DR-RAG: Applying Dynamic Document Relevance to Retrieval-Augmented Generation for Question-Answering

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has recently demonstrated the performance of Large Language Models (LLMs) in the knowledge-intensive tasks such as Question-Answering (QA). RAG expands the query context by incorporating external knowledge bases to enhance the response accuracy. However, it would be inefficient to access LLMs multiple times for each query and unreliable to retrieve all the relevant documents by a single query. We have found that even though there is low relevance between some critical documents and query, it is possible to retrieve the remaining documents by combining parts of the documents with the query. To mine the relevance, a two-stage retrieval framework called **D**ynamic-**R**elevant **R**etrieval-**A**ugmented **G**eneration (**DR-RAG**) is proposed to improve document retrieval recall and the accuracy of answers while maintaining efficiency. Additionally, a compact classifier is applied to two different selection strategies to determine the contribution of the retrieved documents to answering the query and retrieve the relatively relevant documents. Meanwhile, DR-RAG call the LLMs only once, which significantly improves the efficiency of the experiment. The experimental results on multi-hop QA datasets show that DR-RAG can significantly improve the accuracy of the answers and achieve new progress in QA systems.

1 Introduction

Large language models (LLMs) have recently made significant improvement in the field of Natural Language Processing (NLP), especially in text generation tasks (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023b; Anil et al., 2023; Ouyang et al., 2022; Touvron et al., 2023a). Although LLMs excel in various application scenarios, challenges remain regarding the accuracy and timeliness of the generated text, especially in real-time domains. LLMs with intrinsic parameter memories may generate inaccurate or even incorrect

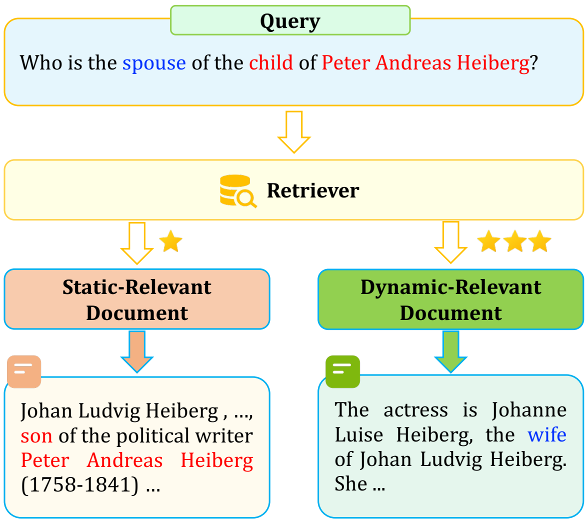

Figure 1: An example shows that retriever easily introduces static-relevant documents due to high relevance (red), but struggles to retrieve dynamic-relevant documents which are of low relevance (blue) but critical for the answer. Stars are levels of retrieval difficulty.

text when faced with up-to-date query (Min et al., 2023; Mallen et al., 2022; Muhlgay et al., 2023). This issue, known as hallucination, occurs when the text generated by LLMs fails to align with real-world knowledge (Ji et al., 2023; Zhang et al., 2023; Kwiatkowski et al., 2019). Therefore, Retrieval-Augmented Generation (RAG) frameworks have been proposed to improve the accuracy of generated text by combining relevant information from external knowledge base with query (Arora et al., 2023; Lewis et al., 2020; Borgeaud et al., 2022). RAG has effectively demonstrated its superiority in knowledge-intensive tasks such as open-domain Question-Answering (QA) and has achieved new progress in the LLMs' performance.

However, irrelevant information reduces the quality of the generated text and further interferes with the ability of LLMs to answer the query in the application (Shi et al., 2023). Moreover, the undifferentiated combining strategy in RAG can lead to mixing in some irrelevant information (Rony et al., 2022). Inconsistent or contradictory information during combining the document may lead to the introduction of incorrect information and have an impact on the accuracy of the generated answers. In the retrieval, we need to select documents that are highly relevant and decisive for the generation of answers (*static-relevant documents*) and documents that are low relevant but also crucial to the generation of answers (*dynamic-relevant documents*). As shown in Fig. 1, an example query is 'Who is the spouse of the child of Peter Andreas Heiberg?', which requires the two most relevant documents to obtain the correct answers. Static-relevant documents is easy to be retrieved due to the high relevance with the query on 'Peter Andreas Heiberg' and 'child/son' (Fig. 1 red). However, dynamic-relevant documents is difficult to be retrieved because it is only related to the query as a 'spouse/wife' (Fig. 1 blue). Moreover, the knowledge base contains too much information about 'spouse', which may cause dynamic-relevant documents to be ranked lower in the retrieval process. There is a high relevance on 'Johan Ludvig Heiberg' and 'wife' between static- and dynamic-relevant documents. If 'spouse/wife' with the query is also taken into account, we can easily retrieve dynamic-relevant documents to get the answer.

Motivated by the above observations, a novel two-stage retrieval framework called **D**ynamic-**R**elevant **R**etrieval-**A**ugmented **G**eneration (**DR-RAG**) is proposed to mine the relevance between the query and documents. In the first-retrieval stage, similarity matching (SM) method is used to obtain a certain percentage of documents based on the query. Subsequently, the documents with the query are concatenated to dig further into more in-depth relevance to dynamic-relevant documents. Moreover, we design a classifier that determines whether the retrieved documents contribute to the current query by a predefined threshold. To optimise the documents, we design two approaches, i.e., forward selection and reverse selection. We aim to ensure that the retrieved documents are highly relevant, thus avoiding redundant documents. Through two-stage retrieval and classifier selection strategies, DR-RAG has the ability to retrieve sufficient relevant documents and address complex and multilevel problems. DR-RAG can make full use of the static and dynamic relevance of documents and enhance the model's performance under diverse queries. To validate the effectiveness of DR-RAG, we conduct extensive experiments by different retrieval strategies on multi-hop QA datasets. The results show that our method can significantly improve the recall and accuracy of the answers.

In short, we summarize the key contributions of this work as follows:

- We design an effective RAG framework named DR-RAG, which is effective in multi-hop QA. Two-stage retrieval strategy is proposed to significantly improve the recall and accuracy of the retrieval results.
- We design a classifier that determines whether the retrieved documents contribute to the current query by setting a predefined threshold. The mechanism can effectively reduces redundant documents and ensures that the retrieved documents are concise and efficient.
- We conduct experiments on three multi-hop QA datasets to validate our DR-RAG. The experimental results show that DR-RAG has the ability to improve recall by 86.75% and improve by 6.17%, 7.34%, 9.36% in the three metrics (Acc, EM, F1). DR-RAG has significant advantages in complex and multi-hop QA and support the performance of the RAG frameworks in QA systems.

2 Method

In this section, we will describe the DR-RAG framework and its design approach in detail. Specifically, in section 2.1 we will define relevant symbols comprehensively, and in section 2.2 we will describe the whole framework.

2.1 Preliminaries

To enrich the knowledge of LLMs, we need to retrieve multiple documents to provide comprehensive answers to complex query. For better clarity, we summarize the key notations in Table 1 and the whole framework can be referred to in Fig. 2.

Our goal is to retrieve the most relevant documents $\boldsymbol{d}^*$ from the retrieved documents $\boldsymbol{d}$ to answer the query and prevent missing key information from the additional knowledge provided to LLMs. However, it is difficult to retrieve all the static- and dynamic-relevant documents through SM method during the retrieval process (Fig. 2). For clearness, we name these two types of relevant documents as $\boldsymbol{d}^*_{stat}$ and $\boldsymbol{d}^*_{dyn}$, respectively.

Table 1: The key mathematical notations.

