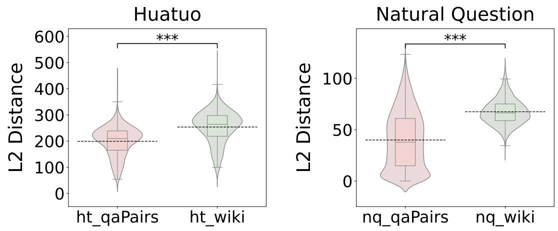
Reasoning Enhanced Missing Knowledge Retrieval Augmented Generation Framework for Domain Specific Question Answering

Yuanjun Shi[1], **Zhaopeng Qiu**[1*]
[1]Individual Researcher
`{switchevans, zp10211059}@gmail.com`

Abstract

Retrieval Augmented Generation (RAG) framework mitigates hallucinations in Large Language Models (LLMs) by integrating external knowledge, yet faces two critical challenges: (1) the distribution gap between user queries and knowledge bases in a specific domain, 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 reasoning-based 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 an end-to-end inference manner while aligning query knowledge distributions, and integrates reasoning traces into answer generation. Experiments on open-domain medical and general question answering (QA) datasets demonstrate significant improvements in context recall and answer accuracy. Our approach achieves effective knowledge supplementation without additional training, offering enhanced interpretability and robustness for real-world QA applications.

Question	May I ask if working in a battery factory during pregnancy has any impact on the fetus?
Wikipedia	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...
LLM Response	Working in... may indeed expose you to various harmful substances. I recommend consulting a doctor as soon as possible... reduce exposure to harmful substances...
QA Pairs	Q: There is an electronics company that produces lithium batteries ... Will this job affect my pregnancy... A: Medical Analysis: During the production... generates some powdered airborne particles, volatile gases, such as NMP ... I recommend that... not directly exposed ... poses risks to your own health. Take proper protective measures or you could change your job .
LLM Response	Hello... 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 ... attend regular prenatal check-ups ...



(a) Huatuo dataset results (b) Natural Question results

Figure 1: *(top)* Example of retrieved text chunks and corresponding responses from Huatuo, where relevant texts are highlighted in red , LLM responses in blue , showing QA pairs are more consistent with user queries in terms of intents and semantics. *(Bottom)* Euclidean distances L_2 between queries and their top-1 ranked retrieved text chunks across knowledge bases demonstrate semantic proximity. Each subplot presents paired distributions: query-to-QA pair distances *(left)* versus query-to-Wikipedia passage distances *(right)*, with dashed lines indicating mean values. Lower L_2 metrics signify higher semantic similarity.

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), demonstrates 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, the Retrieval Augmented Generation (RAG) framework emerges as a promising solution for knowledge-intensive tasks (Lewis et al.,

*Corresponding author.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1361–1379
December 20-24, 2025 ©2025 Association for Computational Linguistics

2020), integrating external knowledge bases with LLMs to enhance answer reliability. While the standard RAG pipeline retrieves contextually relevant documents to ground LLM outputs, two critical challenges persist:

(1) **The semantic discrepancies** between user queries and heterogeneous knowledge bases, which undermine retrieval relevance. As illustrated in the Figure 1 *(top)*, a significant disparity exists between colloquial or non-professional user descriptions and the standardized terminology characteristic of formal medical literature. Besides, as indicated by the average distance (L_2) between user queries and the top-1 retrieved text chunks in Figure 1 *(below)*, the Question Answer (QA) pairs more closely align with the user's query in terms of both semantics and intent, making them valuable resources for addressing user queries.

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). Although these methods demonstrate partial success, they overlook the historical QA pair, which is a naturally aligned knowledge source that inherently matches user query distributions in the domain-specific scenario.

(2) **Missing knowledge** could be attributed to the limited understanding of such queries and thus requires precise clarification of the actual user intent, particularly in scenarios where retrieved contexts fail to fully cover the knowledge required to address queries. Existing works lie in two paths. One line of works propose 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., 2025a). The other line of works propose to generate evidence and criticism with special tokens in one single pass adaptively (Islam et al., 2024; Asai et al., 2024).

Although existing approaches achieve promising results, the rapid progress in the reasoning capabilities of LLMs (DeepSeek-AI et al., 2025) suggests richer opportunities with reasoning chains. Given a user query, retrieved contexts, and preliminary predictions, a reasoning module could analyze the missing knowledge and follow-up queries step by step, while analyzing whether inconsistencies exist in contexts, such as knowledge conflicts and hallucinations simultaneously. Moreover, most prior works treat the detection of missing knowledge and the assessment of query relevance as two separate procedures. This decoupled design disregards the intrinsic interdependence among the user query, retrieved contexts, missing knowledge, and the corresponding follow-up queries, which might lead to cumulative error propagation, and ultimately achieve suboptimal performance.

In this work, we propose a reasoning-enhanced missing knowledge RAG framework that systematically addresses both challenges through a unified pipeline with an end-to-end manner. Unlike prior methods, our approach explicitly considers the intrinsic relationship among the query, retrieved contexts, missing knowledge, and follow-up queries with a reasoner module, utilizing the reasoning content to verify knowledge conflicts or hallucinations. Furthermore, the generated reasoning traces are seamlessly incorporated into the final answer generation process, ensuring both interpretability and accuracy.

- We propose a reasoning enhanced end-to-end RAG framework that generates queries for missing knowledge retrieval and leverages the reasoning content to verify knowledge, thereby improving generation accuracy.
- We comprehensively explore methods with heterogeneous knowledge bases to bridge the gap between colloquial user queries and formal passages without additional training.
- We conduct extensive experiments on two open-domain question answering datasets, evaluating on both general and domain-specific scenarios.

