDA even perform worse than those without knowledge overshadowing detection. This indicates that CoDA's token-removing approach may disrupt the reasoning chain in multi-hop questions, thereby limiting its effectiveness.

Sensitivity and Interpretability Analysis of GaP. We conduct a sensitivity analysis to investigate how varying the standard deviation σ of the Gaussian noise used for keyphrase perturbation affects the performance of our proposed ActiShade. All experiments in this analysis are conducted using the Llama-3-8B-Instruct, and the results are evaluated based on the F1 metric. As shown in Figure 3, we vary σ in the range of [0.05, 0.5] and observe its impact on the final performance. Experimental results show that as σ decreases, the model performance first improves, reaching a peak at $\sigma = 0.1$, and then gradually declines. This indicates that a moderate level of noise can effectively help detect overshadowed keyphrases, while excessive noise causes large output distribution shifts for all candidate keyphrases, reducing the effectiveness of detection. Nevertheless, the overall performance remains relatively stable across a wide range of σ values, suggesting that our method exhibits low sensitivity to this hyperparameter.

To interpret how GaP detects overshadowed keyphrases, we also conduct a visualization analysis of output distribution similarity across different keyphrases. We randomly select two queries and apply Gaussian noise of varying standard deviation to each candidate keyphrase. For each combination of keyphrase and noise level σ, we compute the similarity between the model's output distributions before and after perturbation. A high similarity suggests the keyphrase has little influence on the output and is likely overshadowed. Figure 4 shows that moderate perturbation best separates salient from overshadowed keyphrases, while stronger noise disrupts all outputs, lowers similarities across the board, and weakens detection.

Analysis of Retriever Training

We analyze the effectiveness of retriever training in the ActiShade framework by evaluating both the retrieval capability and the downstream QA performance under different training strategies.

Model	Pos		Semi		Pos&Semi	
	R@1	R@3	R@1	R@3	R@1	R@3
Base	29.20	50.40	12.57	25.42	18.29	36.78
SCL	57.84	69.21	40.12	59.99	**38.21**	50.29
FCL	**75.33**	**84.80**	**43.21**	**61.42**	38.14	**52.72**

Table 3: Comparison of Recall@1 and Recall@3 for different retrievers.

Model / Dataset	MuSiQue	HotpotQA	2WikiMQA
ActiShade	**26.94**	**56.33**	**46.02**
-w/ SCL	24.10	54.25	44.97
-w/o FCL	25.68	53.89	44.61

Table 4: Evaluation of retriever training strategies in ActiShade. The performance is evaluated using the F1 score.

We first assess the retrieval ability of three retriever variants: (1) a retriever without task-specific fine-tuning (Base), (2) our proposed retriever trained with fine-grained contrastive learning (FCL), and (3) a retriever trained using standard contrastive learning (SCL) that only distinguishes between positive and negative examples. As shown in Table 3, our method achieves the highest Recall@k scores on both positive and semi-positive document retrieval, demonstrating its effectiveness in capturing multi-level document relevance critical for multi-hop reasoning. When distinguishing between positive&semi-positive and negative examples, our method performs comparably to the retriever trained with standard contrastive learning on Recall@1, while outperforming it on Recall@3. This indicates that our improved contrastive learning objective helps the retriever better distinguish positive from semi-positive examples, while still effectively discriminating negative ones.

We then examine how these retrievers affect the final QA performance. As presented in Table 4, our proposed retriever achieves the best results across three datasets. This shows that training the retriever to distinguish varying degrees of relevance is beneficial not only for retrieval capability but also for downstream answer generation. Notably, even without retriever training, ActiShade still outperforms previous baselines, highlighting the effectiveness of the Knowledge Overshadowing Detection and Query Formulation modules in the overall framework.