Notation	Description
$\boldsymbol{q}$	the user's input query
$\boldsymbol{q}^*$	the query combined with retrieved document after the first-retrieval stage
D	the knowledge base for storing documents
C	the trained classifier
k	the total number of documents retrieved from D
k_1	the total number of documents retrieved from D in the first-retrieval stage
k_2	the total number of documents retrieved from D in the second-retrieval stage
n	the number of documents critical for correctly answering $\boldsymbol{q}$
$\boldsymbol{d}$	the documents retrieved from D
$\boldsymbol{d}^*$	the relevant documents for correctly answering $\boldsymbol{q}$
$\boldsymbol{d}^{\triangle}$	the irrelevant documents for correctly answering $\boldsymbol{q}$
$\boldsymbol{d}^*_{\boldsymbol{stat}}$	the documents with static relevance to the query
$\boldsymbol{d}^*_{\boldsymbol{dyn}}$	the documents with dynamic relevance to the query

A common approach is to increase the value of k to expand the possibility of retrieving $\boldsymbol{d}^*_{\boldsymbol{dyn}}$. For instance, in MuSiQue, increasing k from 3 to 6 only raises the recall rate from 58% to 76%, leaving many relevant documents unretrieved. Furthermore, irrelevant documents will provide LLMs with redundant information. Motivated by the problem, the main research objective of our work is to improve the document recall rate of $\boldsymbol{d}^*_{\boldsymbol{dyn}}$ based on dynamic relevance with the same top-k.

2.2 DR-RAG

In this section, we will give a comprehensive description about the DR-RAG framework, a new two-stage retrieval method compared to traditional reranking methods (NetEase Youdao, 2023; Chen et al., 2024). From Fig. 2, we retrieve k_1 documents through SM method (first-retrieval stage) and employ a classifier C to model the dynamic relevance between documents (selection process) to enhance the recall rate of the remaining k_2 documents. The classifier C lies in assessing the dynamic relevance between documents to determine whether the information from the documents is crucial to answer the present query.

2.2.1 Query Documents Concatenation

As mentioned before, due to the low relevancy between dynamic-relevant documents and the query, the documents are difficult to be retrieved. Moreover, the only relevant information 'spouse/wife' between them is also obscured by the mixed information in the knowledge base because too many documents in D will contain 'spouse'. Therefore, Query Documents Concatenation (QDC) method aims to employ the sentence to match for more useful and relevant information. After the first-retrieval stage, we will obtain k_1 static-relevant documents and concatenate $\boldsymbol{q}$ with each document to form multiple $<\boldsymbol{q}, \boldsymbol{d_i}, i \in k_1>$ pairs. Moreover, dynamic-relevant documents from D can be retrieved by corresponding $<\boldsymbol{q}, \boldsymbol{d_i}, i \in k_1>$ pair in the second-retrieval stage. As the case in Fig 2, when $\boldsymbol{q}$ and $\boldsymbol{d}^*_{\boldsymbol{stat}}$ are concatenated, the query contains both the 'Johan Ludvig Heiberg' and the relationship 'spouse/wife', which is essentially similar to $\boldsymbol{d}^*_{\boldsymbol{dyn}}$. Therefore, $\boldsymbol{d}^*_{\boldsymbol{dyn}}$ is more clearly related to the query and thus easily retrieved. The whole process is:

$$\begin{aligned}
Cnt &= \{\} \\
\{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{k_1}\} &= \texttt{Retriever}(\boldsymbol{q}) \\
Cnt &= Cnt \cup \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{k_1}\} \\
\boldsymbol{q}_i^* &= \texttt{Concat}(\boldsymbol{q}, \boldsymbol{d}_i) \\
\{\boldsymbol{d}'_{i,1}, \ldots, \boldsymbol{d}'_{i,k_2}\} &= \texttt{Retriever}(\boldsymbol{q}_i^*) \\
Cnt &= Cnt \cup \{\boldsymbol{d}'_{i,j} \mid \boldsymbol{d}'_{i,j} \notin Cnt \wedge first\} \\
answer &= \texttt{LLM}(\texttt{Concat}(\boldsymbol{q}, Cnt))
\end{aligned} \tag{1}$$

where $k_1 + k_2$ is equal to k. `Retriever` is a common SM method. $\boldsymbol{d}$ and $\boldsymbol{d'}$ are the relevant document retrieved from D in the first and second-retrieval stage. Cnt is a context containing multiple documents. $Cnt = Cnt \cup \{\boldsymbol{d}'_{i,j} \mid \boldsymbol{d}'_{i,j} \notin Cnt \wedge first\}$ means that for for a given $\boldsymbol{d'}$, the first $\boldsymbol{d}'_{i,j}$ in the second-retrieval stage that is not already part of Cnt will be placed into Cnt. `LLM` is a large language model to obtain the answer. $answer$ is the output to answer the query.

2.2.2 Classifier for Selection

While QDC method significantly improves document recall and answer accuracy, there are two key issues to consider: 1) There may be redundant information in the k retrieved documents, which

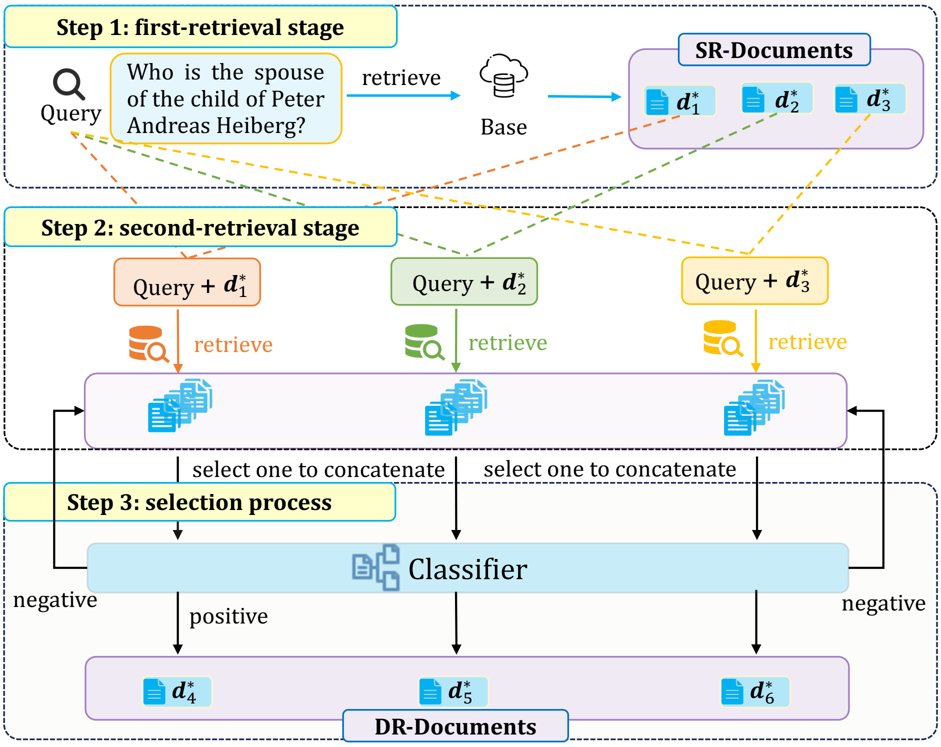

Figure 2: An overview of DR-RAG. In step 1, we retrieve static-relevant documents (SR-Documents) due to high relevance with the query. Then we concatenate SR-Documents with the query to retrieve multiple dynamic-relevant documents (DR-Documents) in step 2. Finally, we select each of DR-Documents in turn to concatenate with the query and SR-Documents and feed them into the classifier to select the most relevant DR-Document.

may affect the response of LLMs; 2) How to determine whether a document retrieved in the second-retrieval stage is valid for an answer to further optimise document recall. Motivated by the issues, two pipelines are designed to dig into in-depth document relevance and solve the issues: 1) Classifier Inverse Selection (CIS): in this pipeline, after the second-retrieval stage we exclude some irrelevant documents from the k retrieved documents; 2) Classifier Forward Selection (CFS) : we set a judgment condition to each retrieved document in the second-retrieval stage to filter out irrelevant documents which are useless or even play a negative role in the answer. In addition, we will train a classifier C by a small model with millisecond-level runtime to prevent excessive delays in our pipelines. DR-RAG involves a small binary-classification model where the input consists of $\boldsymbol{q}$ and two documents. The training objective is to determine the potential contribution of the documents to answering $\boldsymbol{q}$. The specific settings are as follows:

$$\begin{aligned}
\mathrm{C}(\boldsymbol{q}, \boldsymbol{d}^*, \boldsymbol{d}^*) &= positive \\
\mathrm{C}(\boldsymbol{q}, \boldsymbol{d}^*, \boldsymbol{d}^{\triangle}) &= negative \\
\mathrm{C}(\boldsymbol{q}, \boldsymbol{d}^{\triangle}, \boldsymbol{d}^{\triangle}) &= negative
\end{aligned} \tag{2}$$

where C represents the classifier. $positive$ and $negative$ indicate whether the two documents are critical for the query.

