Chain-of-Rewrite: Aligning Question and Documents for Open-Domain Question Answering

Chunlei Xin[1,2], **Yaojie Lu**[1,*], **Hongyu Lin**[1], **Shuheng Zhou**[3], **Huijia Zhu**[3], **Weiqiang Wang**[3], **Zhongyi Liu**[3], **Xianpei Han**[1], **Le Sun**[1]
[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Ant Group
{chunlei2021, luyaojie, hongyu, xianpei, sunle}@iscas.ac.cn,
{shuheng.zsh, huijia.zhj, weiqiang.wwq, zhongyi.lzy}@antgroup.com

Abstract

Despite the advancements made with the retrieve-then-read pipeline on open-domain question answering task, current methods still face challenges stemming from term mismatch and limited interaction between information retrieval systems and large language models. To mitigate these issues, we propose the Chain-of-Rewrite method, which leverages the guidance and feedback gained from the analysis to provide faithful and consistent extensions for effective question answering. Through a two-step rewriting process comprising Semantic Analysis and Semantic Augmentation, the Chain-of-Rewrite method effectively bridges the gap between the user question and relevant documents. By incorporating feedback from the rewriting process, our method can self-correct the retrieval and reading process to further improve the performance. Experiments on four open-domain question answering datasets demonstrate the effectiveness of our system under zero-shot settings.

1 Introduction

Open-Domain Question Answering (ODQA) (Voorhees and Tice, 2000; Chen et al., 2017; Izacard and Grave, 2021b) is a long-standing task aimed at answering a wide variety of user questions without providing specific background documents. Recently, with the emergence of large language models (LLMs), ODQA systems typically employ a *retrieve-then-read* paradigm (Lee et al., 2019; Karpukhin et al., 2020; Lewis et al., 2020), where an information retrieval system is used to identify relevant contexts from knowledge bases, and then an LLM such as InstructGPT (Ouyang et al., 2022) and LLaMA (Touvron et al., 2023) is used to read the context and generate comprehensive and accurate answers. Due to its effectiveness and simplicity, the retrieve-then-read paradigm performs well on the ODQA task and has received considerable attention (Yang and Seo, 2021; Shi et al., 2024; Zhang et al., 2023b).

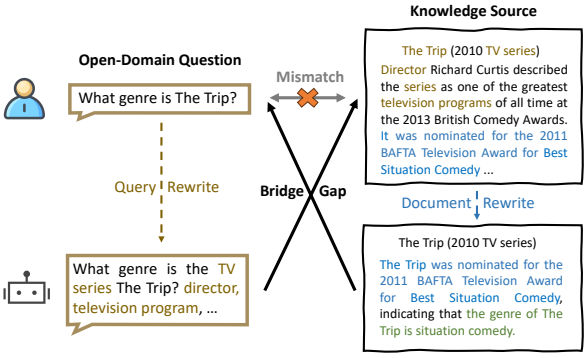

Figure 1: An example illustrates how term mismatch between an open-domain question and relevant documents can be mitigated through query rewriting and document rewriting. Query expansions and their corresponding matches in the document are shown in brown, while key sentences that point to the answer are shown in blue.

Unfortunately, the *retrieve-then-read* paradigm faces challenges stemming from term mismatch and limited interaction between IR systems and LLMs. Firstly, in open-domain question answering, term mismatch often occurs due to ambiguous user intent and limited overlap between user questions and the corpus (Custis and Al-Kofahi, 2007; Zhao and Callan, 2012). This mismatch can negatively impact both the retrieval of relevant documents and the effectiveness of using these documents to answer questions. For example, as shown in Figure 1, "What genre is The Trip?" shares few words with the relevant document "The Trip (2010 TV series)" and misses the important keyword "TV series", making it difficult to match questions with relevant documents. Secondly, the disparate preferences between LLMs and retrievers further hinder the effectiveness of current retrieve-then-read paradigm (Guu et al., 2020; Ke et al., 2024), as there is limited interaction between the retrieval and reading

*Corresponding Author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1884–1896

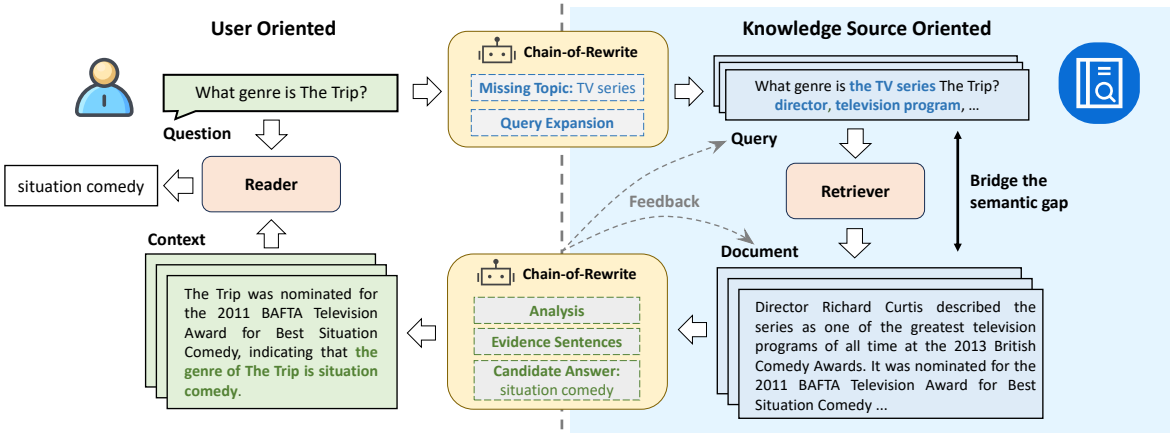

Figure 2: Overall architecture of our proposed Chain-of-Rewrite method, which aims to address the term mismatch and limited interaction problems in ODQA scenarios.