Conclusion

In this paper, we introduce ActiShade, a novel multi-round retrieval framework for multi-hop reasoning. ActiShade iteratively detects overshadowed keyphrases in the query, retrieves documents relevant to both the query and the overshadowed keyphrase, and generates a new query based on the retrieved documents for the next iteration, thereby reducing the error accumulation caused by *knowledge overshadowing*. Extensive experiments demonstrate the effectiveness of ActiShade across multiple datasets and LLMs.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0210800) and the National Natural Science Foundation of China (Grant No. 62476025).

References

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 675–718. Nusa Dua, Bali: Association for Computational Linguistics.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; van den Driessche, G.; Lespiau, J.; Damoc, B.; Clark, A.; de Las Casas, D.; Guy, A.; Menick, J.; Ring, R.; Hennigan, T.; Huang, S.; Maggiore, L.; Jones, C.; Cassirer, A.; Brock, A.; Paganini, M.; Irving, G.; Vinyals, O.; Osindero, S.; Simonyan, K.; Rae, J. W.; Elsen, E.; and Sifre, L. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 2206–2240. PMLR.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Amanda, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Tom, H.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Benjamin, C.; Clark, J.; Berner, C.; Sam, M.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *arXiv: Computation and Language,arXiv: Computation and Language*.

Cao, S.; Zhang, J.; Shi, J.; Lv, X.; Yao, Z.; Tian, Q.; Hou, L.; and Li, J. 2023. Probabilistic Tree-of-thought Reasoning for Answering Knowledge-intensive Complex Questions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 12541–12560. Association for Computational Linguistics.

Chu, Z.; Chen, J.; Chen, Q.; Wang, H.; Zhu, K.; Du, X.; Yu, W.; Liu, M.; and Qin, B. 2024. BeamAggR: Beam Aggregation Reasoning over Multi-source Knowledge for Multihop Question Answering. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 1229–1248. Association for Computational Linguistics.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.; and Li, Q. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In Baeza-Yates, R.; and Bonchi, F., eds., *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, 6491–6501. ACM.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; and Wang, H. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR*, abs/2312.10997.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 3929–3938. PMLR.

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.

Honnibal, M. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *(No Title)*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *CoRR*, abs/2311.05232.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Izacard, G.; Lewis, P. S. H.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *J. Mach. Learn. Res.*, 24: 251:1–251:43.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12): 248:1–248:38.

Jiang, Z.; Xu, F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; and Neubig, G. 2023. Active Retrieval Augmented Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7969–7992. Singapore: Association for Computational Linguistics.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1601–1611. Association for Computational Linguistics.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A. P.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguistics*, 7: 452–466.

Meta, A. 2024a. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Meta, A. 2024b. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711. Singapore: Association for Computational Linguistics.

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 Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9248–9274. Singapore: Association for Computational Linguistics.

Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and Yih, W. 2024. REPLUG: Retrieval-Augmented Black-Box Language Models. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 8371–8384. Association for Computational Linguistics.

Su, W.; Tang, Y.; Ai, Q.; Wu, Z.; and Liu, Y. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12991–13013. Bangkok, Thailand: Association for Computational Linguistics.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 539–554.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. 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)*, 10014–10037. Toronto, Canada: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zhang, Y.; Li, S.; Qian, C.; Liu, J.; Yu, P.; Han, C.; Fung, Y. R.; McKeown, K.; Zhai, C.; Li, M.; et al. 2025. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. *arXiv preprint arXiv:2502.16143*.

Zhang, Z.; Zhang, X.; Ren, Y.; Shi, S.; Han, M.; Wu, Y.; Lai, R.; and Cao, Z. 2023. IAG: Induction-Augmented Generation Framework for Answering Reasoning Questions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 1–14. Association for Computational Linguistics.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhuang, Z.; Zhang, Z.; Cheng, S.; Yang, F.; Liu, J.; Huang, S.; Lin, Q.; Rajmohan, S.; Zhang, D.; and Zhang, Q. 2024. EfficientRAG: Efficient Retriever for Multi-Hop Question Answering. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3392–3411. Miami, Florida, USA: Association for Computational Linguistics.