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

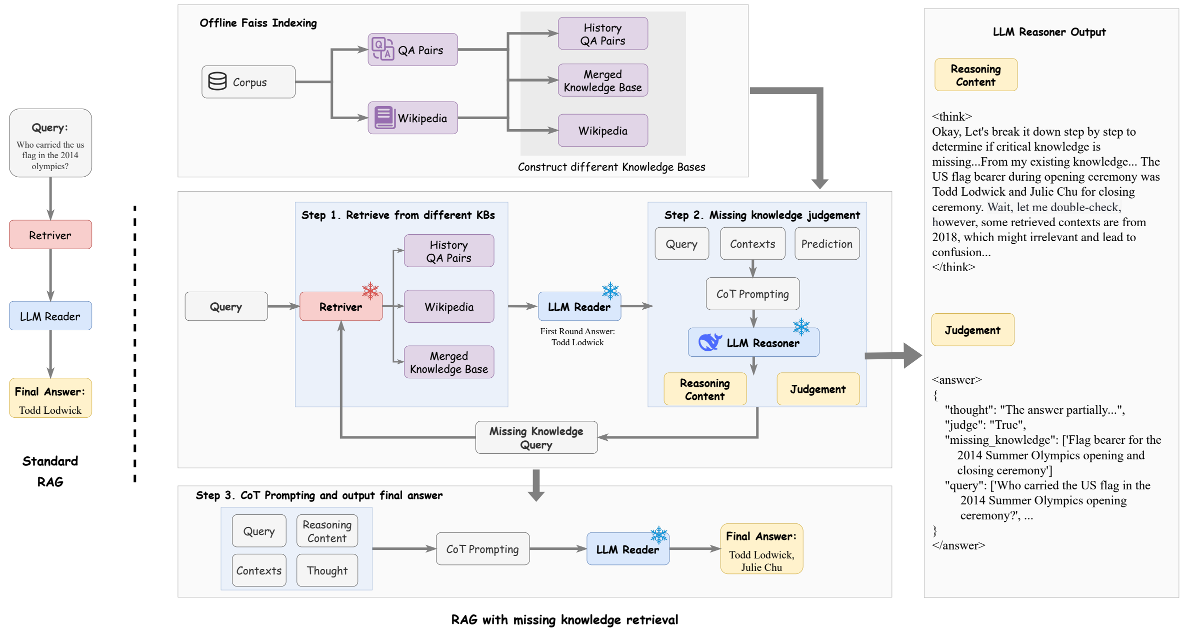


Figure 2: Illustration of our Missing Knowledge RAG framework. Our pipeline first retrieves from knowledge bases and prompts open-sourced LLMs to give a draft answer. Second, LLM needs to decide whether any knowledge is missing. Different from the standard RAG pipeline, the LLM needs to generate missing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

are closely related to user query, and then generate a final answer $\mathcal{Y}$ based on query and grounded knowledge. The final goal of a RAG system is to ensure the comprehensive and precise retrieval of contexts needed to address the query, thereby generating the response aligned with ground truth $\mathcal{A}$.

2.2 Reasoning Enhanced RAG with Missing Knowledge Framework

In our proposed framework, we initially retrieve top-K grounded knowledge text chunks $\mathcal{C}$, sourced from the offline indexing process, to formulate a preliminary answer $\mathcal{Y}_1$. Then, we employ a reasoner LLM to judge if there exists any missing knowledge in the reasoning process $\mathcal{R}$. If deficiencies in knowledge are identified, iterative retrieval is performed with the generated queries corresponding to missing knowledge. Finally, we prompt the reader LLM to produce the final response $\mathcal{Y}_{final}$.

Retrieve from Different Knowledge Bases To mitigate the distribution gap between queries and knowledge bases shown in Figure 1, previous studies propose to utilize query rewriting (Ma et al., 2023), which requires efforts in data annotation. We propose to retrieve from historical QA Pairs, which contain more colloquial descriptions and exhibit closer semantic and intent alignment with user queries, especially in domain-specific scenarios. And we further investigate retrieval from various knowledge sources (i.e., Wikipedia, historical QA pairs, or the merged knowledge base).

As presented in Figure 2, we first employ a dense retriever to obtain top-K text chunks using the encoded query embedding. Subsequently, this approach enables us to acquire contextual information characterized by diverse semantic structures.

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

where $\mathcal{D}$ represents the retrieved knowledge base with text chunks, $\mathbf{V}$ and $\mathcal{C}$ refer to encoded vector set processed through offline indexing and top-K selected contexts according to L2 metric, respectively.

Then we could get the first round answer with the LLM reader:

$$
\mathcal{Y}_1 = LLM_\theta(x_i | \mathcal{Q}, \mathcal{C}, x_{<i}, i = 1, ..., t), \tag{2}
$$

where x_i denotes the i-th token during generation.

Reasoning Enhanced Missing Knowledge Judgment However, the retrieved contexts may not necessarily contain all the knowledge required to answer the query, and they also contain knowledge conflicts or hallucinations. Previous works (Wang et al., 2025a) therefore propose a pipeline approach separately performing missing information retrieval and knowledge entailment classification, which might cause error propagation or misalignment within different modules. Qian et al. (2024) proposes to train a specific generative module, which incurs substantial manual efforts.

To mitigate these limitations, we introduce a single-pass reasoning framework that simultaneously generates missing knowledge and formulates the requisite follow-up queries. The judging process comprises two principal components:

- **Reasoning Content** This exploits the model's reasoning capacity to verify information within given contexts, thereby guiding the subsequent retrieval of any missing knowledge.
- **Missing Knowledge Judgment** This module accurately identifies knowledge gaps and constructs the necessary follow-up queries step by step, considering the intrinsic interconnections among the initial query, retrieved contexts, and the missing knowledge,

Specifically, for example, given query $\mathcal{Q}$: `Who carried the US flag in the 2014 Olympics?`, retrieved contexts $\mathcal{C}$ and first round prediction $\mathcal{Y}_1$, we prompt the reasoner model to generate a judgment about the missing knowledge and reasoning content $\mathcal{C}_{reason}$, which contains step-by-step verification of factual information. The JSON formatted answer includes reflective thought about the first round answer, judgment, missing knowledge cues, and the corresponding queries "`who carried... opening ceremony?`" and "`who carried... closing ceremony?`" aligned with retrieval contexts. The detailed prompts for missing knowledge query and answer generation are presented in the Appendix 4.