steps. This lack of interaction makes it difficult to self-correct the retrieval and reading processes by incorporating feedback from both components. For example, the LLM can guide the retriever with the important keyword "TV series" by identifying the strong connections between <genre, TV series> and <The Trip, TV series>. In turn, the retriever can refine the query based on the feedback indicating the candidate answers provided by the reader.

To address the term mismatch and limited interaction problems, this paper proposes a Chain-of-Rewrite method. This method can effectively bridge the gap between the open-domain question and relevant documents through a two-step rewriting process. As shown in Figure 2, to capture the underlying intent behind the user question and map it to relevant documents, we use Chain-of-Rewrite to transform the user question into clarification search queries. These queries can further express the user intent to improve the retrieval effectiveness. To bridge the expression mismatch between the user question and retrieved documents, we use Chain-of-Rewrite to remove irrelevant content and highlight relevant information to make the original documents better match the question. Meanwhile, by leveraging the feedback from the chain-of-rewrite process, our method can self-correct the retrieval and reading process.

Specifically, the Chain-of-Rewrite method consists of two main modules: Semantic Analysis and Semantic Augmentation. In Semantic Analysis, we prompt the LLM to identify the topics and key information embedded in the original question and document. This goes beyond surface-level matching and aims to capture the underlying relationships within the query and relevant documents. Building upon the insights gained from Semantic Analysis, Semantic Augmentation focuses on enhancing the semantic representation of the original question and document, leading to more consistent and informative extensions. Together, these steps ensure that the rewritten content is aligned with user intent while maintaining semantic consistency. To further self-correct the potential errors caused by distracting documents, we leverage the feedback from the rewriting process to prioritize passages that point to candidate answers and to refine search queries for collecting more relevant documents to correct potentially incorrect answers.

Overall, our main contributions can be summarized as follows:

- We propose an unsupervised rewriting method, Chain-of-Rewrite, to address the term mismatch and the limited interaction problems in ODQA scenarios.
- We design two novel modules, Semantic Analysis and Semantic Augmentation, to ensure that the rewritten content matches the user intent while maintaining semantic consistency.
- We incorporate feedback from the rewriting process to effectively refine search queries and re-rank documents to self-correct the retrieval and reading process.

2 Proposed Method

Our method is based on the retrieve-then-read pipeline (Lee et al., 2019; Karpukhin et al., 2020; Lewis et al., 2020), which first uses a retriever to retrieve relevant documents from external knowledge source, and then uses a reader to predict the

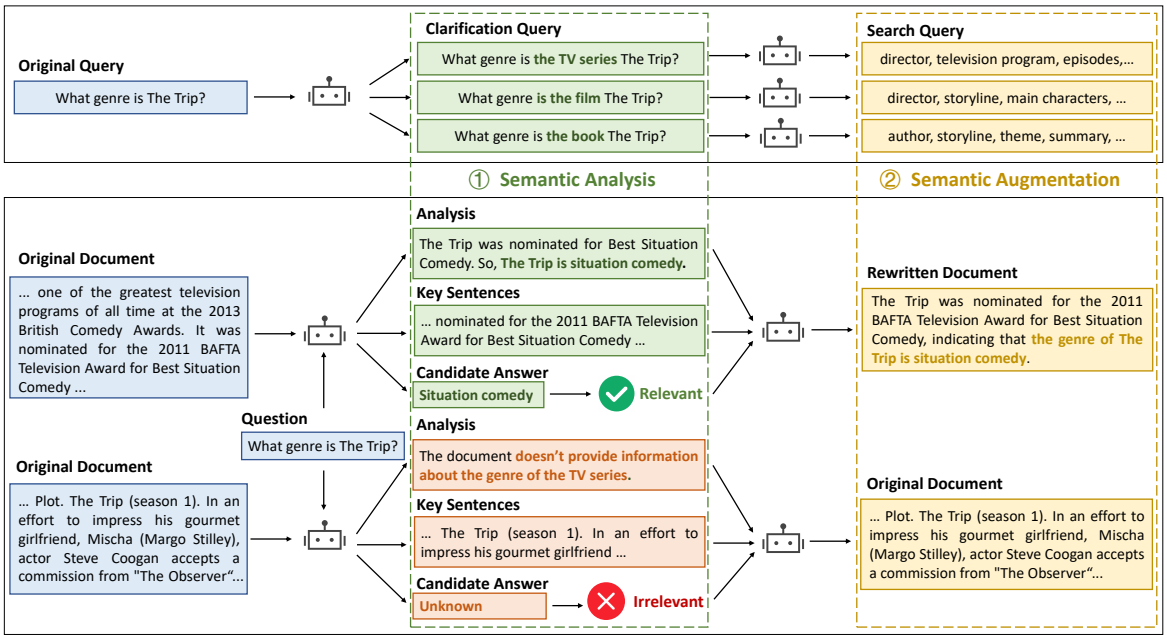

Figure 3: The Chain-of-Rewrite method comprises two main modules: Semantic Analysis and Semantic Augmentation.