Classifier Inverse Selection In this approach, we selectively exclude some irrelevant documents from the retrieved k documents to minimize document redundancy. Specifically, after obtaining k documents in stages, we pair them as <$\boldsymbol{q}, \boldsymbol{d_m}, \boldsymbol{d_n}$> and get C_k^2 pairs. The pairs, with the current query $\boldsymbol{q}$, are collectively fed into the classifier C. Similarly, when the classification result of a document and the remaining k-1 documents is negative, then we consider the document as redundant and should be removed. The whole process is:

$$\begin{aligned}
&Cnt = Cnt \cup \{\boldsymbol{d}'_{i,j} \mid \boldsymbol{d}'_{i,j} \notin Cnt \wedge first\} \\
&P_{i,j} = \begin{cases} 1 & \text{if } \exists i, C(\boldsymbol{q}, \boldsymbol{d'}_{i,j}, \boldsymbol{d}_i) = positive \\ 0 & otherwise \end{cases} \\
&Cnt = Cnt - \{\boldsymbol{d'}_{i,j} \mid P_{i,j} = 0\} \\
&answer = \mathsf{LLM}(\mathsf{Concat}(\boldsymbol{q}, Cnt))
\end{aligned} \tag{3}$$

where - represents complement. $Cnt = Cnt - \{\boldsymbol{d'}_{i,j} \mid P_{i,j} = 0\}$ means $\boldsymbol{d'}_{i,j}$ in the second-retrieval stage is classified as $negative$ combined with all $\boldsymbol{di}$ in the first-retrieval stage, then $\boldsymbol{d'}_{i,j}$ will be removed.

Classifier Forward Selection Unlike the CIS

Table 2: Results on different datasets with Llama3-8B as LLM. Adaptive Retrieval and Self-RAG conduct the retrieval module only under specific conditions (unpopular query entities or special retrieval tokens), so their time overhead is much less than other methods. We emphasize our results in bold.

	MuSiQue					HotpotQA					2Wiki				
Methods	EM	F1	Acc	Step	Time	EM	F1	Acc	Step	Time	EM	F1	Acc	Step	Time
Single-step Approach	13.80	22.80	15.20	1.00	1.00	34.40	46.15	36.40	1.00	1.00	41.60	47.90	42.80	1.00	1.00
Adaptive Retrieval	6.40	15.80	8.00	0.50	0.55	23.60	32.22	25.00	0.50	0.55	33.20	39.44	34.20	0.50	0.55
Self-RAG	1.60	8.10	12.00	0.73	0.51	6.80	17.53	29.60	0.73	0.45	4.60	19.59	38.80	0.93	0.49
Adaptive-RAG	23.60	31.80	26.00	3.22	6.61	42.00	53.82	44.40	3.55	5.99	40.60	49.75	44.73	2.63	4.68
Multi-step Approach	23.00	31.90	25.80	3.60	7.58	44.60	56.54	47.00	5.53	9.38	49.60	58.85	55.40	4.17	7.37
DR-RAG(Ours)	**26.97**	**38.90**	**34.03**	**1.00**	**1.54**	**48.58**	**62.87**	**55.68**	**1.00**	**1.40**	**49.60**	**55.62**	**55.18**	**1.00**	**1.43**

method, CFS method aims to remove the irrelevant dynamic-relevant documents in the second-retrieval stage. To achieve this goal, we search for a document $\boldsymbol{d_n}$ from D according to the <$\boldsymbol{q}, \boldsymbol{d_m}$> pair, and feed both the query and documents into C. When the classification result is negative, we will exclude the dynamic-relevant document in the current retrieved documents, and search for the next dynamic-relevant document which can be classified as positive with $_m$. The whole process is:

$$\begin{aligned} P_{i,j} &= \begin{cases} 1 & \text{if } C(\boldsymbol{q}, \boldsymbol{d}_i, \boldsymbol{d}'_{i,j}) = positive \\ 0 & otherwise \end{cases} \\ Cnt &= Cnt \cup \{\boldsymbol{d}'_{i,j} \mid P_{i,j} = 1 \wedge first\} \\ answer &= \mathtt{LLM}(\mathtt{Concat}(\boldsymbol{q}, Cnt)) \end{aligned} \tag{4}$$

where $Cnt = Cnt \cup \{\boldsymbol{d}'_{i,j} \mid P_{i,j} = 1 \wedge first\}$ means that for a given $\boldsymbol{d_i}$, the first $\boldsymbol{d}'_{i,j}$ in the second-retrieval stage classified as $positive$ combined with $\boldsymbol{d_i}$ will be considered as dynamic-relevant document and placed into Cnt.

3 Experiment Settings

The experimental details will be described in this section. Due to space constraints, the descriptions of implementation details, retrieval strategy and baseline can seen in Appendix A.1, A.2 and A.3.

3.1 Dataset

We verify the effectiveness of our proposed framework on three multi-hop QA datasets, including HotpotQA, 2Wiki and MuSiQue (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022b). The datasets require the system to comprehensively collect and contextualize information from multiple documents to answer more complex queries.

4 Results and Analysis

4.1 Main Results

Table 2 and 3 present the performance of DR-RAG in answering multi-hop query, and highlight the advantages of our approach compared to the sota RAG framework (Jeong et al., 2024; Asai et al., 2024) across multiple metrics, which is in line with our expectations. Table 5 shows the performance of DR-RAG across various retrieval strategies.

As shown in Table 2, when retrieving the same k documents, DR-RAG can achieve a higher recall rate and a higher percentage of correct answers. From the results, DR-RAG achieves better performance than other baseline RAG frameworks (self-RAG and Adaptive-RAG) on all three metrics. Moreover, DR-RAG is also less than other RAG frameworks in terms of the number of LLMs responses and the time consumed in QA systems.

4.2 Analysis

Ablation Study We propose a two-stage retrieval and classifier selection strategies to mine the dynamic relevance of documents. As shown in Table 3, we apple two classification methods based on QDC, and the experimental results have achieved further improvement. Table 4 shows the comparison of the effect of DR-RAG with and without QDC. Quantitatively, CIS and CFS can improve DR-RAG's performance by 2.3% and 4.7% on Acc metric against QDC, while DR-RAG reduces performance by 1.1% and 0.7% on Acc metric without QDC. The results demonstrate that the two strategies are able to efficiently extract document relevance and achieve more accurate answers.

Effects of Classifier and LLM Compared to gpt-3.5-turbo, gpt-4-turbo with better document comprehension has the ability to accurately capture the critical information to answer a query. As for textual responses, gpt-4-turbo generates responses

Table 3: Results on different LLMs and strategies compared to Adaptive-RAG. We set gpt-3.5-turbo and Llama3-8b as the base LLM. We emphasize our best results in bold. Top-k means the total number of retrieved documents.