Finally, we utilize the aligned missing knowledge query to retrieve missing information from the specified knowledge base, subsequently appending it to $\mathcal{C}$. Following this, we apply a straightforward deduplication function using MD5 hashing to remove redundant text chunks. Then, we consolidate them with part of the reasoning information into a structured prompt the reader to generate the final answer:

$$\mathcal{Y}_{final} = LLM_\theta(x_i|\mathcal{Q}, \mathcal{C}, \mathcal{C}_{reason}, x_{<i}), \quad (3)$$

including query $\mathcal{Q}$, contexts with missing knowledge $\mathcal{C}$, the reasoning content $\mathcal{C}_{reason}$ and reflective thought about first round answer.

3 Experiment

3.1 Experimental 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 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 out proposed method, which contain a large volume of high-quality colloquial QA pairs, reflecting the natural distribution of user interactions across various real-world QA systems.

We utilize the open-sourced Llama3.3-70B (Grattafiori et al., 2024) and Qwen2.5-70B-Instruct (Qwen et al., 2025) as our reader LLM in the framework for English and Chinese benchmark, respectively, owing to their success on the LMSys leaderboard under specific categories[1]. We use DeepSeek-R1 (DeepSeek-AI et al., 2025) as a reasoner LLM owing to its powerful reasoning ability. We employ bge-en-large (Xiao et al., 2023) and bge-large-zh-v1.5 to encode QA Pairs and Wikipedia chunks, following instructions on the website.[2] For the purpose of similarity searching, we implement the Faiss index (Johnson et al., 2019), specifically the IndexFlat with L2 distance metrics.

For NQ evaluation, we use all the documents provided in the NQ dataset directly without any HTML tags to construct a 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 4.8M text chunks, followed by a deduplication process with MD5 hash, shown in Table 2. To provide a fair comparison, we also leverage all the QA pairs in the training set of NQ

[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

Methods	Natural Questions				Huatuo-26M						
	EM	F1	Precision	Recall	ROUGE_1	ROUGE_2	ROUGE_L	BLEU_1	BLEU_2	BLEU_3	BLEU_4
DirectGen	22.27	17.11	18.86	20.59	13.91	1.46	9.73	9.38	2.80	0.99	0.30
CoT	24.22	15.99	19.77	17.48	13.9	1.81	8.57	11.13	3.6	1.44	0.53
MIGRES	30.00	17.20	18.57	19.78	17.19	1.77	12.72	14.70	4.67	1.69	0.48
STD-RAG Wiki	26.17	20.19	20.12	27.19	15.06	1.84	10.15	12.01	3.77	1.38	0.48
-top-K=4	26.69	21.71	21.63	29.13	16.70	2.25	10.57	14.35	4.73	1.91	0.77
STD-RAG with QA Pairs	30.08	24.43	25.43	31.45	15.20	1.78	10.11	12.81	3.86	1.53	0.56
-top-K=4	31.25	24.56	25.64	31.26	16.75	2.22	10.90	14.49	4.74	1.83	0.74
STD-RAG with 2-way retrieval	31.64	23.87	26.77	25.48	17.54	2.12	11.40	17.24	5.23	2.04	0.71
STD-RAG with Merged KB	35.55	26.75	28.13	30.63	18.06	2.33	12.18	19.03	6.07	2.30	0.74
Misknow-RAG with Wiki	32.03	21.31	23.76	23.56	18.32	2.05	11.62	18.28	5.47	1.86	0.54
Misknow-RAG with QA Pairs	33.20	25.22	28.22	27.16	18.78	2.43	12.18	18.88	6.02	2.19	0.82
Misknow-RAG with Wiki+QA Pairs	33.20	24.45	27.03	27.22	18.06	2.38	11.54	18.22	5.92	2.15	0.79
Misknow-RAG with QA Pairs+Wiki	35.94	**27.27**	**29.55**	**30.02**	18.44	2.45	11.70	18.51	6.12	2.48	0.90
Misknow-RAG with Merged KB	**41.41**	26.80	28.83	27.78	**18.84**	**2.52**	**12.64**	**19.61**	**6.43**	**2.82**	**1.30**

Table 1: A comparison of results from different baselines on the NQ and Huatuo test set. Our framework retrieves from different knowledge bases (i.e., Wikipedia(Wiki), history QA pairs(QA Pairs), or a merged knowledge base (Merged KB)). The symbol "-" indicates that the result is not available. We **bold** the best performance.

and prompt the LLM to summarize documents into pseudo QA pairs in the test set. This process yields 128K 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 7.3M text chunks, followed by a similar deduplication process.

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 QA pairs. Furthermore, we merged all the encyclopedias with consultant records, getting 9M text chunks for the merged knowledge base.

During evaluation, we use normalized exact match (EM) and word-level F1-score to compare the final prediction with the ground truth. For the medical open-domain QA task, we use ROUGE and BLEU scores to evaluate. As the ground-truth QA pairs are not available across most datasets, we apply the LLM-as-a-Judge paradigm (Gu et al., 2025) to comprehensively evaluate if the retrieved contexts contains all the necessary information to answer user query. Specifically, we develop the context recall metric with DeepSeek-V3[3] (i.e. DeepSeek (DeepSeek-AI et al., 2024)[4]). First, we prompt DeepSeek-V3 with $\mathcal{Q}$ and $\mathcal{A}$ to get labeled ground truth contexts $GT_{\mathcal{C}}$, which contain all the necessary knowledge responding to the user query.

[3]It is worth noting that we choose DeepSeek-V3 as it is much cheaper and classified outputs are relatively small compared with DeepSeek-R1

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

Then, we further prompt it to independently judge if the retrieved context could be attributed to the $GT_{\mathcal{C}}$. 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:

$$context\_recall = \frac{\sum_{i=1}^{|\mathcal{C}|} \mathbb{1}_{attributed[i]}(\mathcal{C}_i)}{|GT_{\mathcal{C}}|}, \quad (4)$$

where $\mathcal{C}$ represent the retrieved contexts.

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.

3.2 Baselines

Since our primary focus is on exploring the distribution gap between queries and different knowledge bases, alleviating missing knowledge issues, thereby improving the accuracy of open-domain question answering and the completeness of retrieved contexts. We primarily consider the following category of baselines: (1) **DirectGen**, whose answer is directly generated by prompting the reader LLM. (2) **CoT** (Wei et al., 2022) which generate responses in a chain-of-thought paradigm following "Let's think step by step". (3) **MIGRES**, Wang et al. (2025a) introduce a state-of-the-art pipeline-based framework for determining missing information and its entailment. To ensure a fair comparison, we employ the identical backbone LLM. (4) **STD-RAG**, which is the standard RAG framework using various knowledge bases (i.e., Wikipedia, QA pairs, or the merged KB), which might affect retrieval and generation performance due to the distribution gap. This includes different retrieval combinations: (i) Simply retrieve from Wikipedia, history QA pairs, or the merged knowledge base with top-K contexts. (ii) 2-way retrieval, denotes separately retrieve $\frac{top-K}{2}$ text chunks from Wikipedia and QA pairs. (5) **Misknow-RAG** extends RAG by explicitly identifying missing knowledge and checking knowledge conflicts simultaneously with a reasoner LLM during retrieval and generation, considering the interconnection among query, contexts, missing knowledge, and its corresponding query.

3.3 Main Results

As shown in Table 1, we observe that our approach, which leverages explicit reasoning based on missing knowledge retrieval, achieves gains of 11.41% in EM and 1.65% in ROUGE-1 on both general and medical QA benchmarks, outperforming the current state-of-the-art.

Furthermore, compared with the standard RAG pipeline, our reasoning-based missing knowledge retrieval method yields a 5.34% EM and 1.62% ROUGE-1 improvement when sourcing from Wikipedia, and a 1.95% EM and 2.03% ROUGE-1 improvement when sourcing from QA pairs, under the same retrieval budget. As presented in the Appendix A, this might be attributed to the effective missing knowledge retrieval while verifying information during reasoning (Dhuliawala et al., 2024).

Moreover, replacing Wikipedia with QA pair retrieval delivers additional gains of 3.91% EM and 4.24% F1 on the NQ dataset, and 0.14% ROUGE-1 and 0.80% BLEU-1 on Huatuo. This improvement could be attributed to the reduced semantic gap between the query and knowledge bases, as QA pairs tend to be more colloquial and therefore more closely aligned with user intents. As shown in Table 7, compared with the medical passages retrieved from Wikipedia, QA pairs more directly address the user's query concerning "`Ejiao syrup`" and "`improving my condition`" and provide more detailed therapeutic recommendations, such as "`do more exercise`" or "`take Astragalus granules`". Interestingly, we also observe that our approach improves at most 4.3% EM and 0.9% ROUGE-1 score compared with the naive 2-way retrieval methods, which implies the effectiveness of our missing knowledge retrieval in supplementing critical information and thus enhancing overall accuracy.

Finally, by leveraging a merged knowledge base comprising both Wikipedia passages and QA pairs, our framework attains an EM score of 41.41% and a ROUGE-1 score of 18.84%. This performance gain could be attributed to the enhanced richness of the knowledge base, which facilitates more effective retrieval of relevant passages.

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 3: Results of context recall for QA pairs and Wikipedia chunks across different methods on NQ and Huatuo. Notably, the number of retrieved text chunks is identical for evaluation.

3.4 Benefits with Missing Knowledge Retrieval

In this subsection, we explore the role of missing knowledge retrieval with context recall described in Eq.4 under two settings: 2-way retrieval, where we calculate metrics for QA pairs and Wiki chunks separately, and retrieval from the same source, where we compare naive context recall metrics.

Specifically, under the standard 2-way retrieval setting, two passages are drawn from Wikipedia and two from QA pairs. In contrast, in the missing knowledge retrieval setting (i.e., Mis Wiki + QA pairs in Table 3), we first retrieve two passages from Wikipedia and then utilize missing knowledge retrieval to extract the remaining two from the QA pairs. Table 3 shows that compared with the 2-way retrieval method, we could obtain 8.09% and 1.93% context recall improvement with missing knowledge retrieval from QA pairs and Wiki chunks for the NQ dataset, respectively. And we also obtain 2.41% and 1.7% context recall improvement retrieved from QA pairs and Wiki chunks for the Huatuo dataset, which directly improves the relevancy of retrieved passages within the reasoning process. More generally, Table 4 also presents an overall performance gain in recall.

More generally, Table 4 also presents an overall performance gain in recall while retrieving from the same source. To be specific, we evaluate the recall of retrieved contexts between the original query and the missing knowledge query with the same number of retrieved text chunks. Our approach improves 2.69% and 0.82% recall while utilizing Wiki and QA pairs on NQ, respectively, and 0.45% and 2.17% on Huatuo.