answer based on these documents. Specifically, the retrieve-then-read pipeline can be represented as $p(a|q) = \sum_i p(a|d_i, q)p(d_i|q)$, marginalizing over all possible documents. In practice, the k highest ranked documents are used to approximate the sum over d, yielding $p(a|q) = \sum_{i=1}^{k} p(a|d_i, q)p(d_i|q)$. While the retrieve-then-read frameworks show remarkable performance on the ODQA task, there exist mismatch between the user question q and documents d in knowledge source, posing challenges for existing ODQA systems to accurately understand the user intent and utilize relevant documents to correctly answer the question.

To solve the above mismatch problem, we explore a two-step rewriting process, Chain-of-Rewrite, to align the user question and relevant documents. As shown in Figure 2, given an open-domain question q, we use Chain-of-Rewrite to transform the user question into clarification search queries q^r. These clarification queries can further express the user intent and contain additional relevant terms to improve the retrieval effectiveness. This can be represented as $p(q^r|q, \theta_{cow})$, where θ_{cow} represents the parameters of the Chain-of-Rewrite model. Query q^r is then fed into the retriever to retrieve a collection of relevant documents d. To remove irrelevant content and highlight relevant information in document d, we use Chain-of-Rewrite to transform the original document d into context c based on the question q. This document rewriting process can be represented as $p(c_i|d_i, q, \theta_{cow})$. Therefore, we reformulate the retrieve-then-read pipeline as $p(a|q) = \sum_i p(a|c_i, q)p(c_i|d_i, q, \theta_{cow})p(d_i|q^r)p(q^r|q, \theta_{cow})$.

Next, we will describe the two main modules in 2.1 and 2.2, and discuss the feedback in 2.3.

2.1 Step 1: Semantic Analysis

In this step, our primary objective is to prompt the LLM to identify the topics and key information embedded in the original question and document, which forms the basis for the subsequent Semantic Augmentation process.

For query reformulation, we prompt the LLM to add potentially missing topics and details to the original question q to obtain clarification queries q^c that can further express the user intent. This is important for open domain question answering, where questions often do not have a clear domain or topic scope, typically resulting in ambiguous user intent. By utilizing the world knowledge stored in the model parameters, this step can provide reliable guidance for subsequent query expansion.

For document reformulation, as shown in Figure 3, we first prompt the LLM to provide the relationship between the original question and retrieved documents through a step-by-step analysis, and to extract the evidence sentences from the document to support its analysis. This step narrows the question-document mismatch in two ways. Firstly, the step-by-step analysis directly provides the re-

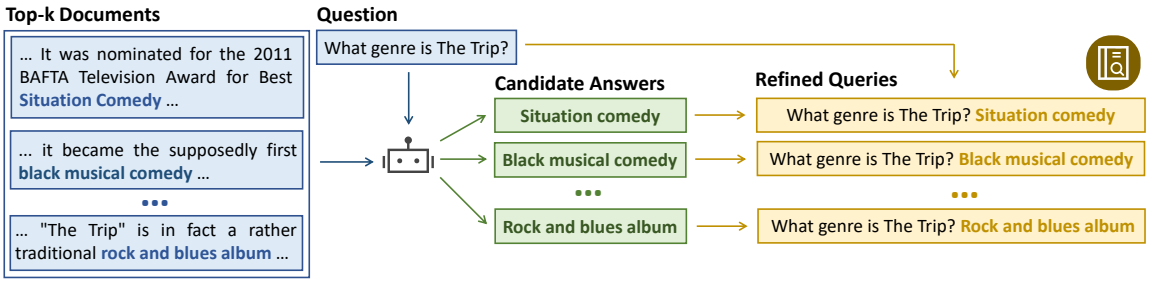

Figure 4: Illustration of incorporating automatic feedback to refine search queries. Candidate answers extracted from top-k re-ranked documents are concatenated with the original question as refined queries to collect relevant documents.

lationship between the document and the question, identifying highly relevant documents from irrelevant ones. Secondly, by identifying the key sentences relevant to the given question, LLM can effectively remove distracting and irrelevant content, hence enabling the LLM to hit the correct answer with higher precision and less effort. Based on the analysis, we prompt the LLM to identify the candidate answer from the given document, and offer an option (e.g. "unknown") to reject irrelevant documents.

2.2 Step 2: Semantic Augmentation

The focus of this step is to bridge the term mismatch between the IR system and LLMs by taking into account the semantic analysis obtained in step 1 to produce faithful and consistent extensions.

For query reformulation, we prompt the LLM to provide a list of highly relevant expanded queries based on the clarification query q^c, from which we can extract various keywords to expand the original question q. For each clarification query q^c, we form query expansions by randomly sampling 4 to 8 keywords from all the generated expansions. These sampled keywords are concatenated with the clarification query q^c to form an expanded query q^r, which is then used to retrieve a set of m relevant documents d. The relevant documents retrieved by all expanded queries are aggregated by de-duplicating, with only the highest-scoring version of each duplicate retained. Finally, the top N documents, ranked by their retrieval scores, are selected as the final retrieval results.

