Missing Knowledge in Retrieval-Augmented Generation: Aligning User Queries with Knowledge Base

Anonymous ACL submission

Abstract

Retrieval Augmented Generation (RAG) frameworks mitigate hallucinations in Large Language Models (LLMs) by integrating external knowledge, yet face two critical challenges: (1) the distribution gap between user queries and knowledge bases, and (2) incomplete coverage of required knowledge for complex queries. Existing solutions either require task-specific annotations or neglect inherent connections among query, context, and missing knowledge interactions. We propose a Missing Knowledge RAG Framework that synergistically resolves both issues through Chain-of-Thought reasoning. By leveraging open-source LLMs, our method generates structured missing knowledge queries in a single inference pass while aligning query knowledge distributions, and integrates reasoning traces into answer generation. Experiments on open-domain medical and general QA datasets demonstrate significant improvements in context recall and answer accuracy. The framework achieves effective knowledge supplementation without additional training, offering enhanced interpretability and robustness for real-world question answering applications.

1 Introduction

The rapid advancement of Large Language Models (LLMs), exemplified by various architectures (OpenAI et al., 2024; Grattafiori et al., 2024; Qwen et al., 2025; DeepSeek-AI et al., 2024), has demonstrated remarkable improvements on a wide range of natural language processing tasks. However, their reliance on static, parametric knowledge often leads to hallucinations, factual inaccuracies, and outdated responses, particularly in dynamic or domain-specific scenarios. To mitigate these limitations, Retrieval-Augmented Generation (RAG) frameworks have emerged as a promising paradigm for knowledge intensive tasks (Lewis et al., 2020), integrating external knowledge bases with LLMs to enhance answer reliability. While standard RAG pipelines retrieve contextually relevant documents to ground LLM outputs, two critical challenges persist:

(1) Distribution Gap between user queries and different knowledge bases, which undermines re trieval relevance. As illustrated in the upper portion of Figure 1, a significant disparity exists between colloquial/non-professional user descriptions and formal medical literature. Previous approaches primarily focus on training memory networks to generate task-specific cues (Qian et al., 2024) or employ adaptive evidence retrieval (Li et al., 2024) to bridge semantic gaps, which often require additional annotation efforts. Alternative solutions involve query rewriting (Ma et al., 2023) or query decomposition enhanced by Monte-Carlo Tree Search (MCTS) (Jiang et al., 2024). While these methods demonstrate partial success, they overlook a naturally aligned knowledge source that inherently matches user query distributions: historical user Q&A pairs. As shown in Fig. 1, compared with specialized medical texts, these historical pairs exhibit stronger alignment with user queries and intents, making them valuable resources for addressing user inquiries.

(2) Missing Knowledge, where retrieved contexts fail to fully cover the knowledge required to answer complex queries. Existing works lies in two paths. One lines of work proposes iterative retrieval directly using the first round answer (Shao et al., 2023) or the intrinsic reasoning capabilities of LLMs (i.e., GPT-3.5) to separately generate missing information and new queries for subsequent retrieval (Wang et al., 2025). The other line of work propose to generate evidence and critics with special tokens in one single pass adaptively (Islam et al., 2024; Asai et al., 2024). Although effective, prior works might neglect the internal connection among user query, context, missing knowledge and derived missing knowledge query with distribu-

User Query: 请问怀孕期间在电池厂制片车间负极上班对胎儿有影响吗？
May I ask if working in the negative electrode section of the production workshop at a battery factory during pregnancy has any impact on the fetus?

Retrieved wiki chunks:
怀孕应该注意哪些: 怀孕时应避免做放射性检查......孕妇应注意生活中每个细小的环节，并耐心地应对，让孕期生活轻松愉快。
During pregnancy, it is important to avoid radiological examinations......Pregnant women should pay attention to every small detail in daily life and respond with patience, ensuring a relaxed and enjoyable pregnancy.

Prediction:
在电池厂制片车间工作确实可能接触多种有害物质......建议您尽快咨询医生，并详细说明您的职业环境，以便得到专业指导和必要监测。此外，尽量采取有效措施减少工作中有害物质的暴露，确保母婴安全。

Retrieved QA Pairs:
Q: 有个电子公司，生产锂电池，应聘我去做质检一职，会影响我怀孕吗？或者我怀孕了上班锂电池的辐射会影响胎儿正常发育吗？
Q: There is an electronics company that produces lithium batteries, and they are offering me a position in quality inspection. Will this job affect my pregnancy? Or, if I am pregnant, will the radiation from the lithium batteries affect the normal development of the fetus?
A: 病情分析：你好配料工序会产生一些粉状漂浮物，主要是原材料粉，涂布工序会产生一些挥发性气体，比如NMP，注液工序会有电解液气体......指导意见：......
Medical Analysis: During the production process of lithium batteries, the ingredient mixing stage generates some powdered airborne particles, primarily raw material powders. The coating process produces some volatile gases, such as NMP, and the electrolyte filling process releases electrolyte gases... Guidance: ...