			MuSiQue			**HotpotQA**			**2Wiki**		
top-k	**LLMs**	**Methods**	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
		Adaptive-RAG	23.60	31.80	26.00	42.00	53.82	44.00	40.60	49.75	46.40
3	**gpt-3.5**	**Query Document Concatenation**	21.40	32.20	29.90	37.80	51.56	52.60	36.20	48.99	51.40
		Classifier Inverse Selection	23.70	33.70	29.40	41.20	53.91	53.40	38.00	51.48	54.20
		Classifier Forward Selection	**26.00**	**36.20**	**35.00**	**43.80**	**58.83**	**55.00**	**48.40**	**60.13**	**64.20**
	Llama3-8B	**Query Document Concatenation**	21.30	32.00	27.10	43.85	56.38	49.88	40.84	48.61	43.76
		Classifier Inverse Selection	20.90	31.90	27.00	44.88	57.05	50.94	42.71	50.82	46.45
		Classifier Forward Selection	**26.50**	**37.40**	**32.60**	**48.71**	**62.42**	**55.26**	**50.48**	**59.29**	**56.07**
4	**gpt-3.5**	**Query Document Concatenation**	25.37	25.05	35.70	42.15	55.79	53.31	50.60	59.99	62.20
		Classifier Inverse Selection	25.80	36.50	35.60	42.00	56.10	54.36	49.40	60.40	65.00
		Classifier Forward Selection	**25.80**	**37.60**	**38.60**	**45.00**	**60.55**	**57.40**	**52.40**	**63.95**	**69.60**
	Llama3-8B	**Query Document Concatenation**	25.10	37.10	32.00	46.05	59.98	53.09	45.79	54.62	49.70
		Classifier Inverse Selection	25.70	37.50	32.70	48.02	61.43	54.36	**50.52**	**59.54**	**55.94**
		Classifier Forward Selection	**27.10**	**39.30**	**34.30**	**48.30**	**62.81**	**55.22**	50.30	59.40	55.85
6	**gpt-3.5**	**Query Document Concatenation**	28.80	41.30	38.60	45.41	60.09	61.60	48.20	62.58	67.00
		Classifier Inverse Selection	**31.20**	**42.70**	40.20	45.80	60.38	60.80	**52.60**	**65.93**	**71.20**
		Classifier Forward Selection	28.40	41.10	**40.60**	**48.20**	**63.83**	**63.60**	49.80	63.74	68.40
	Llama3-8B	**Query Document Concatenation**	25.90	38.10	33.70	46.19	60.21	54.14	44.02	53.08	48.73
		Classifier Inverse Selection	**27.50**	39.40	34.60	47.34	61.30	54.55	**50.59**	**59.58**	**56.29**
		Classifier Forward Selection	27.30	**40.00**	35.20	**48.73**	**63.38**	**56.56**	48.02	57.17	53.63

Table 4: Ablation study on HotpotQA by Llama3-8B.

top-k	**LLMs**	**EM**	**F1**	**Acc**
3	**CFS**	**48.71**	**62.42**	**55.26**
	w/o QDC	46.26	59.37	52.60
	CIS	**44.88**	**57.05**	**50.94**
	w/o QDC	44.82	57.00	50.88
4	**CFS**	**48.30**	**62.81**	**55.22**
	w/o QDC	47.78	61.63	54.73
	CIS	**48.02**	**61.43**	**54.36**
	w/o QDC	46.54	59.28	52.80
6	**CFS**	**48.73**	**63.38**	**56.56**
	w/o QDC	48.00	63.10	56.29
	CIS	**47.34**	**61.30**	**54.55**
	w/o QDC	46.30	60.00	53.95

of higher quality and more accurate content. Quantitatively, as shown in Table 7, gpt-4-turbo improve by an average of 9.07%, 10.63%, and 12.73% against gpt-3.5-turbo on three metrics. As shown in Table 6, when switching to different kinds or sizes of classifiers, the difference in the metrics is negligible (the extreme difference of EM, F1, and Acc is less than 2%), which suggests that our approach is applicable to different classifiers and that the classifier has little impact on our framework.

Effects of Recall Rate The ability of LLMs to answer domain-specific query correctly almost depends on whether all the necessary information is included in the prompt context. When relevant information is missing, it is difficult for LLMs with the hallucination problem to accurately answer the query. Table 8 illustrates the answers of the query with and without sufficient information provided to LLMs. As seen in Table 5, in 2Wiki, our retrieval strategy already achieves a recall rate of 98% when top-k is 6. When we feed enough relevant information into LLMs, the accuracy of their answers can be improved accordingly. CFS method achieves higher recall rate by 26.4% and 8.6% than BM25 and SM methods, respectively, which proves the feasibility of DR-RAG.

Effects of Redundant Information We hypothesise that if there is less redundant information in the contextual knowledge, LLMs can fully understand the query to reduce the hallucination. Therefore, CIS method is devised to validate this hypothesis. Invalid information may increase by about 30% as the number of documents fed into LLMs increases, but LLMs fail to judge the information when answering. LLMs may refer to redundant information and provide an answer with incorrect information. The results all validate our hypothesis that we should provide LLMs as little redundant or incorrect information as possible

Table 5: Recall rate and actual numbers under different retrieval strategies. Actual numbers represents the actual numbers of documents that we feed into LLMs. A smaller number means fewer redundant documents.

		MuSiQue		**HotpotQA**		**2Wiki**	
top-k	**Retrieval Strategies**	Recall rate	Actual numbers	Recall rate	Actual numbers	Recall rate	Actual numbers
3	**BM25**	37.57	3.00	64.67	3.00	57.46	3.00
	Similarity Matching	58.31	3.00	80.33	3.00	74.34	3.00
	Query Document Concatenation	58.44	3.00	86.12	3.00	74.48	3.00
	Classifier Inverse Selection	57.42	**2.82**	79.80	**2.74**	74.34	**2.60**
	Classifier Forward Selection	**66.45**	2.95	**88.89**	2.89	**87.81**	2.83
4	**BM25**	43.52	4.00	70.72	4.00	63.95	4.00
	Similarity Matching	66.20	4.00	85.57	4.00	80.05	4.00
	Query Document Concatenation	70.30	4.00	90.20	4.00	89.89	4.00
	Classifier Inverse Selection	66.45	**3.61**	88.37	**3.43**	89.71	**3.03**
	Classifier Forward Selection	**73.83**	3.82	**92.28**	3.75	**94.13**	3.41
6	**BM25**	51.69	6.00	79.93	6.00	73.95	6.00
	Similarity Matching	76.78	6.00	92.59	6.00	88.83	6.00
	Query Document Concatenation	79.46	6.00	94.00	6.00	95.04	6.00
	Classifier Inverse Selection	77.13	**5.23**	93.24	**4.72**	94.79	**3.72**
	Classifier Forward Selection	**83.01**	5.69	**96.27**	5.35	**98.04**	4.63

Table 6: Results of different classifier on HotpotQA dataset as Llama3-8B.

top-k	**Classifier**	**EM**	**F1**	**Acc**
3	**Bigbird-base(125M)**	48.71	62.42	55.26
	Bigbird-large(355M)	**49.20**	**62.89**	**55.55**
	Longformer(147M)	49.00	62.80	55.34
4	**Bigbird-base(125M)**	48.30	62.81	55.22
	Bigbird-large(355M)	**49.12**	**63.70**	**56.00**
	Longformer(147M)	50.27	64.41	56.90
6	**Bigbird-base(125M)**	48.73	63.38	56.56
	Bigbird-large(355M)	**49.02**	**63.58**	**56.52**
	Longformer(147M)	50.08	64.69	57.30

Table 7: Results of 500 samples sampled on HotpotQA dataset based on gpt-3.5-turbo and gpt-4-turbo.