For document reformulation, we aim to provide a concise and straightforward context that focuses on the given question while maintaining semantic and factual consistency. Specifically, we concatenate key sentences and a sentence that explicitly points to the answer as a rewrite of the relevant document, while leaving the irrelevant document unchanged. The top k rewritten documents are then concatenated as a context c, which is then fed into the reader with the original question q to produce the answer a.

2.3 Automatic Feedback

Incorporating semantic analysis into the rewriting process not only provides guidance to maintain consistency, but can also provide automatic feedback to self-correct the retrieval and reading process by effectively refining queries and re-ranking retrieved documents.

2.3.1 Refine Query

In many cases, retrieved documents may be relevant but not sufficient to point to the correct answer. These documents are likely to mislead the LLM into producing an incorrect or incomplete answer, thereby reducing the effectiveness and robustness of the ODQA system. To collect more relevant documents to refine potentially incorrect answers, as shown in Figure 4, candidate answers extracted from top k relevant documents can be used as automatic feedback to augment the original question. By providing more relevant documents focused on the candidate answer, incorrect answers extracted from misleading documents can be refined, leading to more accurate and reliable responses.

2.3.2 Re-rank Documents

Given the retrieved documents d, the re-ranker is expected to rank the documents based on the query-document relevance. For unsupervised passage re-ranking, a widespread method UPR (Sachan et al., 2022) is to rank documents according to the average log-likelihood of the question tokens conditioned on the passage:

$$p(d_i|q) \propto \frac{1}{|q|} \sum_t \log p\left(q_t|q_{<t}, d_i; \Theta\right)$$

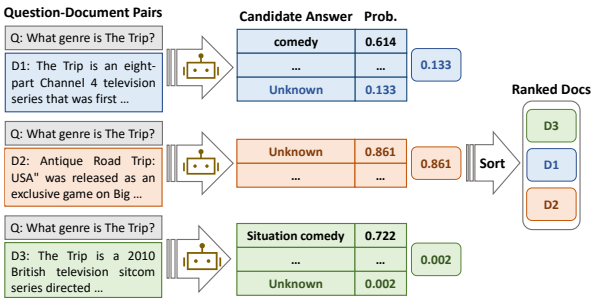

Figure 5: Illustration of incorporating automatic feedback to re-rank documents. The output probability of "unknown" is used to estimate the relevance of the question-document pair.

where Θ denotes the parameters of the re-ranker model and $|q|$ denotes the number of question tokens. However, due to a significant mismatch between this relevance assessment method and the training objective of next-word prediction in LLMs (Zhang et al., 2023a), UPR is suboptimal.

To effectively distinguish truly relevant documents from distracting ones that may superficially overlap with the question, we estimate the query-document relevance at a deeper semantic level. As shown in Figure 5, we utilize the output probability of "unknown" when predicting candidate answer as an automatic feedback, to estimate the relevance of the question-document pair:

$$p(d_i|q) \propto 1 - p(\text{unknown}|q, d_i, g)$$

where g represents the analysis provided in step 1.

Note that our method complements the UPR method by focusing on deep semantic relationships rather than surface-level token matching. Based on the prediction likelihood of the input question conditioned on a passage, UPR favors passages that contain question tokens, overlooking whether these passages actually help in answering the question. In contrast, our approach focuses on this critical aspect. In our main experiments, we employ the standard Reciprocal Rank Fusion (RRF) approach to combine the re-ranking results of our method with those of UPR.

3 Experiments

3.1 Experiment Setup

Dataset. We conduct extensive experiments on four open-domain question answering datasets, including Natural Question (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (WebQ) (Berant et al., 2013) and PopQA (Mallen et al., 2023), which focuses on questions about long-tail entities. We use the same splits as previous approaches (Karpukhin et al., 2020; Izacard and Grave, 2021b).

Retrieval System. We employ the representative sparse retriever BM25 (Robertson and Zaragoza, 2009) to collect relevant documents from the English Wikipedia dump from December 20, 2018. This choice is based on the observation that the performance of dense retrieval methods like DPR (Karpukhin et al., 2020) degrades rapidly on rarer entities, whereas BM25 exhibits less sensitivity to entity frequency (Sciavolino et al., 2021). Since our experiments focus on open-domain questions, particularly on datasets like PopQA with long-tail entities, BM25 was selected for its robustness and effectiveness across diverse datasets.

Implementation. We choose Vicuna-13B-v1.5 (Chiang et al., 2023) as the base model for chain-of-rewriting open-domain questions and documents. For query reformulation, we collect up to 5 clarification queries for each user question. Each expanded query is used to retrieve a set of m=30 documents. After de-duplication, top N=100 documents with highest retrieval scores are selected as the final retrieval results. From these, top k=5 evidence passages are selected and fed to the reader. We employ Vicuna-13B-v1.5 and GPT-3.5-Turbo (OpenAI., 2022) as the reader in main experiments.