Methods	Natural Questions		Huatuo-26M	
	OrigQ	MisQ	OrigQ	MisQ
Wiki	54.79	57.48	34.91	35.36
QA Pairs	36.29	37.11	39.07	41.24

Table 4: Evaluation on the retrieved contexts utilizing original query and the missing knowledge queries with recall scores across different datasets.

3.5 Impact of the Reasoning Enhanced Method

In this subsection, we conduct ablation studies to demonstrate the effectiveness of of reasoning framework. As shown in Table 5, the experimental results indicate that leveraging reasoner models to generate chain-of-thought reasoning content significantly improves the response accuracies across all retrieval settings. This improvement can be attributed to the generated reasoning chain might incorporate factual knowledge, validating the correctness of the retrieved passages or the explicit analysis of the missing knowledge, which yields more precise queries.

Table 6 presents an example that illustrates the above analysis in detail. Considering the question "`Who has the most all-star mvp awards?`", the reasoner model generates "`the provided contexts don't explicitly list the number of awards each player has`" and "`players like Kobe Bryant or LeBron James could have more`". By incorporating such external knowledge, the LLM reader identifies missing knowledge such as "`Confirmation of the current record for All-Star MVP awards`" and generates a more accurate response "`Bob Pettit, Kobe Bryant, LeBron James` compared with the first-round answer.

Table 6 furnishes a detailed example that elucidates the foregoing analysis. Taking the question "Who has the most All-Star MVP awards?" as an illustrative case, the reasoner model's missing-knowledge analysis observes that "the provided contexts do not explicitly list the number of awards each player has" and further conjectures that "players like Kobe Bryant or LeBron James could have more." By incorporating such external knowledge, the model successfully identifies the precise information gap—namely, confirmation of the current record for All-Star MVP awards.

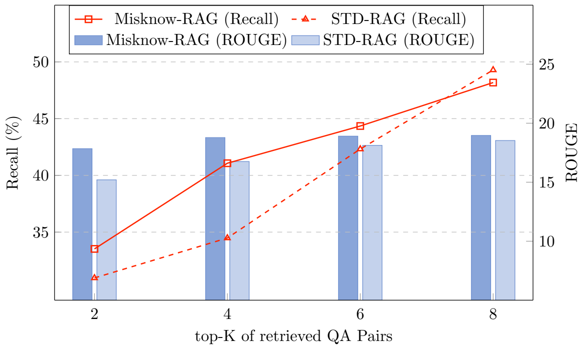
Methods	Natural Questions		Huatuo-26M	
	EM	F1	ROUGE_1	BLEU_1
Wiki	32.03	21.31	18.32	18.28
w/o reasoning	30.47	21.29	18.05	17.87
QA Pairs	33.20	25.22	18.78	18.78
w/o reasoning	32.81	24.65	18.12	17.89
Wiki+QA Pairs	33.20	24.45	18.06	18.22
w/o reasoning	32.42	23.85	17.94	17.91
QA Pairs+Wiki	35.94	27.27	17.92	18.18
w/o reasoning	35.16	27.20	17.66	18.03
Merged KB	41.41	26.80	18.84	19.61
w/o reasoning	37.11	25.74	18.63	19.13

Table 5: Ablation results on reasoning content of missing knowledge RAG framework.

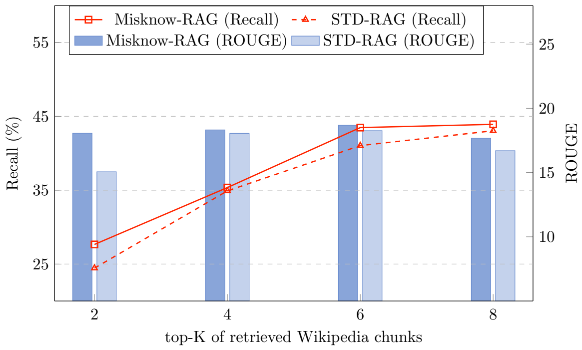
3.6 Comparative Analysis of Missing Knowledge and Standard top-K Retrieval

In this subsection, we offer a fair comparison between standard RAG and our proposed method to demonstrate the improved coverage of retrieved text chunks, which leads to more accurate final responses.

Specifically, we perform domain-specific experiments on Huatuo-26M with missing knowledge RAG and standard RAG, and then evaluate how comprehensively the retrieved passages encompass the pertinent knowledge to answer questions, as well as the accuracy of the generated responses, using context recall in Eq.4 and ROUGE-1 scores. As shown in Figure 3, we find that RAG with missing knowledge queries achieves better results in recall and ROUGE, especially for retrieving from QA pairs, which could be explained by Wikipedia passages offering extensive knowledge compared with QA pairs, which are short but more targeted to the queries. And as top-K becomes larger, the gap in performance between missing knowledge RAG and standard RAG becomes smaller. This could be rationalized by the sufficient context to answer the user's query.



(a) Retrieve from history QA pairs.



(b) Retrieve from Wikipedia knowledge base.

Figure 3: Comparison results of missing knowledge and standard RAG on context recall and ROUGE-1 scores across different top-K.

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 the model how to think utilizing internal parameter knowledge. Qian et al. (2024) proposes to use a parameteric memory module to generate context cues before retrieval, 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 in optimizing reasoning through an external process with a powerful LLM such as GPT-3.5 to generate follow-up thinking steps (Press et al., 2023; Yao et al.). Kim et al. (2024) proposes a RAG system with query decomposation and expansion. Jiang et al. (2024) and Wang et al. (2024) utilize Monte-Carlo Tree Search to find optimal chunk combinations, and Feng et al. (2025) aims at enhancing the model's self-consistency at test time.