top-k	**LLMs**	**EM**	**F1**	**Acc**
3	**gpt-3.5-turbo**	43.80	58.83	55.00
	gpt-4-turbo	**51.60**	**68.93**	**67.20**
4	**gpt-3.5-turbo**	45.00	60.55	57.40
	gpt-4-turbo	**53.40**	**70.75**	**69.80**
6	**gpt-3.5-turbo**	48.20	63.83	63.60
	gpt-4-turbo	**59.20**	**75.43**	**77.20**

throughout the RAG process. CIS method is effective in removing redundant information, but it may reduce the quality of responses when the reduction in recall is too large. Even though we feed all the relevant documents into LLMs, it is still possible to fail to get the right answer. In Table 5, on dataset 2Wiki, when the number of documents k provided to LLMs at 4 and 6, there is only a slight increase from CIS to CFS in the recall and instead a decrease

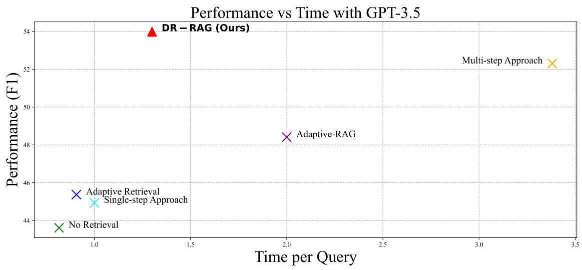

Figure 3: QA performance (F1) and time for different RAG frameworks. We use the GPT-3.5-turbo as the base LLM on the multi-hop QA datasets (MuSiQue, HotpotQA and 2Wiki).

in the metrics. Therefore, CFS method is propsed to balance redundant and relevant information.

Increase Recall with Lower Documents In CFS method, it seems impossible to find a match for every <q, d> pair in the second-retrieval stage because the documents we need have been retrieved. Therefore, there will be cases where the total number of our retrieved documents is less than k. For instance, in the HotpotQA dataset, when k is set to 6, the average number of documents actually provided to LLMs is 5.35, thereby reducing irrelevant information to some extent. CFS method in Table 5 achieves a higher recall rate while retrieving fewer actual numbers of documents compared to QDC method. CFS method yields higher scores across the three metrics in our experiments and achieves more significant retrieval capabilities with lower redundant inputs than other methods.

Time for One Response Compared to previous RAG frameworks, DR-RAG also achieves

Table 8: Case study with Llama3-8B, where we present the factual error in red and the accurate infomation in blue.

Datasets	Queries	Traditional RAG	DR-RAG (Ours)
MuSiQue	Who is the spouse of the Green performer?	**Retrieval strategy**: SM **Reasoning**: The spouse of the Green performer is Debbie Allen because she is the sister of the Green performer's sister. The answer is: Debbie Allen.	**Retrieval strategy**: CFS **Reasoning**: The Green performer mentioned in the question is Steve Hillage. Then we identify that Miquette Giraudy is mentioned as his partner. The answer is: Miquette Giraudy.
HotpotQA	what year was the cover artist of Multiverse: Exploring Poul Anderson's Worlds born ?	**Retrieval strategy**: SM **Reasoning**: The book "Multiverse: Exploring Poul Anderson's Worlds" was first published in hardcover in May 2014 by Subterranean Press. And assuming Bob Eggleton was around 50 years old. We can calculate his birth year is 1964. The answer is: 1964.	**Retrieval strategy**: CFS **Reasoning**: the cover artist of "Multiverse: Exploring Poul Anderson's Worlds" is Bob Eggleton. We also know that Bob Eggleton was born on September 13, 1960. The answer is: 1960
2Wiki	What nationality is the director of film Downpour (Film)?	**Retrieval strategy**: SM **Reasoning**: we know that "Downpour" is directed by Bahram Bayzai. the nationality of Bahram Bayzai is not explicitly mentioned in the provided contexts. However, based on my knowledge, Bahram Bayzai is an Iranian director. The answer is: Iranian	**Retrieval strategy**: CFS **Reasoning**: It is mentioned that the film "Downpour" is directed by Bahram Bayzai. And Bahram Beyzaie is associated with persian letters, arts. So the director of the film "Downpour" is of persia nationality. The answer is persia.

better time optimization during the whole process. Other RAG frameworks may call LLMs multiple times, resulting in high computational cost. In fact, the inference time of LLMs is also a worthwhile optimization in the applications. It takes a lot of time to call LLMs once, and calling them multiple times presents a catastrophic challenge in terms of time overhead. Therefore, we attempt to design a small model with relatively few parameters to achieve better optimization rather than calling LLMs multiple times. In Fig. 3 and Table 2, compared to Adaptive-RAG, we have achieved an average 74.2% reduction in time overheads. Therefore, we can conclude that we can achieve better experimental efficiency and the time overhead makes DR-RAG valuable in applications.

Case Study We conduct a case study to qualitatively compare our DR-RAG against the traditional RAG. Table 8 demonstrates the specific inference cases on the multi-hop datasets. For example, in MuSiQue dataset, our DR-RAG identifies the answer to the query by only using the LLM's parametric knowledge about 'partner'. Traditional RAG sometimes generate incorrect responses due to the inclusion of irrelevant information about 'sister'. Meanwhile, faced with a complex query, DR-RAG can first retrieve static-relevant documents based on 'cover artist' and 'Multiverse: Exploring Poul Anderson's Worlds' to get the name 'Bob Eggleton'. Then, in the second-retrieval stage, by combining the name 'Bob Eggleton' with 'born' in the query, dynamic-relevant documents can be retrieved to obtain the answer '1960'.

5 Conclusion

This paper presents DR-RAG, an innovative RAG framework designed to enhance document retrieval accuracy by leveraging the relevance of different documents in various QA scenarios. Throughout this research, we explore diverse retrieval strategies and conduct comprehensive experimental comparisons. Ultimately, we adopt CFS as the final framework, which not only reduces the number of redundant document but also achieves the most superior performance. Additionally, we analyze the utilization of dynamic document relevance under constrained training resources. The experimental results demonstrate that DR-RAG significantly improves answer quality and reduces the time required for QA systems.

6 Limitations

While DR-RAG has demonstrated excellent performance across multiple datasets for multi-hop QA, its implementation requires the prior training of a distinct classifier. It is uncertain whether our classifier will be effective with niche domains. Therefore, DR-RAG can serve as an invaluable inspiration to train a classifier with private data. In the future, we will collect more comprehensive data to train a more applicable classifier for various QA tasks.

7 Ethics Statement

DR-RAG substantiates its efficacy in real-world scenarios, which are characterized by diverse user queries. However, given the potential variability in user inputs, which may span a range from benign to offensive, it is imperative to consider scenarios where inputs might be detrimental. Such instances could facilitate the retrieval of objectionable content and lead to unsuitable responses by retrieval-augmented LLMs. Addressing this concern necessitates the development of robust methodologies to detect and mitigate offensive or inappropriate content in both user inputs and the documents retrieved within the RAG framework. This area represents a critical part for future research.

References

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

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. 2023. Gar-meets-rag paradigm for zero-shot information retrieval. *arXiv preprint arXiv:2310.20158*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with knn-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

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

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the preference gap between retrievers and llms. *arXiv preprint arXiv:2401.06954*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

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

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

Ye Liu, Semih Yavuz, Rui Meng, Meghana Moorthy, Shafiq Joty, Caiming Xiong, and Yingbo Zhou. 2023. Exploring the integration strategies of retriever and large language models. *arXiv preprint arXiv:2308.12574*.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*.

Fan Luo and Mihai Surdeanu. 2023. Divide & conquer for entailment-aware multi-hop evidence retrieval. *arXiv preprint arXiv:2311.02616*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Inc. NetEase Youdao. 2023. Bcembedding: Bilingual and crosslingual embedding for rag. https://github.com/netease-youdao/BCEmbedding.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. 2023. Visconde: Multi-document qa with gpt-3 and neural reranking. In *European Conference on Information Retrieval*, pages 534–543. Springer.

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

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialokg: Knowledge-structure aware task-oriented dialogue generation. *arXiv preprint arXiv:2204.09149*.

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

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022b. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 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)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Search-in-the-chain: Towards accurate, credible and traceable large language models for knowledgeintensive tasks. *CoRR, vol. abs/2304.14732*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences. *Preprint*, arXiv:2007.14062.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Appendix

A.1 Implementation Details

We follow the standard evaluation approach (Jeong et al., 2024) and validate our DR-RAG for QA systems by multiple metrics including F1, EM, and Accuracy (Acc). These metrics provide an objective measure between the prediction results and ground truth. In addition, the efficiency is also another issue we have to tackle. Most existing RAG frameworks (Asai et al., 2024; Jeong et al., 2024) require multiple calls to LLMs for inference. Therefore, we consider the number of inferences by LLMs and the time required for responses as our evaluation. To eliminate the effects of different LLMs, we select gpt-3.5-turbo (Achiam et al., 2023; Brown et al., 2020) and Llama3-8B (Liu et al., 2024) as base LLMs, and accurately acquire the answers to query based on retrieval documents. For the classifier C, we fine-tune bigbird-roberta-base (Zaheer et al., 2021) by the entire training set to accommodate longer input tokens. Due to the imbalance between positive and negative samples in the datasets, we sample the positive and negative examples to construct the datasets with the ratio of 1:1. In addition, we sample about 2300 pieces of data in each dataset, which exceeds the existing experiment (Jeong et al., 2024) in sample numbers.

A.2 Retrieval Strategy

DR-RAG aims to solve the problem of low recall in document retrieval. Therefore, five different retrieval strategies are designed to verify the effectiveness of our proposed DR-RAG.

- BM25: A method to measure the relevance between $\boldsymbol{q}$ and the documents.
- SM: The retrieval documents will be embedded and stored in D and the similarity between $\boldsymbol{q}$ and D is calculated to extract the k most relevant documents.
- QDC: We first retrieve k_1 documents from D, concatenate $\boldsymbol{q}$ with the documents to form multiple pairs and retrieve the k_2 most relevant documents for $<\boldsymbol{q}, \boldsymbol{d}>$ pairs, respectively, until the number of retrieved documents = k.

- CIS: To minimize document redundancy in QDC, all k documents retrieved are pairwise combined, concatenated with $\boldsymbol{q}$ and then fed into C to filter out irrelevant documents.

- CFS: To remove irrelevant dynamic-relevant documents, after retrieving k_1 documents, $<\boldsymbol{q}, \boldsymbol{d}>$ pairs are matched one by one with the remaining documents for similarity. Simultaneously, they have been fed into C for classification. If classified as negative, the process have been extended to the next document. Otherwise, positive instances will be included in the document set.

A.3 Baseline

We conduct a comprehensive comparison of our retrieval strategies against other RAG frameworks. In DR-RAG, we calculate the recall with different retrieval strategies and then evaluate the accuracy of the answers. Therefore, we select BM25 and SM methods (Lewis et al., 2020; Chan et al., 2024) as baselines. Moreover, we choose self-RAG (Asai et al., 2024) and Adaptive-RAG (Jeong et al., 2024), which are effective RAG frameworks for multi-hop QA, to validate the performance of our DR-RAG. In addition, we add the experimental results of Non-retrieval, original RAG and multi-step approach (Trivedi et al., 2023) to enrich our comparisons.

A.4 Related Works

A.4.1 RAG for Multi-hop QA

RAG is a popular framework for LLMs and has received much attention to many tasks, such as QA systems. RAG (Lewis et al., 2020) combined a sequence-to-sequence model with external knowledge bases to significantly improve the quality of quizzing and summarization tasks. The decomposition of a complex query (Khattab et al., 2022; Press et al., 2022; Pereira et al., 2023; Khot et al., 2022; Zhou et al., 2022) into a series of simpler sub-queries might inevitably require multiple calls to LLMs, resulting in high computational cost. Adaptive-RAG (Jeong et al., 2024) evaluated the complexity of the problem by a classifier and selects the most appropriate retrieval strategy based on the classification results. RQ-RAG (Chan et al., 2024) aimed to improve the performance of models by optimising search query, including rewriting, decomposition and disambiguation. However, it would be inefficient to access LLMs multiple times for each query and unreliable to retrieve all dynamic-relevant documents by a single query.

A.4.2 Retriever in RAG

The retriever in a RAG system is the key to verify how the retriever can obtain revelant instant contexts from external knowledge bases and alleviate the hallucination of LLMs. Fan et al. (2021) combined K Nearest Neighbor (KNN) retrieval with a traditional transformer model to dynamically access historical data and provide enough information by a composite memory. Cheng et al. (2024) proposed Selfmem to make the generated text more relevant to the retrieved information through a self-memory mechanism. Recent research has highlighted the potential applications of LLMs, which can be considered as supervised signals for training retrieval components, even as retrieval components. These findings provide us with new avenues for exploring the ability of retrievers to improve the efficiency of information retrieval based on the document relevance. In our work, we retrieve multiple relevant documents based on the query by a two-stage strategy and design a classifier to determine whether the documents can answer the query, and the remaining relevant documents are fed into LLMs with the query to obtain the answer.

A.5 More Analysis

Optimization under Resource Constraints The classifier C of document relevance requires certain hardware conditions and resources for data annotation. However, QDC method indicates that dynamic documents relevance can still be utilized without C. As seen in Table 5, compared to SM method, across all datasets, when top-k is 4 or 6, there is a significant increase of the retrieval recall by 3.84%. Yet, when top-k is 3, there is also 6% increase on HotpotQA and a slight increase on the other two datasets. This suggests that by making reasonable choices about top-k, even in cases where resources are limited, the performance of retrieval can be optimized by leveraging the relevance of relevant documents, thus improving LLMs' performance in QA tasks.

More Cases Table 9 shows the prompts we provide to LLMs. Contexts contain the documents (Document i) after retrieving and selecting. Moreover, we show the case of the classifier for selection in CIS and CFS methods in Table 10, the output cases compared to Adaptive-RAG in Table 11, the case of documents retrieved by QDC and CFS

Algorithm 1 Classifier Forward Selection (CFS)

Require:

Classifier C

Retrieval Function $\texttt{Retriever}$

Input query $\boldsymbol{q}$

Generated response $answer$

Initialize empty context: $Cnt = \{\}$

Retrieve k_1 documents: $\{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{k_1}\} = \texttt{Retriever}(\boldsymbol{q})$

Update context: $Cnt = Cnt \cup \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{k_1}\}$

for $i = 1$ to k_1 **do**

    Construct a new query: $\boldsymbol{q_i}^* = \texttt{concat}(\boldsymbol{q}, \boldsymbol{d_i})$

end for

Retrieve full set of documents for each new query:

$\{\boldsymbol{d}'_{i,1}, \boldsymbol{d}'_{i,2}, , \ldots, \boldsymbol{d}'_{i,k_2}\} = \texttt{Retriever}(\boldsymbol{q}_i^*),$ for $i = 1, 2, \ldots, k_1$

for $i = 1$ to k_1 **do**

    for $j = 1$ to k_2 **do**

        if $\boldsymbol{d}'_{i,j} \notin Cnt$ and $C(\boldsymbol{q}, \boldsymbol{d}_i, \boldsymbol{d}'_{i,j}) = positive$ **then**

            Update context: $Cnt = Cnt \cup \{\boldsymbol{d}'_{i,j}\}$

        end if

    end for

end for

Combine the input question with the updated context: $input = \texttt{concat}(\boldsymbol{q}, Cnt)$

Generate the answer using a large language model: $answer = \texttt{LLM}(input)$

return $answer$

methods in Table 12, the case of documents retrieved by QDC and CIS methods in Table 13.

Table 9: A case of our prompt provided to LLMs.

You are a reading comprehension expert, and you need to complete a reading comprehension task.

Contexts

Document 1:
Walk, Don't Run is a 1966 Technicolor comedy film directed by Charles Walters and starring Cary Grant in his final film role, Samantha Eggar, and Jim Hutton. The film is a remake of the 1943 film "The More the Merrier" and is set during the Olympic Games

Document 2:
Douglas Sirk(born Hans Detlef Sierck; 26 April 1897 – 14 January 1987) was a German film director best known for his work in Hollywood melodramas of the 1950s. Sirk started his career in Germany as a stage and screen director, but he left to Hollywood in 1937 because his Jewish wife was persecuted by the Nazis. In the 1950s, he achieved his greatest commercial success with film melodramas like "Imitation of Life All That Heaven Allows Written on the WindMagnificent Obsession" and "A Time to Love and a Time to Die". While those films were initially panned by critics as sentimental women's pictures, they are today widely regarded by film directors, critics and scholars as masterpieces. His work is seen as "critique of the bourgeoisie in general and of 1950s America in particular", while painting a" compassionate portrait of characters trapped by social conditions". Beyond the surface of the film, Sirk worked with complex mise enscenes and lush Technicolor colors to subtly underline his message.

Document 3:
The Mall, The Merrier is a 2019 Philippine musical family comedy film directed by Barry Gonzales, starring Vice Ganda and Anne Curtis. The film is co-produced by Star Cinema and Viva Films under the working title" Momalland". The film premiered in Philippine cinemas on December 25, 2019 as one of the official entries to the 2019 Metro Manila Film Festival. " The Mall, The Merrier" marks the first on- screen collaboration between Anne Curtis and Vice Ganda, both of whom are regular hosts in the noontime variety show" It's Showtime".

Document 4::
Robert Wallace Russell(January 19, 1912 – February 11, 1992) was an American writer for movies, plays, and documentaries. He was nominated for two Academy Awards for Best Writing, Original Story and Best Writing, Screenplay on the 1943 film "The More the Merrier". He died in 1992 in New York City, shortly after his 80th birthday.

Document 5:
Sleep, My Love is a 1948 American film noir directed by Douglas Sirk and starring Claudette Colbert, Robert Cummings and Don Ameche.

Document 6:
The More the Merrier is a 1943 American comedy film made by Columbia Pictures which makes fun of the housing shortage during World War II, especially in Washington, D.C. The picture stars Jean Arthur, Joel McCrea and Charles Coburn. The movie was directed by George Stevens. The film was written by Richard Flournoy, Lewis R. Foster, Frank Ross, and Robert Russell, from" Two's a Crowd", an original story by Garson Kanin(uncredited). This film was remade in 1966 as" Walk, Don't Run", with Cary Grant, Samantha Eggar and Jim Hutton.

After reading the documents above, answering the following question. Reasoning step by step. At last, you should output the final result via the following format:
Answer: <your answer based on the documents>;
Please answer the question directly.

Question
Which film has the director who died later, The More The Merrier or Sleep, My Love?

Give your analysis process first, and then output your answer in a specified format.

Table 10: Case of the classifier we train for selection in CIS and CFS methods. We mark relevant information that can influence classification results in blue.

Dataset	classified as *positive*	classified as *negative*
MuSiQue	**Query** In True Grit, who did the star play? **Document 1** True Grit is a 1969 American western film. It is the first film adaptation of Charles Portis' 1968 novel of the same name. The screenplay was written by Marguerite Roberts. The film was directed by Henry Hathaway and starred Kim Darby as Mattie Ross and John Wayne as U.S. Marshal Rooster Cogburn. Wayne won his only Academy Award for his performance in this film and reprised his role for the 1975 sequel Rooster Cogburn. **Document 2** In October 2015, TCM announced the launch of the TCM Wineclub, in which they teamed up with Laithwaite to provide a line of mail-order wines from famous vineyards such as famed writer-director-producer Francis Ford Coppola\'s winery. Wines are available in 3 month subscriptions, and can be selected as reds, whites, or a mixture of both. From the wines chosen, TCM also includes recommended movies to watch with each, such as a "True Grit" wine, to be paired with the John Wayne film of the same name.	**Query** In True Grit, who did the star play? **Document 1** True Grit is a 1969 American western film. It is the first film adaptation of Charles Portis' 1968 novel of the same name. The screenplay was written by Marguerite Roberts. The film was directed by Henry Hathaway and starred Kim Darby as Mattie Ross and John Wayne as U.S. Marshal Rooster Cogburn. Wayne won his only Academy Award for his performance in this film and reprised his role for the 1975 sequel Rooster Cogburn. **Document 2** The Iberian frog, Iberian stream frog or rana patilarga ("Rana iberica") is a species of frog in the family Ranidae found in Portugal and Spain. Its natural habitats are rivers, mountain streams and swamps. It is threatened by habitat loss, introduced species, climate change, water contamination, and increased ultraviolet radiation.
HotpotQA	**Query** The 2000 British film Snatch was later adapted into a television series for what streaming service? **Document 1** Snatch is a British/American television series based on the film of the same name, which debuted on Crackle on 16 March 2017. The show has been renewed for a second season. **Document 2** Snatch (stylised as snatch.) is a 2000 British crime comedy film written and directed by Guy Ritchie, featuring an ensemble cast. Set in the London criminal underworld, the film contains two intertwined plots: one dealing with the search for a stolen diamond, the other with a small-time boxing promoter (Jason Statham) who finds himself under the thumb of a ruthless gangster (Alan Ford) who is ready and willing to have his subordinates carry out severe and sadistic acts of violence.	**Query** The 2000 British film Snatch was later adapted into a television series for what streaming service? **Document 1** Snatch is a British/American television series based on the film of the same name, which debuted on Crackle on 16 March 2017. The show has been renewed for a second season. **Document 2** Orange Is the New Black (sometimes abbreviated to OITNB) is an American comedy-drama web television series created by Jenji Kohan for Netflix. The series is based on Piper Kerman\'s memoir, "" (2010), about her experiences at FCI Danbury, a minimum-security federal prison. "Orange Is the New Black" premiered on July 11, 2013 on the streaming service Netflix. In February 2016, the series was renewed for a fifth, sixth, and seventh season. The fifth season was released on June 9, 2017. The series is produced by Tilted Productions in association with Lionsgate Television.
2Wiki	**Query** Where was the composer of film Love Story 1999 born? **Document 1** Devanesan Chokkalingam, popularly known as Deva, is an Indian film composer and singer. He has composed songs and provided background music for Tamil, Telugu, Malayalam and Kannada films in a career spanning about 20 years. Many know his gaana songs, written mostly using Madras Tamil. He is known as the "Father of Gaana Genre" in the Tamil film industry. Deva has composed music for many films. He debuted as a film music director in the film "Manasukkeththa Maharaasa" in 1989. In the intervening years he has composed music for a total of more than 400 films. **Document 2** Love Story 1999 is a 1998 Indian Telugu-language romantic comedy film directed by K. Raghavendra Rao. The film had an ensemble cast starring Prabhu Deva, Vadde Naveen, Ramya Krishna, Rambha and Laila in the lead roles.	**Query** Where was the composer of film Love Story 1999 born? **Document 1** Devanesan Chokkalingam, popularly known as Deva, is an Indian film composer and singer. He has composed songs and provided background music for Tamil, Telugu, Malayalam and Kannada films in a career spanning about 20 years. Many know his gaana songs, written mostly using Madras Tamil. He is known as the "Father of Gaana Genre" in the Tamil film industry. Deva has composed music for many films. He debuted as a film music director in the film "Manasukkeththa Maharaasa" in 1989. In the intervening years he has composed music for a total of more than 400 films. **Document 2** Tarcisio Fusco was an Italian composer of film scores. He was the brother of the composer Giovanni Fusco and the uncle of operatic soprano Cecilia Fusco.

Table 11: Cases that the query can be answered correctly in DR-RAG, and can not in Adaptive-RAG. We present wrong answer in red and the right answer in blue.

Dataset	Query and Answer	Answer by Adaptive-RAG	Answer by DR-RAG
MuSiQue	**Query** What is the birth date of the By Your Side performer? **True Answer** 1 January 1986.	The performer of "By Your Side" is James Cottriall. James Cottriall was born on 20 May 2011. So the answer is: 20 May 2011.	It is mentioned that James Cottriall was born on 1 January 1986. Additionally, in context, it is stated that "By Your Side" is a song by James Cottriall.Putting this information together, we can determine that the birth date of James Cottriall, the performer of the song "By Your Side," is 1 January 1986. Answer: <1 January 1986>
HotpotQA	**Query** Stephen susco wrote a movie produced by which american film actor? **True Answer** Do Kyung-soo	The member of the South Korean-Chinese boy group EXO who stars in the upcoming movie "The Underdog" is Sehun. So the answer is: Sehun.	It is mentioned that Do Kyung-soo, also known by his stage name D.O., is a member of EXO and has starred in various television dramas and movies. In context, it is mentioned that the upcoming South Korean animated film "The Underdog" stars Do Kyung-soo. Therefore, based on the information provided, Do Kyung-soo stars in the upcoming movie "The Underdog". Answer: <Do Kyung-soo>
2Wiki	**Query** Where was the place of death of the director of film The First Day Of Freedom? **True Answer** Florida	The director of The First Day of Freedom is Aleksander Ford. Aleksander Ford died in Warsaw, Poland. So the answer is: Warsaw, Poland.	In context, it is mentioned that Aleksander Ford committed suicide in 1980 in Naples, Florida, United States. Given the information provided, the place of death of Aleksander Ford, the director of the film "The First Day of Freedom," was Naples, Florida, United States. Answer: <Naples, Florida, United States.>

Table 12: A case of documents retrieved by QDC and CFS on the MuSiQue dataset, where the necessary documents are in blue, and the top-k is 4.

Query	Documents retrieved by QDC	Documents retrieved by CFS
Who is the spouse of the Green performer?	**Document 1:** Green is the fourth studio album by British progressive rock musician Steve Hillage. Written in spring 1977 at the same time as his previous album, the funk-inflected "Motivation Radio" (1977), "Green" was originally going to be released as "The Green Album" as a companion to "The Red Album" (the originally intended name for "Motivation Radio"). However, this plan was dropped and after a US tour in late 1977, "Green" was recorded alone, primarily in Dorking, Surrey, and in London. **Document 2:**: "Little Green" is a song composed and performed by Joni Mitchell. It is the third track on her 1971 album "Blue". **Document 3:**: The Main Attraction is an album by American jazz guitarist Grant Green featuring performances recorded in 1976 and released on the Kudu label. **Document 4:**: Grant's First Stand is the debut album by American jazz guitarist Grant Green featuring performances by Green recorded and released on the Blue Note label in 1961. Earlier recordings made by Green for Blue Note were released as "First Session" in 2001.	**Document 1:**: Green is the fourth studio album by British progressive rock musician Steve Hillage. Written in spring 1977 at the same time as his previous album, the funk-inflected "Motivation Radio" (1977), "Green" was originally going to be released as "The Green Album" as a companion to "The Red Album" (the originally intended name for "Motivation Radio"). However, this plan was dropped and after a US tour in late 1977, "Green" was recorded alone, primarily in Dorking, Surrey, and in London. **Document 2:**: "Little Green" is a song composed and performed by Joni Mitchell. It is the third track on her 1971 album "Blue". **Document 3:**: Miquette Giraudy (born 9 February 1953, Nice, France) is a keyboard player and vocalist, best known for her work in Gong and with her partner Steve Hillage. She and Hillage currently form the core of the ambient band System 7. In addition to her performances in music, she has also worked as an actress, film editor and writer. In each role, she has used different stage names.

Table 13: A case of documents retrieved by QDC and CIS on the HotpotQA dataset, where the necessary documents are in blue, and the top-k is 4.

Query	Documents retrieved by QDC	Documents retrieved by CIS
Who is the child of Caroline LeRoy's spouse?	**Document 1:**: Caroline LeRoy Webster (September 28, 1797 in New York City – February 26, 1882) was the second wife of 19th Century statesman Daniel Webster. Her father was Herman LeRoy, who was once head of the commercial house of Leroy, Bayard, McKiven Co., a largetrading company that operated in different partsof the world. Her father was also the first Holland Consul to the United States. Caroline's mother was Hannah Cornell, daughter of the last Royal Attorney General of the State of North Carolina. Caroline was a descendant of Thomas Cornell. **Document 2:**: Pierre Paul Leroy-Beaulieu (9 December 1843 in Saumur – 9 December 1916 in Paris) was a French economist, brother of Henri Jean Baptiste Anatole Leroy-Beaulieu, born at Saumur, Maine -et-Loire on 9 December 1843, and educated in Paris at the Lycée Bonaparte and the École de Droit. He afterwards studied at Bonn and Berlin, and on his return to Paris began to write for "Le Temps", "Revue nationale" and "Revue contemporaine". **Document 3:**: Daniel Fletcher Webster, commonly known as Fletcher Webster (July 25, 1813 in Portsmouth, New Hampshire – August 30, 1862) was the son of renowned politician Daniel Webster and Grace Fletcher Webster. He was educated at Harvard College. During his father's first term as Secretary of State, Fletcher served as Chief Clerk of the United States State Department which, at the time, was the second most powerful office in the State Department. As Chief Clerk, he delivered the news of President William Henry Harrison's death to the new President, John Tyler. **Document 4:**: Leroy, also Leeroy, LeeRoy, Lee Roy, LeRoy or Le Roy, is both a male given name in English - speaking countries and a family name of French origin. Leroy (lørwa) is one of the most common surnames in northern France. As a surname it is sometimes written Le Roy, as a translation of Breton Ar Roue. It is an archaic spelling of le roi, meaning "the king "and is the equivalent of the English surname King.	**Document 1:**: Daniel Fletcher Webster, commonly known as Fletcher Webster (July 25, 1813 in Portsmouth, New Hampshire – August 30, 1862) was the son of renowned politician Daniel Webster and Grace Fletcher Webster. He was educated at Harvard College. During his father's first term as Secretary of State, Fletcher served as Chief Clerk of the United States State Department which, at the time, was the second most powerful office in the State Department. As Chief Clerk, he delivered the news of President William Henry Harrison's death to the new President, John Tyler. **Document 2:**: Caroline LeRoy Webster (September 28, 1797 in New York City – February 26, 1882) was the second wife of 19th Century statesman Daniel Webster. Her father was Herman LeRoy, who was once head of the commercial house of Leroy, Bayard, McKiven Co., a largetrading company that operated in different partsof the world. Her father was also the first Holland Consul to the United States. Caroline's mother was Hannah Cornell, daughter of the last Royal Attorney General of the State of North Carolina. Caroline was a descendant of Thomas Cornell. **Document 3:**: Leroy, also Leeroy, LeeRoy, Lee Roy, LeRoyor Le Roy, is both a male given name in English - speaking countries and a family name of French origin. Leroy (lørwa) is one of the most common surnames in northern France. As a surname it is sometimes written Le Roy, as a translation of Breton Ar Roue. It is an archaic spelling of le roi, meaning "the king "and is the equivalent of the English surname King.