Metrics. We use exact match (EM) and F1 scores for NQ, TriviaQA and WebQ, and follow the same normalization process utilized in previous work (Karpukhin et al., 2020; Chen et al., 2017; Lee et al., 2019). For PopQA, we adopt the same evaluation metric, accuracy score, as used by Mallen et al. (2023). A prediction is considered correct if any substring of the prediction exactly matches one of the golden answers.

Baseline Methods. We compare our method with the following zero-shot QA baselines: (1) Close book methods without using retriever: Vicuna-13B-v1.5 (Chiang et al., 2023), GPT-3.5-Turbo (OpenAI., 2022) and GenRead (Yu et al., 2023). (2) Retrieval-augmented LLMs: Vicuna-13B-v1.5 and GPT-3.5-Turbo augmented with BM25 retrieved documents from Wikipedia. (3) Self-Ask (Press et al., 2023): decomposing questions into sub-questions for multi-step reasoning. (4) Distraction-aware Answer Selection (DAS) (Cho et al., 2023): incorporating unanswerable instruction and selecting an answer from candidates. (5) ALLIES (Sun et al., 2023): iteratively refining and expanding the original query.

Models	NQ		WebQ		TriviaQA		PopQA
	EM	F1	EM	F1	EM	F1	ACC
Close book methods without using retriever							
Vicuna-13B	22.1	29.4	23.7	35.9	50.1	56.9	30.2
GPT-3.5-Turbo	29.4	40.7	22.8	40.0	54.8	65.5	37.9
GENREAD (InstructGPT) (Yu et al., 2023)	31.1	44.8	19.1	36.9	59.3	70.7	46.0
Retrieval-augmented LLMs							
Vicuna-13B + BM25	25.5	34.9	20.2	34.2	53.7	64.0	34.6
Vicuna-13B + BM25 + UPR	28.9	39.2	22.4	35.9	56.5	67.0	39.0
GPT-3.5-Turbo + BM25	33.2	42.8	23.3	38.2	57.6	68.7	43.9
Decomposing questions for multi-step reasoning							
Self-Ask (Davinci-002) (Press et al., 2023)	26.4	36.5	15.1	29.5	59.4	68.5	33.6†
Selecting an answer from candidates							
Vicuna-13B w/ DAS (Cho et al., 2023)	27.1†	38.5†	23.4†	35.8†	54.9†	62.3†	31.2†
Iteratively refining the original query							
ALLIES (GPT-3.5-Turbo) (Sun et al., 2023)	38.0	47.8	**28.2**	**45.6**	61.4	70.8	37.6†
Our method, bridging the mismatch between the question and relevant documents							
Chain-of-Rewrite (Reader=Vicuna-13B)	34.0	45.3	24.8	40.4	59.3	70.4	48.0
Chain-of-Rewrite (Reader=GPT-3.5-Turbo)	**40.2**	**49.3**	27.1	41.4	**61.6**	**71.7**	**50.2**

Table 1: Main results on four open-domain question answering benchmarks under zero-shot settings. The best performing pipelines are highlighted in bold. Results marked with † are from our runs with their released code.

3.2 Overall Results

As shown in Table 1, by effectively handling the mismatch between questions and retrieved documents and by self-correcting the retrieval and reranking process, our Chain-of-Rewrite method significantly outperforms the retrieval-augmented LLMs baselines. Specifically, when employing Vicuna-13B-v1.5 as the reader, incorporating Chain-of-Rewrite process can achieve an improvement of over 13 points on PopQA, which focuses on questions about long-tail entities. The same trend can be observed when GPT-3.5-Turbo is used as the reader to answer questions. These experimental results demonstrate the effectiveness of our Chain-of-Rewrite method in improving the ability of LLMs to handle open-domain questions, especially for long-tail questions.

Furthermore, we observe that applying GPT-3.5-Turbo as the reader enhances the performance of our system on all datasets, achieving the highest performance on most open-domain question answering datasets. Notably, on the long-tail ODQA dataset PopQA, baseline methods do not perform as well as expected, with many even falling short of the performance of retrieval-augmented LLMs. This is mainly due to suboptimal retrieval effectiveness on PopQA, negatively impacting subsequent optimizations that depend on these retrieval results. It's important to note that in our Chain-of-Rewrite (Reader=GPT-3.5-Turbo) configuration, GPT-3.5-Turbo is only used for question answering, while the entire Chain-of-Rewrite process is still based on Vicuna-13B. In contrast, ALLIES employs GPT-3.5-Turbo for the entire pipeline, including query generation, question answering, and answer scoring. Despite this, our pipeline achieves comparable or even superior performance to ALLIES, showcasing the effectiveness and efficiency of our system.

3.3 Detailed Analysis

3.3.1 Question Answering Performance

To explore the impact of the guidance and automatic feedback provided in our Chain-of-Rewrite method on improving the question answering performance, we conduct ablation studies on four ODQA datasets, and the results are shown in Table 2. We can see that:

(1) *Bridging the mismatch between the question and the retrieved documents is crucial for open-domain question answering.* After removing the

	NQ		WebQ		TriviaQA		PopQA		
	EM	F1	EM	F1	EM	F1	ACC	EM	F1
Chain-of-Rewrite	**34.0**	**45.3**	**24.8**	**40.4**	**59.3**	**70.4**	48.0	**41.7**	**49.2**
w/o refine query feedback	33.6	45.0	24.4	39.7	58.7	69.7	47.4	41.5	48.9
w/o rerank documents feedback	32.4	43.3	23.9	39.9	58.9	70.3	47.0	40.6	47.9
w/o chain-of-rewrite documents	29.5	39.2	23.1	38.0	58.3	69.4	**48.2**	40.4	47.4
w/o chain-of-rewrite question	29.9	40.4	23.7	38.6	56.7	67.5	38.4	33.0	39.6
Vicuna-13B + BM25	25.5	34.9	20.2	34.2	53.7	64.0	34.6	28.8	34.3

Table 2: Ablation study. Our proposed chain-of-rewrite components and the automatic feedback provided during rewriting can improve overall performance on four ODQA benchmarks. We employ Vicuna-13B-v1.5 as the Reader.