These methods primarily focus on refining the reasoning space, addressing complex problems in the RAG system, which might ignore distribution gaps between queries and knowledge bases. Wang et al. (2025a) proposes to extract missing information and generate queries within separate modules, whereas Trivedi et al. (2023) performs iterative retrieval through multi-step reasoning. Inspired by their works, to solve the missing knowledge problem and explore distribution gaps in a special domain, we build a single-pass way to generate formatted missing knowledge queries from different knowledge sources. Our approach could not only avoid error propagation between modules but also leverage the interconnections among queries, contexts, and the missing knowledge. Furthermore, we utilize the reasoning content produced by the missing knowledge analysis phase, which encompasses verification signals about retrieved contexts to further enhance overall accuracy.

Query generation in RAG To capture user intent accurately with informal spoken expressions in queries, query rewriting seems to be a promising solution. Many efforts focus on training a query rewriting module to better align with user's intent (Ma et al., 2023; Wang et al., 2025b). This might improve the recall of retrieved contexts, further enhancing the overall accuracy of the reader's response. Li et al. (2024) proposes to train a unified model to simultaneously generate fine-grained clues and evidence.

Motivated by these approaches, we explore methods with different knowledge bases for generating appropriate missing knowledge queries while mitigating the gap between informal spoken expressions and professional documentation.

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 reasoning-based missing knowledge RAG framework, which introduces single-pass 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.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the ACL 2024*, pages 3563–3578.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Jingyi Song, and Hao Wang. 2025. Airrag: Activating intrinsic reasoning for retrieval augmented generation using tree-based search. *Preprint*, arXiv:2501.10053.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

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. 2025a. 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.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2025b. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI*, volume 39, pages 25434–25442.

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 Case Study of Reasoning Based Missing Knowledge Retrieval

In this section, we present case studies derived from the NQ and Huatuo datasets, encompassing both general and medical open-domain QA scenarios. For each original query and its corresponding missing knowledge query, we retrieve $\frac{top-K}{2}$ text passages from Wikipedia or QA pairs.

The examples provided in Table 6 and Table 7 demonstrate that QA pairs are more closely aligned with the query intent, particularly in colloquial contexts such as the query "`the most all star mvp awards`". Furthermore, within our proposed framework, the reasoning process involved in analyzing missing knowledge proves beneficial to the final response. This is because the accuracy of the retrieved knowledge is verified during the reasoning process, and any missing knowledge is subsequently retrieved through additional queries. Specifically, in Table 7, the final responses include more detailed treatment recommendations, such as "`examination, Astragalus and goji berries`".

B Prompt Used for Missing Knowledge and Response Generation

In this section, we describe the prompts used for generating missing knowledge queries and the final responses.

B.1 Prompt for Missing Knowledge and Query Generation

Figure 4 illustrates the prompt we employ to generate missing knowledge and the corresponding queries step by step. When provided with the user's query and the retrieved contexts as input, the model returns a JSON-formatted answer.

B.2 Prompt for Response Generation

Figure 5 presents the prompt used for answer generation. It integrates two key components: (1) the missing knowledge represented by the "thought" output from the missing knowledge generation phase shown in Figure 4, which provides a detailed analysis of the contexts, and (2) the reasoning content, which supplies verified information about specific contextual details. The prompt takes the user's query together with the above components as inputs and returns the final prediction.

C Prompt Used for Evaluating Context Recall

Figure 6 illustrates prompts utilized for generating ground truth contexts, which are then used to obtain the classification results and compute the recall defined in Eq. 4. A detailed example is listed in Figure 7, where "`attributed`" denotes whether an individual ground-truth context is presented among the retrieved contexts.

Query	Retrieval Context	Response
Question: Who has the most all star mvp awards? **Gold Answer:** Bob Pettit, Kobe Bryant	**Wiki**: Document[1]: Westbrook (2015 , 2016) are the only players to win consecutive awards... Jordan, Pettit and O'Neal did this multiple times... Document[2]: ... Kareem Abdul - Jabbar has the distinction of playing in the most All - Star Games (18) Document[3]: Pettit (1956) , Bob Cousy (1957) , Wilt Chamberlain (1960) , Bill Russell (1963)... all won the All - Star Game MVP and the NBA Most Valuable Player Award... Document[4]: NBA All - Star game records - wikipedia NBA All - Star game records Jump to : navigation... Bob Pettit was named MVP of the NBA All - Star Game four times...	**First Round Answer:** Bob Pettit holds the record with 4 All-Star MVP awards. **Reasoning Content:** Okay, let's tackle this step by step. The question is asking who has the most All-Star MVP awards. The retrieval contexts provided mention several players... The first-round answer states that Bob Pettit holds the record with 4 All-Star MVPs based on general knowledge. However, the provided contexts don't explicitly list the number of awards each player has ... Therefore, the answer relies on external knowledge not present in the context, indicating missing information. **Missing Knowledge:** Additional knowledge about the total count of All-Star MVP awards for each player, Clear identification of the player(s) with the most All-Star MVP awards **Final Response:** Bob Pettit
	QA Pairs: Q1: Who won the mvp at the all-star game ? A1: LeBron James... 1955 Bill Sharman... 1956 Bob Pettit... 1957 Bob Cousy... 1958 Bob Pettit... 1959 Elgin Baylor... 1960 Wilt Chamberlain... 1961 Oscar Robertson... 1962 Bob Pettit... Q2: Who has the most mvp awards in basketball? A2: ... Both Bill Russell and Michael Jordan won the award five times , while Wilt Chamberlain and LeBron James won the award four times... Q3: Who has won the most mvp awards in the nba? A3: ... Kareem Abdul - Jabbar won the award a record six times... while Bob Pettit , Karl Malone , Tim Duncan , Steve Nash and Stephen Curry have each won it twice... Q4: Who has won the most mvp awards in basketball? A4: ... Only two rookies have won the award : Wilt Chamberlain in the 1959 – 60 season and Wes Unseld in the 1968 – 69 season . Hakeem Olajuwon of Nigeria , Tim Duncan of the U.S. Virgin Islands , Steve Nash of Canada and Dirk Nowitzki of Germany are the only MVP winners considered " international players " by the NBA .	**First Round Answer:** Bob Pettit with 4 awards. **Reasoning Content:** Okay, let's tackle this question step by step. The user is asking who has the most All-Star MVP awards... Notable entries here include Bob Pettit with 4 awards (1956, 1958, 1959, 1962). The first-round answer states Bob Pettit with 4 awards. However, I need to verify if there's any mention of other players with more than 4 awards in the provided contexts. The second context talks about regular season MVP awards, not All-Star MVPs... The NBA's All-Star Game continued beyond 1964, so players like Kobe Bryant or LeBron James could have more , but the retrieval contexts here don't provide that information... Wait, the question is about All-Star Game MVPs, not regular season. The first context lists All-Star MVPs, and Pettit has 4... Hence, according to the provided context, the answer is correct, but there's missing info needed to answer the question fully... **Missing Knowledge:** Additional knowledge about The name of the person(s) with the most All-Star MVP awards, Confirmation of the current record for All-Star MVP awards **Final Response:** Bob Pettit, Kobe Bryant , LeBron James

Table 6: A case study on Natural Questions that presents both the reasoning process and the final response. The key information is annotated in red box .

Query	Retrieval Context	Response
Question: I am 30 years old, an office worker in a sub-health state: I always feel weak and lack energy. My family consulted an expert, who said that with qi and blood deficiency one can take compound Ejiao syrup. I would like to ask the doctor: can taking compound Ejiao syrup really improve my condition **Gold Answer:** For those chronically weakened by long-term illness, it is indeed possible to take compound Ejiao syrup, which can effectively tonify qi and blood and improve bodily weakness. It is recommended to combine it with a balanced diet, engage in practical outdoor exercise, and persist over time in order to strengthen the body.	**Wiki** Document[1]: ... full of qi, blood, and spirit... when these three energies are insufficient, the body will exhibit various "sub-health" symptoms... The simplest method for tonifying qi and blood is ginger and brown sugar water, or porridge made of black glutinous rice, longan, and red dates... Document[2]: How long must one tonify qi and blood before regaining health... avoid picky eating, and achieve nutritional balance ... Document[3]: Cassia seeds... are believed to pacify the liver and subdue yang, clear the liver, and brighten the eyes... Document[4]: Fuzi's efficacy and functions: Fuzi can enhance myocardial contractility, accelerate heart rate, increase cardiac output, and raise myocardial oxygen consumption...	**First Round Answer:** Compound Ejiao Syrup is a commonly used TCM formulation with certain regulatory effects for symptoms such as fatigue and pale complexion caused by deficiency of both qi and blood... It's best to first consult a professional doctor... Additionally, adjusting lifestyle habits is important, such as ensuring adequate rest and a balanced diet... **Reasoning Content:** ... the retrieved text lacks specific knowledge regarding Compound Ejiao Syrup, such as its efficacy and indications... the retrieved text contains general information about qi and blood deficiency... **Missing Knowledge:** The efficacy of Compound Ejiao Syrup, The usage of Compound Ejiao Syrup, The safety of Compound Ejiao Syrup **Final Response:** Compound Ejiao Syrup can effectively help improve symptoms caused by deficiency of both qi and blood , first visit a hospital... combined with a healthy lifestyle and a balanced diet ... don't forget to schedule regular follow-up visits so as to monitor changes in your condition and the efficacy of the medication.
	QA Pairs: Q1: A 28-year-old office worker always feeling very fatigued... A friend recommended I take Compound Ejiao Syrup to tonify qi and blood and improve subhealth— is this really targeted? A1: You can take it; the effect is quite good. You can also take some Astragalus granules together to tonify... Be sure to get plenty of rest and avoid overexertion, and also try to get some sun and do some exercise ... Q2: ... I've heard that taking Compound Ejiao Syrup can effectively eliminate these subhealth symptoms... Is this true? A2: Compound Ejiao Syrup can effectively eliminate subhealth symptoms caused by blood deficiency ... Q3: ... what symptoms deficiency of both qi and blood can lead to? A3: ... seek medical attention promptly to avoid delaying treatment. Q4: ... I feel so tired that I don't want to get out of bed, and have no strength to work. A4: ... If possible, take some blood-nourishing and spleen-fortifying formulas, such as Yangxue Guipi Pills or Guipi Pills...	**First Round Answer:** Compound Ejiao Syrup does have certain beneficial effects on physical weakness caused by deficiency of both qi and blood... adjusting your lifestyle is also very important, such as ensuring adequate sleep, a balanced diet, and moderate exercise... **Reasoning Content:** ... but they require combination with other medicines such as Astragalus granules, as well as exercise and rest... other conditions unsuitable for use ... **Missing Knowledge:** Contraindications of Compound Ejiao Syrup, Specific medical criteria for symptoms suitable for Compound Ejiao Syrup **Final Response:** You can try taking Compound Ejiao Syrup; it helps with qi and blood deficiency . But it's best to go to the hospital for a thorough examination ... under a doctor's guidance for greater safety. At the same time, pay attention to a balanced diet , more exercise , and ensure adequate sleep ... you can also combine it with dietary therapy such as Astragalus and goji berries ...

Table 7: A translated case study on Huatuo. We present both the reasoning process and the final response, where the key information is annotated in red box

.

<|system|>
You are a knowledge expert proficient in JSON formats.
Instructions
Given a question, retrieval contexts, and a predicted answer, your task is to determine if the provided contexts are sufficient to answer the question or if additional knowledge is required. If additional knowledge is needed, rewrite it as a short query closely related to the main entity in the question.
Requirements
Proceed step by step as follows:
- First, based on the question and the initial answer, determine if the answer is incomplete or if the retrieval contexts lack the knowledge required to answer the question.
- Second, if so, list the missing knowledge and generate the corresponding query.
- Third, output your reasoning and a JSON dict containing the fields "thought", "judge", "missing_knowledge", and "query", adhering strictly to the JSON format.
- Queries should be short, precise, and closely related to the main entity in the question.
Output Format
Provide a JSON dict in a markdown code block:
Key-value descriptions:
- thought: Analysis of the correctness and relevance of the retrieved context.
- judge: Whether the knowledge is missing.
- missing_knowledge: List of the missing knowledge points.
- query: List of queries corresponding to the missing knowledge.
Examples
{few-shot examples}

<|user|>
Question
{user query}
Retrieval Context
{retrieved contexts}
First-round answer

Figure 4: Prompts for missing knowledge query generation.

|<system>|
Instructions Please carefully read the following context and briefly answer the question with essential keywords or short phrasesbased 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.
Examples
{one-shot example}

|<user>|
Context
{retrieved contexts}
Missing Knowledge Context
Additional knowledge about {missing knowledge}
keywords from reasoning process: {reasoning content}
{retrieved missing knowledge contexts}
Question
{user query}
Answer

Figure 5: Prompts for reasoning enhanced answer generation.

Generate Ground Truth Contexts:
You are a professional knowledge assistant. Given a question and an answer, analyze the context or evidence knowledge needed to answer the question. Output json with context. Use the same language as the actual task.
The output should be a well-formatted JSON instance that conforms to the JSON schema below.
As an example, for the schema
`{“properties”: {“foo”: {“title”: “Foo”, “description”: “a list of strings”, “type”: “array”, “items”: {“type”: “string”}}}, “required”: [“foo”]}`
the object `{“foo”: [“bar”, “baz”]}` is a well-formatted instance of the schema.
The object `{“properties”: {“foo”: [“bar”, “baz”]}}` is not well-formatted.
Here is the output JSON schema:
{schema}
Examples:
few-shot examples
Your actual task:
Question:
{user query}
Answer:
{ground truth}
Contexts:

Generate Classification Results:
You are a professional knowledge assistant. Given a question, retrieved contexts and ground truth statements
, analyze each sentence in the statements can be attributed to a given context or not. Use only "Yes" (1) or "No" (0) as a binary classification. Output json with reason. Use the same language as the actual task.
The output should be a well-formatted JSON instance that conforms to the JSON schema below.
As an example, for the schema
`{“properties”: {“foo”: {“title”: “Foo”, “description”: “a list of strings”, “type”: “array”, “items”: {“type”: “string”}}}, “required”: [“foo”]}`
the object `{“foo”: [“bar”, “baz”]}` is a well-formatted instance of the schema.
The object `{“properties”: {“foo”: [“bar”, “baz”]}}` is not well-formatted.
Here is the output JSON schema:
{schema}
Examples:
few-shot examples
Your actual task:
Question:
{user query}
Contexts:
{retrieved contexts}
Ground Truth:
{ground truth}
Classification:

Figure 6: Prompts for generating ground-truth contexts and performing entailment classification for retrieval contexts.

Query:
Who plays hannibal in silence of the lambs?

Retrieved Contexts:
"who played hannibal lecter in silence of the lambs Anthony Hopkins, Anthony Hopkins as Dr. Hannibal Lecter", "who played hannibal lecter in silence of the lamb Anthony Hopkins, Anthony Hopkins as Dr. Hannibal Lecter", "who was the actor in silence of the lambs Frank Theodore "Ted" Levine, Frank Theodore Ted Levine (born May 29 , 1957) is an American actor . He is known for his roles as Buffalo Bill in The Silence of the Lambs and as Captain Leland Stottlemeyer in the television series Monk .", "who starred in the silence of the lambs Jodie Foster, Anthony Hopkins, Scott Glenn, Ted Levine, The Silence of the Lambs Theatrical release posterDirected by Jonathan Demme Produced by Kenneth Utt Edward Saxon Ron Bozman Screenplay by Ted Tally Based on The Silence of the Lambs by Thomas Harris Starring Jodie Foster Anthony Hopkins Scott Glenn Ted Levine Music by Howard Shore Cinematography Tak Fujimoto Edited by Craig McKay Production company Strong Heart / Demme Production Distributed by Orion Pictures Release date January 30, 1991 (1991 - 01 - 30) (New York City) February 14 , 1991 (1991 - 02 - 14) (United States) Running time 118 minutes Country United States Language English Budget 19 million Box office 272.7 million"

Ground Truth:
{"context": "Anthony Hopkins is a Welsh actor, director, and producer."}, {"context": "Anthony Hopkins portrayed the character Hannibal Lecter in the 1991 film The Silence of the Lambs."}, {"context": "The Silence of the Lambs is a 1991 American horror-thriller film directed by Jonathan Demme and starring Jodie Foster and Anthony Hopkins."}

Classification Results:

```
{
"statement": "Anthony Hopkins is a Welsh actor, director, and producer.",
"attributed": 0,
"reason": "The context does not mention Anthony Hopkins being a Welsh actor, director, and producer."
},
{
"statement": "Anthony Hopkins portrayed the character Hannibal Lecter in the 1991 film The Silence of the Lambs.",
"attributed": 1,
"reason": "The context explicitly states that Anthony Hopkins played Hannibal Lecter in The Silence of the Lambs."
},
{
"statement": "The Silence of the Lambs is a 1991 American horror-thriller film directed by Jonathan Demme and starring Jodie Foster and Anthony Hopkins.",
"attributed": 1,
"reason": "The context mentions that 'The Silence of the Lambs' stars Jodie Foster and Anthony Hopkins, and is directed by Jonathan Demme."
}
```

Figure 7: An illustrative example of context classification