Prediction:
您好，您在电池厂制片车间工作可能会接触一些化学物质，比如负极材料里的石墨、粘结剂以及溶剂等。建议还是要加强个人防护措施，减少长期处于这种环境下的时间。同时，记得按时做产前检查，确保胎儿健康成长。
Hello, working in the electrode manufacturing workshop of a battery factory may expose you to certain chemical substances, such as graphite, binders, and solvents in the anode materials. It is advisable to enhance personal protective measures and reduce the amount of time spent in such an environment. Additionally, remember to attend regular prenatal check-ups to ensure the healthy development of the fetus.

Gold Answer:
指导意见：你好，根据你的情况来看，怀孕期间在车间工作有一定程度影响，电池带有放射性。
Guidance: Hello, based on your situation, working in the workshop during pregnancy may have some impact, as batteries can emit radiation.

Figure 1: Comparison of retrieved text chunks for the same query from Huatuo-26M (Li et al., 2023), a large scale chinese open domain medical QA dataset, where plain texts and QA pairs come from medical encyclopedia, articles and health websites respectively. Here, we highlight the relevant information in blue with query in red.

tion mismatch between query and knowledge bases, which might lead to suboptimal retrieval outcomes.

In this work, we propose a Missing Knowledge RAG Framework based on Chain-of-Thought (CoT) (Wei et al., 2022) which systematically addresses both challenges through a unified, efficient pipeline. Unlike prior methods, our approach explicitly consider the intrinsic relationships the query, retrieved context, missing knowledge and its corresponding query. With the help of powerful reasoning capabilities of open-sourced LLM (>70B), our framework enables the LLM to generate structured, JSON-formatted missing knowledge queries in a single-step inference, while aligning query and knowledge base distributions. Furthermore, the generated reasoning traces are seamlessly incorporated into the final answer generation process, ensuring both interpretability and accuracy.

1. We comprehensively explore methods to bridge the gap between query and knowledge base distributions by leveraging multi-source knowledge bases within a real world question answering system, without additional training.

2. We introduce an efficient one-pass (RAG) framework incorporating missing knowledge query generation, which explicitly exploits the inherent relationships among queries, contexts, missing knowledge, and their corresponding queries.

3. We conduct extensive experiments on two open-domain question-answering datasets, evaluating both general and domain-specific scenarios. Additionally, we provide a detailed analysis of the effectiveness with our proposed framework through context recall metrics.

2 Method

In the following section, we will first define the problem, followed by a comprehensive analysis of the encountered challenges, and finally propose our RAG framework augmented with missing knowledge integration.

2.1 Problem Statement

Given a user query $\mathcal{Q}$, the task of RAG system is to first retrieve contexts $\mathcal{C} = \{c_1, c_2, ...\}$ which is closely related to user query, and then generate a

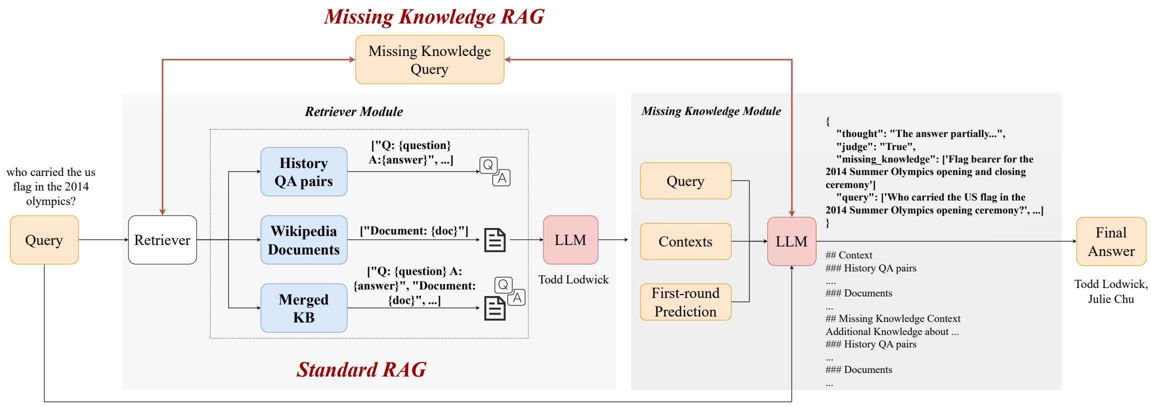


Figure 2: Illustration of our Missing Knowledge RAG framework. Our pipeline first retrive from knowledge bases, and prompt Open sourced LLM to give draft answer. Second, LLM need to decide is there any knowledge missing. Different from standard RAG pipeline, the LLM need to generate missing knowledge and query with JSON format in a single-turn. Finally, after second-time retrieval with generated query, we prompt all the retrieved contexts with explicit knowledge to generate the final answer.

final answer $\mathcal{Y}$ based on query and contexts. The final goal of a RAG system is to ensure the comprehensive and precise retrieval of contexts relevant to the query, thereby facilitating the accurate generation of corresponding answers compared with ground truth $\mathcal{A}$.

2.2 RAG with missing knowledge framework

Retriever Module To address the distribution gap between queries and text chunks in the knowledge base in practical applications, we propose to retrieve from different knowledge bases (i.e. Wikipedia chunks, historical QA pairs or the merged knowledge base).

As presented in Figure 2, in standard RAG pipeline, given a user query Q, we first retrieve top-k chunks with the encoded query embedding and similarity search.

$$
\begin{aligned}
&qE = Encoder(\mathcal{Q}), \\
&tE = \{Encoder(\mathcal{D}_i), i = 1, ..., |\mathcal{D}|\}, \\
&\mathbf{V} = \{\sqrt{\sum_{j=1}^{d}(tE_{ij} - qE_j)^2} | i = 1, ..., |\mathcal{D}|\}, \\
&\mathcal{C} = \{\mathcal{D}[i] | i \in \arg \text{Top-k}(\mathbf{V})\},
\end{aligned} \tag{1}
$$

where, $\mathcal{D}$ denotes the retrieved knowledge base with text chunks, $\mathbf{V}$ and $\mathcal{C}$ refer to encoded vector set and Topk selected contexts according to L2 metric respectively.

Then we could get the first round answer with frozen LLM generation.

$$\hat{Y} = LLM_\theta(x_i | q, \mathcal{C}, x_{<i}, i = 1, ..., t) \tag{2}$$

Missing Knowledge Module To mitigate the potential issue of missing knowledge, we propose a single-pass CoT prompting method to indentify the missing knowledge and generate its corresponding query at the same time, considering the potential distribution gap between missing knowledge query and knowledge bases. Previous efforts need a two-stage missing information pipeline with iterative retrieval (Wang et al., 2025) or train a specific generative module (Qian et al., 2024). Instead, we consider the intrinsic relationships between the query, context, and missing knowledge, and employ an efficient reasoning approach with open-sourced LLM, utilizing the power of CoT to get JSON formatted outputs. Specifically, for example, given query $\mathcal{Q}$ "`who carried the us flag in the 2014 Olympics?`", retrieved contexts $\mathcal{C}$ and first round prediction $\hat{Y}$, we utilize CoT prompt to generate analysis part "thought", judgment about whether knowledge missing or not "judge", missing knowledge cues "`missing_knowledge`", which is "`flag bearer for the 2014 summer Olympics opening and closing ceremony`" and finally corresponding query "`who carried... opening ceremony?`" and "`who carried... closing ceremony?`" aligned with knowledge base distribution. The detailed prompt for missing knowledge query and answer generation is presented in Appendix A.

Finally, we utilize the aligned missing knowledge query to retrieve relevant information from the specified knowledge base. Subsequently, we apply a straightforward deduplication function using md5 hashing to remove redundant text chunks. Then, we consolidate them with part of CoT information into a structured prompt to generate the final answer, which contains "`## Context...## Missing Knowledge Context...{thought}...## Question...`"

3 Experiment

3.1 Experiment Setup

We fairly evaluated our framework under a one-shot setting on two open-domain question answering datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019) for general knowledge question answering, which consists of real-world user queries from search engines. And we further experiment on specific medical domain, Huatuo-26M (Li et al., 2023), which is a large-scale chinese medical QA dataset curated from online healthcare QA websites These datasets are ideal benchmarks to evaluate the robustness of the proposed framework, which contain a large volume of high-quality QA pairs, reflecting the natural distribution of user interactions across various real-world QA systems.

We utilize open-sourced Llama3.3-70B (Grattafiori et al., 2024) and Qwen2.5-70B-Instruct (Qwen et al., 2025) as our backend LLM within all modules in the framework for English and Chinese benchmark respectively, owing to their success on LMSys leaderboard under specific categories[1]. We employ bge-en-large (Xiao et al., 2023) and bge-large-zh-v1.5 with specific prompt to encode query and text chunks, following instructions on the website[2] and use Faiss index (Johnson et al., 2019) IndexFlat with L2 metrics for similarity search.

For NQ evaluation, we use all the documents provided in the NQ dataset directly without any HTML tag to construct wikipedia knowledge base. To accommodate the maximum length constraints of the encoder model, each document is segmented into text chunks containing fewer than 300 words, resulting in 4.8 million text chunks, followed by a deduplication process with md5 hash, as Table 2 shows. To provide a fair comparison, we also leverage all the QA pairs in the training set of NQ and prompt the LLM to summarize documents into pseudo QA pairs in the test set. This process yields 128,000 QA pairs, which undergo a deduplication procedure using md5 hashing. Furthermore, we merge all the document chunks with history QA pairs to get a merged knowledge base with 7.3 million text chunks, followed by a similar deduplication preocess.

For Huatuo-26M evaluation, we use all the provided encyclopedia articles, segmenting them into text chunks of 400 tokens to construct the medical wikipedia knowledge base, while using all the consultant records to construct history QA pairs. Furthermore, we merged all the encyclopedias with consultant records, getting 9 million text chunks for the merged knowledge base.

During evaluation, we use normalized exact match (EM) and word level F1-score to compare final prediction with ground truth answer. For medical open domain QA task, we use ROUGE and BLEU score to evaluate. In order to comprehensively and fairly evaluate if the retrieved contexts contains all the knowledge needed to answer user query, we develop context recall metric with powerful LLM[3] (i.e. DeepSeek (DeepSeek-AI et al., 2024)[4]), as ground truth contexts are not available under this scenario. First, we prompt DeepSeek-V3 with $\mathcal{Q}$ and $\mathcal{A}$ to get labeled ground truth context $GT_{\mathcal{C}}$, which contains all the necessary knowledge to answer the user query. Then, we further prompt it to independently judge if the retrieved context could be attributed in the $GT_{\mathcal{C}}$ within the JSON format. Specifically, the output is a list containing attribution judgement of contexts with reason, which is like "`{'context': string, 'attributed': boolean, 'reason': string}`". Finally, we could calculate the context recall score with the following formula:

$$context\_recall = \frac{\sum_{i=1}^{K} \mathbb{1}_{attributed[i]}(\mathcal{C}_i)}{|GT_{\mathcal{C}}|}, \quad (3)$$

where $\mathcal{C}$, K represent the retrieved contexts and the number of them respectively.

3.2 Baselines

Since our primary focus is on exploring how to use open-sourced LLMs to infer missing knowledge without fine-tuning, thereby improving the accuracy of open-domain question answering and the

[1]Chatbot Arena LLM Leaderboard: https://lmarena.ai/

[2]Instructions for using BGE series models on Hugging Face: https://huggingface.co/BAAI/bge-large-en

[3]It is worth noting that we choose DeepSeek-v3 as it is much cheaper with the MIT license and the difference of classified output is relatively small compared with DeepSeek-R1 (DeepSeek-AI et al., 2025)

[4]Limited by our budget, we randomly sample 256 data from NQ validation set and 128 consultation data from Huatuo-26M for evaluation.

TopK	Knowledge Base	EM	F1	Precision	Recall	Context_Recall_OrigQ	Context_Recall_MisQ	Miss_rate
-	DirectGen	22.27	17.11	18.86	20.59	-	-	-
TopK=2	Wiki	26.17	20.19	20.12	27.19	-	-	-
	QA Pairs	30.08	24.43	25.43	31.45	-	-	-
TopK=4	Wiki	29.69	21.71	21.63	29.13	-	-	-
	Misk Wiki+Wiki	30.47	22.29	23.09	26.79	54.79	57.48	45.7%
	QA Pairs	31.25	24.56	25.64	31.26	-	-	-
	Misk QA Pairs + QA Pairs	32.81	25.65	27.8	29.42	36.29	37.11	53.91%
	2-way retrieval	31.64	23.87	26.77	25.48	-	-	-
	Misk Wiki+QA Pairs	32.42	24.85	26.37	29.19	65.13	69.09	48.44%
	Misk QA Pairs+Wiki	35.16	**27.20**	**28.64**	**32.19**	36.82	62.66	52.34 %
	Merged KB	35.55	26.75	28.13	30.63	-	-	-
	Misk Merged KB	**37.11**	25.74	27.61	28.33	49.65	51.44	46.66%

Table 1: A comparison results from different baselines on the Natural Questions development set. The framework retrieves from different knowledge bases (i.e. Wikipedia(Wiki), history QA pairs(QA Pairs), or a merged knowledge base (Merged KB)). Misk {A+B} denotes the process of retrieving from knowledge base A and B with original query and missing knowledge query respectively, hence Context_Recall_OrigQ and Context_Recall_MisQ specifically refer to the recall scores evaluated using the ground truth context after retrieving the TopK contexts with standard RAG and missing knowledge RAG under this scenario. Miss_rate denotes the proportion of instances that require missing knowledge retrieval. The symbol "-" indicates that the result is not available. We **bold** the best performance.

Dataset	Knowledge Base	#Text Chunks
Natural Questions	Wikipedia	4,760,729
	History QA Pairs	2,500,931
	Merged KB	7,261,660
Huatuo-26M	Wikipedia	231,528
	History QA Pairs	8,802,233
	Merged KB	9,033,761

Table 2: Text chunk statistics about the knowledge bases across different datasets.

completeness of retrieved texts, while also considering the distribution gap between queries and the knowledge base, we primarily consider the following baselines:

DirectGen, where the answer is directly generate by prompting LLMs with parametric knowledge. It is worth noting that we use Llama3.3-70B and Qwen2.5-70B-Instruct as backbone LLM for NQ and Huatuo-26M datasets.

RAG with different knowledge bases, which evaluates the standard RAG framework using multiple knowledge bases (i.e. wikipedia, history QA pairs or the merged KB) to assess the impact of gap between query and different bases, which might affect retrieval and generation performance. We comprehensively explore with different retrieval combinations, and then pack into prompt to generate the final answer: (1) Simply retrieve from wikipedia, history QA pairs or the merged knowledge base with TopK contexts. (2) 2-way retrieval, denotes separately retrieve $\frac{TopK}{2}$ text chunks from wikipedia and history qa pairs knowledge bases.

RAG with Missing knowledge extends RAG by explicitly identifying and addressing missing knowledge with one-pass generation during retrieval, considering the inter connection among query, context, missing knowledge and its corresponding query, aiming to improve answer completeness and accuracy.

3.3 Natural Questions Results

As shown in Table 1, by analyzing the retrieved knowledge bases, we observed that using QA pairs as the knowledge source significantly improved the Exact Match (EM) and F1 scores compared to Wikipedia, with maximum increases of 3.91% and 4.24%, respectively. However, as the value of TopK increased, the performance gap gradually narrowed. For instance, at TopK=4, the improvements reduced to 1.56% (EM) and 2.85% (F1). This phenomenon can be attributed to the distribution gap between queries and knowledge bases. When the number of retrieved documents is small, the likelihood of recalling relevant documents from the history QA pairs KB is higher than that from Wikipedia. However, Wikipedia chunks inherently contain more comprehensive information, leading to diminishing performance differences as TopK increases.

Under the same TopK setting, our method of supplementing missing knowledge retrieval outperformed the standard RAG framework, achieving maximum improvements of 1.56% (EM) and 1.09% (F1). This enhancement is attributed to the

TopK	Knowledge Base	ROUGE_1	ROUGE_2	ROUGE_L	BLEU_1	BLEU_2	BLEU_3	BLEU_4	Context_Recall_OrigQ	Context_Recall_MisQ
-	DirectGen	13.91	1.46	9.73	9.38	2.80	0.99	0.30	-	-
TopK=2	Wiki	15.06	1.84	10.15	12.01	3.77	1.38	0.48	-	-
	QA Pairs	15.20	1.78	10.11	12.81	3.86	1.53	0.56	-	-
TopK=4	Wiki	16.70	2.25	10.57	14.35	4.73	1.91	0.77	-	-
	Misk Wiki+Wiki	18.05	2.09	11.74	17.87	5.52	2.17	0.83	34.91	35.36
	QA Pairs	16.75	2.22	10.90	14.49	4.74	1.83	0.74	-	-
	Misk QA Pairs+QA Pairs	18.12	2.51	12.01	17.89	5.95	2.45	1.03	39.07	41.24
	2-way retrieval	17.54	2.12	11.40	17.24	5.23	2.04	0.71	-	-
	Misk Wiki+QA Pairs	17.94	2.42	11.89	17.91	5.85	2.26	0.76	37.35	41.73
	Misk QA Pairs+Wiki	17.66	**2.53**	11.70	18.03	6.06	2.40	0.84	39.41	41.80
	Merged KB	18.06	2.33	12.18	19.03	6.07	2.30	0.74	-	-
	Misk Merged KB	**18.63**	2.50	**12.38**	**19.13**	**6.31**	**2.70**	**1.21**	38.60	40.48

Table 3: A comparison results of ROUGE and BLEU score from different baselines on the Huatuo-26M medical consultation test set. The framework retrieves from different knowledge bases. Misk {A+B} represents the retrieval augmented with missing knowledge process involving knowledge base A and B. Context_Recall_OrigQ and Context_Recall_MisQ denote the recall scores evaluated for standard RAG and our missing knowledge RAG framework, respectively. The symbol "-" indicates unavailable results. We **bold** the best performance.

model's robust reasoning capability, which identifies missing knowledge and retrieves supplementary information, resulting in a 2.69% increase in context recall. Notably, incorporating the model's analysis of missing knowledge into the prompt indirectly contributed to generating more accurate answers. Nevertheless, due to the inherent distribution gap between queries and knowledge bases, retrieving from history QA pairs still yielded higher performance gains compared to Wikipedia, with improvements of 2.34% (EM) and 3.36% (F1).

Furthermore, to comprehensively investigate the impact of knowledge base selection and combination strategies—which may mitigate the distribution gap between user queries and knowledge bases—we explore a sequential retrieval approach integrating both wikipedia and QA pairs, with missing knowledge augmentation. As shown in Table 1, we observe that retrieving from QA pairs using the user query, followed by retrieving from wikipedia using the missing knowledge query, achieves improvements of 3.52% in Exact Match (EM) and 3.33% in F1 score compared to the 2-way retrieval approach. Additionally, this retrieval combination yields a 25.84% improvement in context recall compared to directly retrieving from QA pairs using the original user query. In contrast, retrieving first from wikipedia and then from QA pairs results in only a 3.96% recall improvement.

One potential explanation for this discrepancy is that wikipedia chunks inherently contain more comprehensive information, and the generated missing knowledge query may effectively mitigate the distribution gap with wikipedia. Another contributing factor could be the inherent distribution gap between the original query and wikipedia, as errors introduced during the initial retrieval from wikipedia may propagate, negatively impacting subsequent retrieval steps.

Finally, our framework achieves an Exact Match (EM) score of 37.11% and an F1 score of 25.74% by leveraging text chunks retrieved from a merged knowledge base that combines wikipedia and QA pairs, further enhanced through missing knowledge retrieval.

To summarize, consider the inherent mismatch between queries and knowledge bases, retrieve QA pairs are more effective than wikipedia knowledge base, and integrating missing knowledge significantly enhances performance, demonstrating the importance of dynamic knowledge integration.

3.4 Medical domain Results

To evaluate our missing knowledge retrieval framework on domain specific tasks, we further evaluate on Huatuo-26M, a chinese medical open-domain question answering dataset. As shown in Table 3, we discover that we could get better results with 0.14% ROUGE_1 and 0.8% BLEU_1 improvement with retrieved QA pairs with standard RAG pipeline. This could be attributed to the distribution gap between user query and knowledge base in the specific domain. The following is a case shows the distribution gap in medical consultation, where we retrieve text chunks from medical encyclopedia and consultant QA pairs respectively.

Furthermore, we investigate the effectiveness of missing knowledge retrieval within this framework. As demonstrated in Table 3, the performance exhibits notable improvements, with increases of 1.37% and 1.35% in ROUGE scores, as well as 3.4% and 3.52% in BLEU scores, when utilizing missing knowledge queries derived from QA pairs and Wikipedia, respectively. A plausible explanation for these enhancements is that the missing knowledge query retrieval mechanism retrieves a higher proportion of relevant text chunks, as evidenced by the 2.17% and 0.45% improvements in context recall compared with the standard RAG pipeline under the same TopK configuration.

> **Case study of distribution gap within Huatuo-26M**
>
> **Query:** As mentioned, I've been taking traditional Chinese medicine for qi deficiency but also engage in intense daily workouts that cause heavy sweating. Could this affect the medicine's efficacy?
>
> **Retrive from QA Pairs:** Q: In summer, even slight physical activity or mildly hot weather causes sweating on my upper body,... who diagnosed me with qi deficiency and prescribed a 5-day herbal treatment. What should I do? A: ...involves frequent sweating, which has become more pronounced after taking the prescribed herbal medicine. This could be due to the presence of qi-tonifying... temporarily stop taking the medication or wait until the weather cools down before resuming...
>
> **Retrieve from Wiki:** Efficacy and Effects of Aconite: Aconite can enhance myocardial contractility, increase heart rate,...

Finally, we conduct an evaluation of various knowledge base retrieval combinations. The results indicate that the highest ROUGE and BLEU scores, achieving 18.63% and 19.13% respectively, are attained when utilizing the merged knowledge base enhanced with missing knowledge retrieval. This approach demonstrates a 1.88% improvement in context recall compared with the standard RAG pipeline.

3.5 Benefits with missing knowledge retrieval

To further explain the role of missing knowledge retrieval. We offer fair comparison among different retrieval methods with the same number of text chunks before evaluation. Table 4 shows that compared with 2-way retrieval method, we could obtain 8.09% and 2.41% recall improvement with missing knowledge query retrieval from NQ and Huatuo-26M QA pairs knowledge base respectively. And we also obtain 1.93% and 1.7% recall improvement with missing knowledge query retrieval from NQ and Huatuo-26M wiki knowledge base respectively.

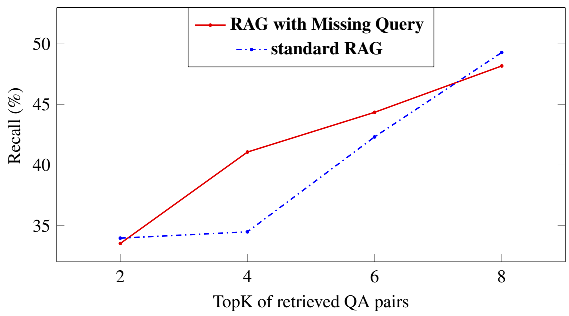
To further explain the role of missing knowledge retrieval, we provide a fair comparison among different retrieval methods by ensuring an identical number of retrieved text chunks prior to evaluation. As illustrated in Table 4, the missing knowledge query retrieval demonstrates significant improvements in recall compared to the 2-way retrieval method. Specifically, recall improvements of 8.09% and 2.41% are observed when utilizing the NQ and Huatuo-26M QA pairs knowledge bases, respectively. Similarly, 1.93% and 1.7% of recall improvements are achieved when employing the NQ and Huatuo-26M Wikipedia knowledge bases, respectively. These results underscore the effectiveness of missing knowledge retrieval in enhancing recall performance across diverse knowledge sources.

Dataset	Method	QA Pairs Recall	Wiki Recall
Natural Questions	2-way-retrieval	48.31	57.83
	Mis Wiki+QA pairs	56.4	57.12
	Mis QA Pairs+Wiki	48.51	59.76
Huatuo-26M	2-way-retrieval	32.87	27.31
	Mis Wiki+QA pairs	35.28	27.23
	Mis QA Pairs+Wiki	33.05	29.01

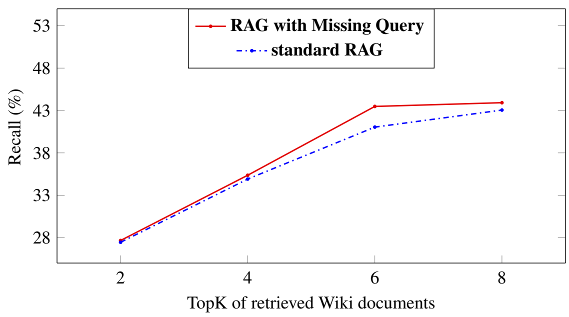
Table 4: Results of context recall for QA pairs and wiki documents across different methods. Notably, the number of retrieved text chunks is identical for evaluation.

3.6 Impact on TopK for missing knowledge retrieval

In order to detect the impact on context recall score across different TopK, we conduct domain specific experiments on Huatuo-26M with missing knowledge RAG and standard RAG. As shown in Figure 3, we find that RAG with missing knowledge query achieves better results, especially for retrieving from QA pairs. While smaller TopK shows different trends, which could be explained by failing to reason with limited contexts. And as TopK becomes larger, the gap of recall score between missing knowledge RAG with standard RAG becomes smaller, which could be explained by enough contexts to answer the user query. To systematically evaluate the influence of varying TopK values on context recall performance, we conducted domain-specific experiments utilizing the Huatuo-26M dataset, comparing the effects of missing knowledge RAG and standard RAG frameworks. Our findings indicate that the missing knowledge RAG approach yields superior results, particularly retrieving from QA pairs. Notably, smaller TopK values exhibit divergent trends, this might attribute to the challenges of reasoning effectively with limited contextual information. Conversely, as TopK increases, the disparity in recall scores between missing knowledge RAG and standard RAG diminishes, this could be rationalized by the provision of sufficient contexts to adequately address user queries.



(a) Retrieve from history QA pairs.



(b) Retrieve from wikipedia knowledge base.

Figure 3: Context Recall scores of RAG variants across TopK values, comparing missing knowledge query (retrieving one text chunk from knowledge base using user query followed by missing knowledge query) with standard RAG (directly retrieving two text chunks from knowledge base) at TopK=2.

4 Related Work

Iterative RAG with missing knowledge To enhance reasoning ability within the RAG pipeline, there are primarily two approaches: One relies on teaching model how to think utilizing internal parameter knowledge. Qian et al. (2024) proposes to use parametirc memory module to generate context cues before retrieval, in order to bridge the gap between query and knowledge base. Islam et al. (2024) proposes to use hybrid adaptive retrieval to effectively determine relevant and supported contexts. Another line lies on optimizing reasoning through external process with powerful LLM such as GPT-3.5 to generate follow up thinking steps (Press et al., 2023; Yao et al.). Kim et al. (2024) propose a RAG system with query decomposation and expansion. Jiang et al. (2024) and Wang et al. (2024) utilizes Monte-Carlo Tree Search to find optimal chunk combinations. Wang et al. (2025) propose to extract missing information and generate query within a two-pass pipeline, whereas Trivedi et al. (2023) employ the CoT sentence to perform iterative retrieval.

Inspired by their works, we build an efficient, single-pass way to generate formatted missing knowledge and its query, which could be interconnected with open-sourced frozen LLMs. Furthermore, we utilize reasoning ability within LLM without training effectively.

Query generation in RAG To improve the accuracy of retrieved contexts, further enhance the accuracy of output responses, and mitigate the distribution gap between queries and the knowledge base, the quality and form of queries are crucial. Many efforts focus on query rewriting to boost the accuracy of question answering. (Li et al., 2024) propose to train a unified model to generate Fine-grained clues and evidence at the same time. (Ma et al., 2023) propose a trainable rewrite-retrieve-read framework.

Inspired by their progress, we further explore the distribution gap between query and different knowledge base. And build an efficient missing knowledge query generation framework with frozen LLMs.

5 Conclusion

We comprehensively explore the distribution gap between query and text chunks with in knowledge bases by leveraging multi-source knowledge bases in a real-world question answering system. Furthermore, to mitigate the missing knowledge problem, we propose a Missing Knowledge RAG Framework leveraging Chain-of-Thought (CoT) reasoning, which introduces one-pass efficient missing knowledge query generation. By explicitly modeling the interconnections among the query, retrieved contexts, missing knowledge, and its corresponding query, our approach enhances the relevance and completeness of retrieved knowledge considering the above distribution gap as well without additional training.

Limitation

Our work primarily focuses on addressing the distribution gap between queries and knowledge bases by exploring different knowledge sources, rather than optimizing the retrieval mechanism itself. To ensure a fair comparison, we employ a widely-used dense retriever, leaving the exploration of advanced retrieval techniques for future work.

Ethical Statement

This study complies with ethical standards by using open sourced data and avoiding sensitive personal information. Our research improves accuracy and reliability for the widely used QA system, ensuring no harm to individuals or communities.

References

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*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik

Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244, Miami, Florida, USA. Association for Computational Linguistics.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *ArXiv*, abs/2412.12881.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Dongkyu Kim, Byoungwook Kim, Donggeon Han, and Matouvs Eibich. 2024. Autorag: Automated framework for optimization of retrieval augmented generation pipeline. *ArXiv*, abs/2410.20878.

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.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *Preprint*, arXiv:2305.01526.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2024. Retrollm: Empowering large language models to retrieve fine-grained evidence within generation. *Preprint*, arXiv:2412.11919.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Conference on Empirical Methods in Natural Language Processing*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

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.

Keheng Wang, Feiyu Duan, Peiguang Li, Sirui Wang, and Xunliang Cai. 2025. LLMs know what they need: Leveraging a missing information guided framework to empower retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2379–2400, Abu Dhabi, UAE. Association for Computational Linguistics.

Ziting Wang, Haitao Yuan, Wei Dong, Gao Cong, and Feifei Li. 2024. Corag: A cost-constrained retrieval optimization system for retrieval-augmented generation. *ArXiv*, abs/2411.00744.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

A Prompt for one-pass missing knowledge query generation and final answer generation

 # Natural Questions
You are an knowledge expert and proficient in JSON formats.
Instructions Given a question, retrieval context and prediction, your task is to determine if the context information is sufficient to answer the question or if additional knowledge is required, rewrite with short query
which is closely related to the main entity in the query.
Missing Knowledge Criterion:
- The Incomplete answer may not cover all key-points and details of the question.
- The answer contain wrong information, and is inconsistant with contexts or facts.
- The answer is fuzzy or irrelevant.
Requirements
Proceed step by step as follows:
- First, based on the question and first round answer, determine if the answer is imcomplete or the retrieval text lacks the knowledge required to answer the question;
- Second, if true, list the missing knowledge and generate the corresponding query;
- Third, output reason and a dict containing the fields 'thought', 'judge', 'missing_knowledge',
and 'query', adhering strictly to the JSON format;
- Query should be short, precise and closely related to query;
Output Format
‘‘‘json
{"thought": str, "judge": bool, "missing_knowledge": list, "query": list}
‘‘‘
Key-value descriptions:
- thought: analysis on the correctness ans relavance of the retrived context;
- judge: if the knowledge is missing;
- missing_knowledge: list of missing knowledge;
- query: list the corresponding query of missing knowledge;
Query
{query}
Context
{contexts}
First Round Answer
{first round prediction}
Output

Figure 4: Prompts for missing knowledge query generation with NQ dataset.

 # Huatuo-26M (Translated)

Translate the following prompt into English while preserving the format:
You are a medical knowledge expert proficient in Chinese and JSON format.
Requirements
Given a query question, retrieved text, and first-round response, analyze step by step whether medical knowledge is missing.
Output markdown-embedded JSON data containing the missing knowledge and rewritten corresponding query questions.
- First, based on the query question, retrieved text, and first-round response, analyze and determine if the retrieved text lacks relevant symptoms and treatment plans required to answer the symptoms;
- Second, if True, list the missing knowledge and generate corresponding queries;
- Third, output data containing judge, missing_knowledge, and query, strictly adhering to JSON format requirements;
Output Format
{"thought": str, "judge": bool, "missing_knowledge": list, "query": list}
##Format Explanation
- thought: brief analysis process;
- judge: whether knowledge is missing (True if missing, False otherwise);
- missing_knowledge: missing knowledge in list format;
- query: corresponding query questions in list format;
Missing Knowledge Example
“‘json
{"thought": "xxx", "judge": true, "missing_knowledge": ["causes of migraines", "methods to alleviate migraines"],
"query": ["causes of migraines", "How to alleviate migraines?"] } “‘
No Missing Knowledge Example
“‘json
{"thought": "No missing knowledge", "judge": false, "missing_knowledge": [], "query": []}
“‘
Query
{query}
Context
{contexts}
First Round Answer
{first round prediction}
Output

Figure 5: Prompts for missing knowledge query generation with Huatuo dataset.

Natural Questions
Instructions
Please carefully read the following context and briefly answer the question with essential keywords or short phrases based on the context.
Requirements
- Ensure that your answer is highly relevant to the provided contexts and missing knowledge contexts.
- The answer should be short, concise, and as accurate as possible without explanation.
- If it is not mentioned in the context, briefly answer with your own knowledge.
Example:
Context
Documents
Document[1]: Google was founded by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University.
History QA pairs
Q: Who is the original CEO of Google?
A: Larry Page and Sergey Brin.
Question
who founded google?
Answer
Larry Page and Sergey Brin
Context
{contexts}
Missing Knowledge Context
{thought}
{context retrieved with missing knowledge query}
Question
{query}
Answer

Huatuo-26M
You are a medical Q&A assistant, and your task is to provide brief advice based on the retrieved text,
missing knowledge, and the example format in response to the question.
Requirements
Use natural language in your response, do not format as a list Keep the response concise, do not repeat
the question content Ensure the response is relevant to the question, retrieved text, and missing knowledge.
Example
Context
Historical Q&A Record
Q: Why do I feel hungry after a nap?
A: If you are chronically sleep-deprived, you may feel hungry because lack of sleep increases cortisol and
ghrelin levels in the body. Ghrelin is the hunger hormone that stimulates your appetite.
Missing Knowledge
Document
Document[1]: Patients with hyperthyroidism have increased thyroid hormone secretion, which speeds up the body's metabolism.
Food is quickly consumed, leading to symptoms like increased
hunger and food intake. In addition to increased appetite, these individuals may also experience
rapid heart rate, excessive sweating, hand tremors, weight loss, irritability, and menstrual irregularities.
Question:
Why do I feel hungry after taking a nap at noon?
Answer:
Based on your description, hyperthyroidism should be ruled out. It is recommended to visit an endocrinology clinic
and undergo blood tests for T3, T4, and TSH to exclude hyperthyroidism.
Ensure adequate sleep and maintain a positive mood, then treat according to the specific condition. Take this seriously.
Context
{context}
Missing Knowledge Context
{thought}
{context retrieved with missing knowledge query}
Question
{query}
Answer

Figure 6: Prompts for answer generation combined with CoT reasoning process.