	PopQA		NQ		WebQ		TriviaQA	
	R@5	R@20	R@5	R@20	R@5	R@20	R@5	R@20
Baseline	37.41	50.12	49.25	67.45	50.30	68.11	68.67	77.70
Direct Expansion	48.02	52.11	62.35	75.10	61.12	73.87	75.95	80.71
Chain-of-Rewrite	**59.80**	**65.71**	**63.57**	**77.40**	**61.71**	**74.51**	**76.81**	**81.31**

Table 3: Recall@5 and Recall@20 on test sets across four open-domain question answering datasets.

chain-of-rewrite documents component, the EM and F1 scores decrease significantly for all datasets. Specifically, the EM and F1 scores decrease by 4.5 and 5.9 points respectively for the NQ dataset. A similar trend can be observed when the chain-of-rewrite query component is removed.

(2) *Expanding the user question with candidate answers obtained during chain-of-rewriting can effectively provide more relevant documents.* When incorporating automatic feedback to refine the search queries, our method further improves both EM and F1 scores across all datasets.

(3) *The "unknown" output probability based on the step-by-step analysis is an effective estimate of query-document relevance.* As shown in Table 2, while applying the Chain-of-Rewrite method solely to the top-5 documents (w/o rerank documents feedback) leads to significant improvements compared to Vicuna-13B + BM25, incorporating automatic feedback to re-rank documents yields further enhancements across all datasets. This observation confirms that utilizing the output probability of "unknown" to measure the query-document relevance can effectively distinguish truly relevant documents from distracting documents.

3.3.2 Passage Retrieval Performance

For detailed analysis of passage retrieval performance, we incorporate Recall@K (R@K) as the evaluation metric, which calculates the percentage of top-K retrieved documents that contain the correct answer. The results are shown in Table 3, and we can see that our method outperforms Direct Expansion on all ODQA datasets. Specifically, we see the Recall@5 and Recall@20 improvements of +11.8 and +13.6 (24.6% and 26.2% relative improvement) respectively on PopQA, which focuses on open-domain questions about long-tail entities. This demonstrates that by adding clear topic and details to the open-domain question, our Chain-of-Rewrite process can provide reliable guidance to generate detailed and consistent query expansions, leading to improved retrieval effectiveness.

3.3.3 Passage Ranking Performance

To evaluate the passage ranking performance, we compute the top-K retrieval accuracy score following Sachan et al. (2022). It is defined as the proportion of questions for which at least one passage within the top-K passages contains a span that matches the answer. We report the top-1, top-5 and top-10 retrieval accuracy on three ODQA datasets in Table 4, and report the top-1 and top-10 retrieval accuracy on different question categories in PopQA in Figure 6.

As shown in Table 4, incorporating rerank document feedback into the BM25 ranking or BM25+UPR ranking can provide consistent passage ranking performance improvements across all datasets, with particularly significant performance gains in top-1 retrieval accuracy and top-5 retrieval accuracy. In particular, BM25 + Rerank Feedback

	NQ			WebQA			TriviaQA		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
BM25	24.02	47.65	60.30	25.59	51.13	62.75	45.41	64.87	72.10
+ Rerank Feedback	40.64	65.87	72.94	37.11	65.40	**73.13**	59.81	76.91	79.85
BM25 + UPR	36.12	63.57	72.05	36.91	61.71	68.90	61.95	76.81	79.46
+ Rerank Feedback	**46.76**	**69.20**	**75.01**	**43.45**	**66.24**	72.39	**66.46**	**78.08**	**80.49**

Table 4: Top-{1, 5, 10} retrieval accuracy on test sets across three open-domain question answering datasets.

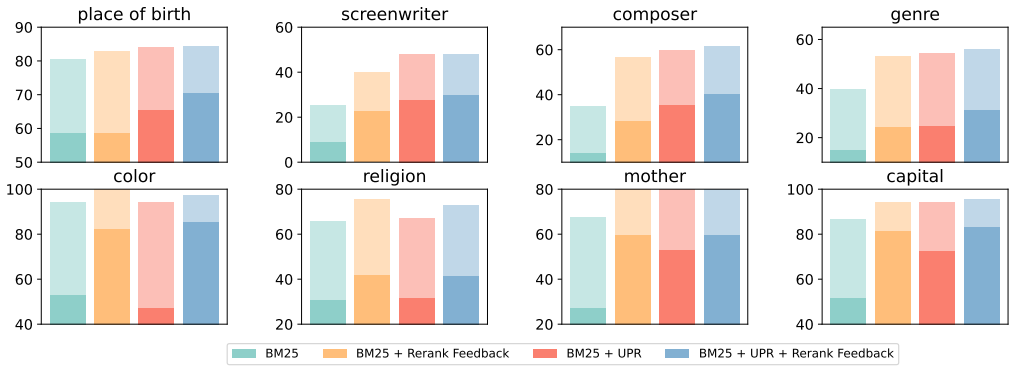

Figure 6: Top-1 and Top-10 retrieval accuracy for different question categories in the PopQA dataset. Top-1 retrieval accuracy is shown in darker color and Top-10 retrieval accuracy is shown in lighter color.

improves the top-1 retrieval accuracy by +16.62 (69.2% relative improvement) and top-5 retrieval accuracy by +18.22 (38.2% relative improvement) on NQ dataset. Furthermore, BM25 + Rerank Feedback outperforms BM25 + UPR on all datasets, demonstrating the effectiveness of our method for prioritizing highly relevant documents to the top compared to UPR. The same trend can be verified in Figure 6, where incorporating rerank feedback leads to an improvement in passage ranking performance on different question categories in PopQA.

4 Related Work

4.1 Open-Domain Question Answering

Open-domain question answering can be categorized into two settings: the closed-book setting and the open-book setting. In the closed-book setting, pre-trained language models answer open-domain questions directly without access to the external corpus. In the open-book setting, the ODQA system typically consists of a retriever and a reader component (Izacard and Grave, 2021a; Yang and Seo, 2021; Zhang et al., 2023b). The retriever finds relevant information from a corpus such as Wikipedia (Chen et al., 2017; Izacard and Grave, 2021b) or web pages (Nakano et al., 2021; Lazaridou et al., 2022), followed by a reader that focuses on answering the question based on the retrieved information.

Recently, the emergence of large language models has demonstrated their potential to be used for open-domain question answering. With no training data or external corpus, LLMs are able to provide answers with direct prompts (Brown et al., 2020; Wei et al., 2022; Chiang et al., 2023; Yang et al., 2024). At the retrieval stage, the LLM can be used as query rewriter to refine input questions to better express user intent (Liu et al., 2022; Chuang et al., 2023; Qin et al., 2023), or as knowledge sources to provide relevant contextual documents that increase the likelihood of covering the correct answer (Yu et al., 2023; Li et al., 2024). At the reader stage, LLMs can effectively reduce distractions from irrelevant documents to improve the context quality (Levine et al., 2022; Cho et al., 2023).

4.2 Query and Document Expansion

In order to overcome the term mismatch problem between open-domain questions and relevant documents in ODQA scenarios, various query expansion and document expansion approaches are proposed. Document expansion enhances each document with additional information, such as additional terms selected from a corpus (Billerbeck and Zobel, 2005; Dai and Callan, 2020), or expanded queries generated by pre-trained language models

based on the original document (Nogueira et al., 2019a,b). On the other hand, query expansion adds additional terms selected from the top-ranked documents retrieved by the initial query (Abdul-Jaleel et al., 2004; Metzler and Croft, 2005, 2007) or expands the original query with lexical-level (Zukerman and Raskutti, 2002) or phrase-level (Riezler et al., 2007) paraphrases.

Recently, leveraging LLMs for expansion has proven to be a promising solution to address the term mismatch problem (Wang et al., 2023; Lee et al., 2023). Among them, Chen et al. (2024) propose the AGR method, which prompts the LLM to generate potential answers as query expansions after analyzing the question. In contrast, our query reformulation method focuses more on narrowing the expression gap between the user question and the corpus, without requiring the LLM to directly generate answers. This reduces reliance on the LLM's internal knowledge and demonstrates improved generalization and robustness, especially for long-tail questions.

5 Conclusion

We propose an unsupervised rewriting method, Chain-of-Rewrite, which utilizes pre-trained language models to effectively address the term mismatch and limited interaction problems. We design two modules, Semantic Analysis and Semantic Augmentation, to progressively bridge the mismatch between the user question and relevant documents while maintaining semantic consistency. During the two-step rewriting process, we incorporate automatic feedback into the retrieval and reading process to self-correct the potential errors. Extensive experiments on four ODQA datasets demonstrate the effectiveness of our system in performing question answering, passage retrieval and passage ranking tasks under zero-shot settings.

6 Limitations

In this work, we propose an effective unsupervised rewriting method, Chain-of-Rewrite, to tackle the challenges in open-domain question answering scenarios. The limitations of the proposed method are as follows:

(1) Due to the need for in-depth analysis and reconstruction, our approach requires more computational resources and time than using a retrieve-then-read pipeline that takes the original questions and documents as input. However, in situations where computational resources are constrained, there are strategies to mitigate the computational overhead while still maintaining considerable performance gains. One such strategy involves selectively applying the chain-of-rewrite process to the top-ranked documents, rather than all retrieved documents. As indicated in Table 2 under the "w/o rerank documents feedback" condition, the performance decrease is marginal compared to rewriting all retrieved documents.

(2) Due to constraints in computational resources and time, our experiments were limited to using the Vicuna-13B model as the base model for chain-of-rewriting open-domain questions and documents retrieved from Wikipedia. However, our framework can be easily applied to models with varying scales and architectures, and we are interested in investigating the effectiveness of our approach on other LLMs.

7 Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by Beijing Natural Science Foundation (L243006), the Natural Science Foundation of China (No. 62106251 and 62306303), and Ant Group Research Fund.

References

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. *Computer Science Department Faculty Publication Series*, page 189.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Bodo Billerbeck and Justin Zobel. 2005. Document expansion versus query expansion for ad-hoc retrieval. In *Proceedings of the 10th Australasian document computing symposium*, pages 34–41.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11908–11922, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Sukmin Cho, Jeongyeon Seo, Soyeong Jeong, and Jong Park. 2023. Improving zero-shot reader by reducing distractions from irrelevant documents in open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3145–3157, Singapore. Association for Computational Linguistics.

Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. 2023. Expand, rerank, and retrieve: Query reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12131–12147, Toronto, Canada. Association for Computational Linguistics.

Tonya Custis and Khalid Al-Kofahi. 2007. A new approach for evaluating query expansion: query-document term mismatch. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 575–582, New York, NY, USA. Association for Computing Machinery.

Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1897–1907. ACM / IW3C2.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the preference gap between retrievers and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10438–10451, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *ArXiv*, abs/2203.05115.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yejoon Lee, Philhoon Oh, and James Thorne. 2023. Knowledge corpus error in question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9183–9197, Singapore. Association for Computational Linguistics.

Yoav Levine, Ori Ram, Daniel Jannai, Barak Lenz, Shai Shalev-Shwartz, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022. Huge frozen language models as readers for open-domain question answering. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024. Self-prompting large language models for zero-shot open-domain QA. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 296–310, Mexico City, Mexico. Association for Computational Linguistics.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2014–2029, Seattle, United States. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Donald Metzler and W. Bruce Croft. 2005. A markov random field model for term dependencies. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 472–479. ACM.

Donald Metzler and W. Bruce Croft. 2007. Latent concept expansion using markov random fields. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 311–318. ACM.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to doctttttquery. *Online preprint*, 6:2.

Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *CoRR*, abs/1904.08375.

OpenAI. 2022. Introducing chatgpt. *See https://openai.com/index/chatgpt*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. WebCPM: Interactive web search for Chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8968–8988, Toronto, Canada. Association for Computational Linguistics.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471, Prague, Czech Republic. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 8371–8384. Association for Computational Linguistics.

Hao Sun, Xiao Liu, Yeyun Gong, Yan Zhang, Daxin Jiang, Linjun Yang, and Nan Duan. 2023. Allies: Prompting large language model with beam search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3794–3805, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Sohee Yang and Minjoon Seo. 2021. Designing a minimal retrieve-and-read system for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5856–5865, Online. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023a. Rankinggpt: Empowering large language models in text ranking with progressive enhancement. *CoRR*, abs/2311.16720.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023b. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

Le Zhao and Jamie Callan. 2012. Automatic term mismatch diagnosis for selective query expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 515–524, New York, NY, USA. Association for Computing Machinery.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

A Detailed Prompts

In this subsection, we provide a thorough description of the prompts utilized in our proposed Chain-of-Rewrite method. Prompts used for our question reformulation and document reformulation are shown in Section A.1 and A.2, respectively. Section A.3 shows a prompt template to instruct LMs to answer the given question based on retrieved documents.

It is important to note that these prompts are configured for the Vicuna-13B-v1.5 model, and slight adaptability adjustments may be necessary when applying them to other models.

A.1 Query Reformulation

For query reformulation, we first prompt the LLM to add potentially missing topics and details to the original question to obtain clarification queries with the following prompt:

> Given a query that might be ambiguous or unclear, provide a bullet-point list of explicit queries that the ambiguous query could pertain to.
>
> Query: {query}
> List of explicit queries:

Then, we prompt the LLM to provide a list of highly relevant expanded queries based on the clarification query, from which various terms and phrases are extracted to expand the original question. The detailed prompt is:

> Based on the given query, generate a bullet-point list of diverse related queries that will find relevant documents. Next to each point, extract keywords from it that are closely related to the original query.
>
> Query: {clarification query}
> List of related queries:

A.2 Document Reformulation

For document reformulation, we first prompt the LLM to provide the step-by-step analysis of whether the document is relevant to the given question and extract the evidence sentences. The detailed prompt is:

> In the first paragraph, give a step-by-step analysis of whether the document provides information to answer the question. And in the second paragraph, you should give the evidence sentences extracted from the document.
>
> Document: {document}
> Question: {question}
> Output:

Based on the analysis, we prompt the LLM to identify the candidate answer based on the given document, and offer an option (e.g. "unknown") to reject irrelevant documents.

> Based on the analysis, extract the answer entity from the document to answer the following question. Output "unknown" as the answer if there is no relevant information in the document.
>
> Question: {question}
> Document: {document}
> Analysis: {analysis}
> Answer:

Note that in this step, we collect not only the answers predicted by the LLM, but also the probability of LLM predicting 'unknown', which is used as an automatic feedback to to rerank documents.

A.3 Question Answering

In our experiments, we use the following prompt to ask the reader to answer the given question based on evidence passages.

> Passages: {evidence passages}
>
> Extract the single answer entity from the passages to answer the following question. Only output the answer entity.
>
> Question: {question}
> Answer: