REHKS-QA:Reflection-Enhanced Complex Question Answering for LLMs with Heterogeneous Knowledge Sources

Anonymous ACL submission

Abstract

Large Language Models (LLMs) still struggle with knowledge-intensive complex question answering, which requires reasoning over multiple knowledge facts. Existing approaches commonly use question decomposition with retrieval-augmented generation, where LLM first decomposes a complex question into sub-questions and then retrieves relevant information from external knowledge sources for sequential answering. Nevertheless, such methods suffer from error propagation, primarily due to negative retrieval, where irrelevant or missing knowledge misleads the LLMs' responses. To address these challenges, we propose **REHKS-QA** (Reflection-Enhanced Complex Question Answering for LLMs with Heterogeneous Knowledge Sources), a novel framework that integrates unstructured knowledge, structured knowledge and LLMs' parametric knowledge through a stepwise reflection mechanism. Specifically, REHKS-QA first decomposes complex questions into sub-questions, retrieves relevant external knowledge from both structured and unstructured sources, and generates preliminary answers. To mitigate misleading information, LLMs then explicitly reflect on the faithfulness of each answer by identifying supporting evidence. If no valid evidence is found, LLMs either revise their responses or use their parametric knowledge. Experimental results on two CQA benchmarks demonstrate that REHKS-QA not only outperforms state-of-the-art methods but also improves the explainability and verifiability of answers.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Dubey et al., 2024), supported by extensive training data and parameters (Minaee et al., 2024), have demonstrated impressive capabilities for various downstream tasks. Nevertheless, they still face challenges in knowledge-intensive complex question answering (CQA) (Cao et al., 2023; Tan et al., 2023), which requires LLMs to reason over multiple knowledge facts to obtain the final answer.

Currently, a widely adopted method integrates question decomposition with retrieval-augmented generation, requiring LLM to first break down complex questions into a series of atomic sub-questions and then leverage external knowledge sources to solve these sub-questions sequentially (Press et al., 2023; Trivedi et al., 2023). However, these decomposition-based methods suffer from the risk of error propagation (Cao et al., 2023). As illustrated in Figure 1, an incorrect answer to the first sub-question inevitably leads to errors in answering subsequent sub-questions that rely on previous output. Though a few of these errors are caused by inappropriate question decomposition (only 4%, see Sec. 6), we find the main cause of errors lies in **negative retrieval**: On the one hand, when only a single external knowledge source is used (e.g., Wikipedia), there is a possibility that no helpful knowledge can be found due to the underperforming retriever or the limited coverage of knowledge source; On the other hand, even though the evidence knowledge is recalled or the LLM can generate correct answer solely based on its internal parametric knowledge (e.g., "Ruslana Lyzhychko" in Figure 1), it would be easily misled by other retrieved distractive knowledge and finally produce an incorrect response.

To address these challenges, we are inspired by recent knowledge-fusion-related works (Zhang et al., 2023; Cao et al., 2023; Chu et al., 2024) and OpenAI's o-series models (Jaech et al., 2024), where the former shows the complementarity of different knowledge sources, and the latter demonstrates the benefit of continuous reflection in solving complex tasks. In light of this, We propose **REHKS-QA** (Reflection-Enhanced Complex Question Answering with Heterogeneous Knowledge Sources), a novel framework that incorporates

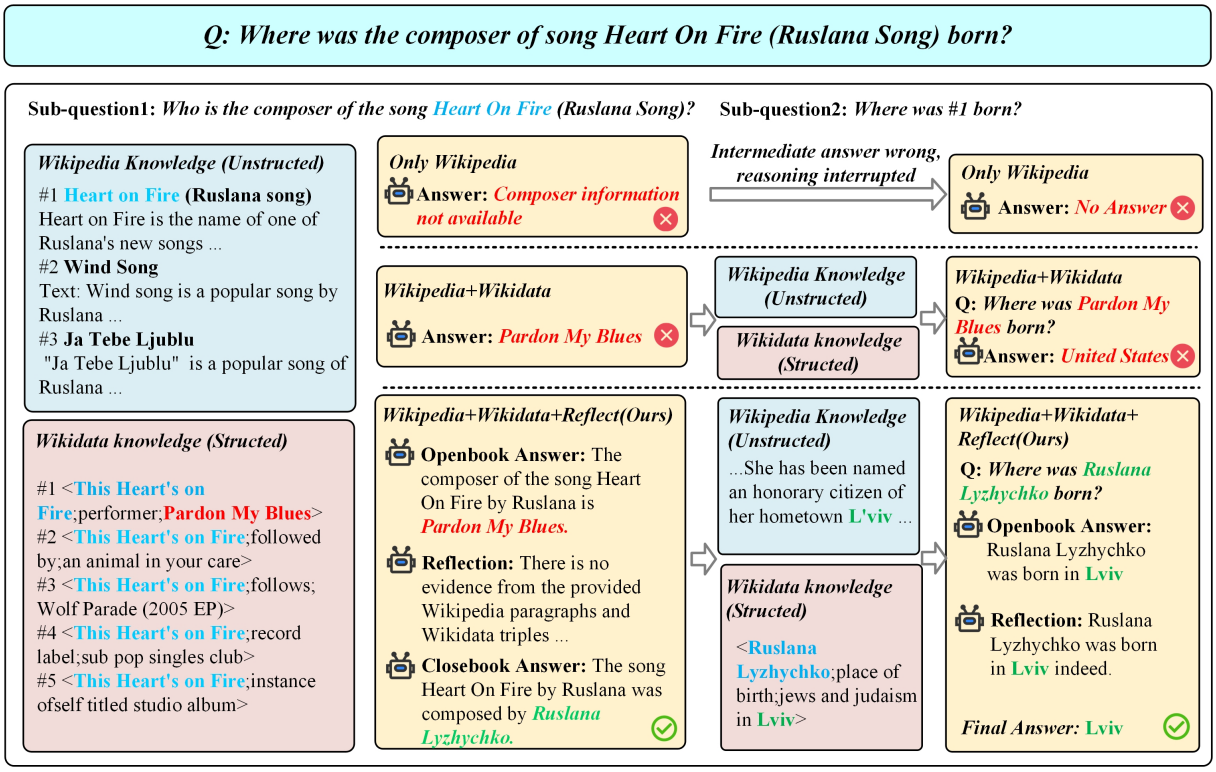

Figure 1: An example compares the use of different knowledge sources and response strategies.

three types of heterogeneous knowledge sources — unstructured knowledge (i.e., text corpus), structured knowledge (i.e., knowledge graphs), and LLMs' parametric knowledge — along with a stepwise reflection mechanism to mitigate the negative retrieval problem. Specifically, REHKS-QA begins by leveraging a backbone LLM to decompose the complex question into simpler sub-questions. For each sub-question, REHKS-QA first retrieves relevant external knowledge from both unstructured and structured knowledge sources, and employs the LLM to integrate them to obtain a preliminary answer. Then, to avoid misleading by distractive knowledge, the LLM is required to reflect on the faithfulness of the preliminary answer, i.e., explicitly pinpointing the evidence that supports the answer. Finally, the LLM needs to re-adjust the answer according to the located evidence, or directly leverage internal parametric knowledge to answer the sub-question if no evidence is found.

In summary, Our method cleverly uses a reflection mechanism to adaptively integrate heterogeneous knowledge and alleviate the influence of negative retrieval, thereby better guaranteeing the correctness of answers to each sub-question and mitigating error propagation. Besides, the pinpointed evidence also explicitly shows the sources of intermediate answers, improving the explainability and verifiability of REHKS-QA. Our evaluations on two CQA datasets show that our method not only enhances performance on benchmark datasets but also provides more interpretable and trustworthy responses

Our contributions include: (1) proposing REHKS-QA, a novel CQA framework that integrates heterogeneous knowledge fusion with a stepwise reflection mechanism, well alleviate the problems of negative retrieval and error propagation; (2) conducting thorough evaluations to demonstrate the superiority of REHKS-QA over SoTA CQA methods; (3) proving the effect of each component of our framework with careful ablation studies.

2 Related Work

2.1 Retrieval-Augmented Generation Question Answering

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Izacard and Grave, 2021) enhances LLMs by integrating external knowledge sources, showing great potential for knowledge-intensive and general language tasks (Ram et al., 2023). Knowledge sources include internal knowledge sources and external knowledge sources. The parametric knowledge within LLMs acts as an internal

source, providing significant potential for solving complex tasks (Yu et al., 2022).External knowledge sources can typically be classified into two primary categories according to their structural format: unstructured sources and structured sources. Unstructured sources leverage non-structured text, such as documents, web pages, or articles, to extract rich and contextually diverse information. However, the varied and inconsistent phrasing, as well as potential noise in unstructured text, can introduce challenges, often misleading models and resulting in inaccurate or incomplete answers, particularly when handling complex multi-step queries (Jo et al., 2021). This phenomenon exacerbates the unreliability of reasoning chains, where further leading to issues of hallucination (Huang et al., 2023). On the other hand, structured sources, like knowledge graphs (Hogan et al., 2021), offer highly organized and precise factual data, making them more reliable for extracting specific information (Chen et al., 2020).

Existing approaches mainly rely on unstructured text for retrieval, limiting effectiveness in complex reasoning tasks (Glass et al., 2022; Wang et al., 2023). Compared to existing methods, our framework offers significant advantages by leveraging three heterogeneous knowledge sources: parameter knowledge, unstructured knowledge and structured knowledge.

2.2 Reasoning Enhancement in LLMs

LLMs demonstrate impressive capabilities in reasoning tasks, but they also encounter hallucination problem (Xu et al., 2024). In order to solve the hallucination problem and enhance the reasoning ability of LLMs, many methods have proposed different reasoning ability enhancement methods. Chain of Thought (CoT) (Wei et al., 2022) guides the model to express its reasoning process to output more accurate answers. Self-consistency (Wang et al., 2022) enhances reasoning ability by selecting the most consistent answer from diverse reasoning paths. Error correction mechanisms are equally important for mitigating the errors during reasoning. Since LLMs possess inherent error-checking mechanisms, they are able to perform self-assessment and rectify potential errors throughout the generation process, thereby enhancing the accuracy and coherence of their outputs (Li et al., 2023). Ji et al. (2023) propose to generate and evaluate medical knowledge, prompting LLMs to self-correct and improve accuracy when discrepancies with established knowledge are identified. DUAL-REFLECT (Chen et al., 2024) improves the translation ability of LLM by comparing the differences between the back-translated results and the initial source input, revealing translation biases.

Inspired by these methods, our framework introduces a reflection-enhanced error correction mechanism, enabling LLM to improve its outputs by seeking evidence within the retrieved knowledge, thereby enhancing reasoning accuracy.

3 Task Definition

Given a complex question Q, our task is to employ an LLM $\mathcal{M}$ to answer the question by utilizing three types of knowledge: unstructured knowledge from a text corpus $\mathcal{K}_u = \{d_i\}$, structured knowledge from a knowledge graph $\mathcal{K}_s = \{(h_i, r_i, t_i)\}$, and parametric knowledge from $\mathcal{M}$ itself. Here, d_i denotes a textual paragraph, and (h_i, r_i, t_i) denotes a factual triple, where h_i, r_i, and t_i is the head entity, relation, and tail entity, respectively.

4 Methodology

We illustrate the overall framework of REHKS-QA in Figure 2. It consists of two stages: **question decomposition** and **reflection-enhanced multi-source reasoning**. In the question decomposition stage, the LLM decomposes the original multi-hop question into single-hop sub-questions. In the answer reasoning stage, each sub-question is addressed sequentially. For each sub-question, relevant information is first retrieved from both structured and unstructured knowledge sources, and the LLM generates a openbook answer. Then, this answer along with the retrieved knowledge is then fed back into LLM, where it reflects on the evidence and revises the openbook answer. If LLM fails to locate supporting evidence, it generates the final answer via closebook answer. Finally, in the answer selection stage, a strategy is implemented to choose the final answer.

4.1 Question decomposition

Recent studies have shown that LLM can decompose complex questions into simpler sub-questions, enabling step-by-step reasoning and improved question-solving performance (Ye et al., 2023). Following previous work (Chu et al., 2024), we leverage LLMs to decompose complex questions. Specifically, we decompose the original complex

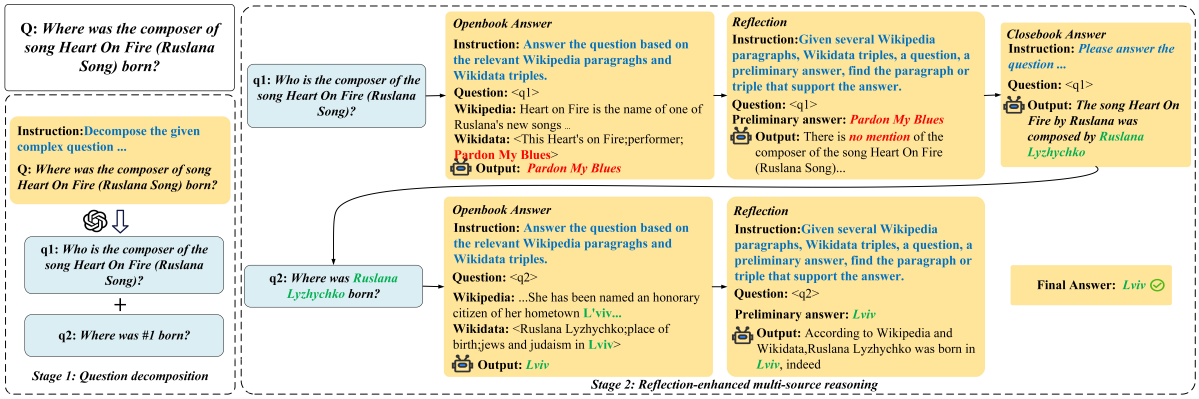

Figure 2: REHKS-QA consists of two steps: (1) Question decomposition (2) Reflection-enhanced multi-source reasoning.

question Q into a list of simpler sub-questions list q_{list}, thereby establishing a structured reasoning process to solve each sub-question sequentially and enhance overall reasoning performance.

$$q_{list} = \mathcal{M}(Q) \tag{1}$$

Considering the dependencies between these sub-questions, a placeholder $\#i$ is used to represent intermediate questions. During the processing of each sub-question, the placeholder is dynamically replaced with the answer from the previously resolved sub-question, facilitating a step-by-step reasoning progression. As illustrated in Figure 2 stage 1, the LLM decomposes the complex question into two sub-questions.

4.2 Heterogeneous Knowledge Sources

For external knowledge sources, we introduce unstructured and structured knowledge sources. By leveraging multiple types of knowledge sources, our method can acquire more comprehensive information, improving understanding and reasoning capabilities for complex questions and ensuring the accuracy and reliability of answers.

4.2.1 Unstructured Knowledge Source

Given a sub-question q, the unstructed knowledge base retriever $Retriever_u$ evaluates each paragraph's relevance by analyzing the relationship between question keywords and term frequency within the document. Then, it ranks paragraphs by relevance scores.

$$d = Retriever_u(q, K_u) \tag{2}$$

d represents the set of paragraphs in unstructed knowledge base K_u deemed most relevant to q. We use d as our unstructured background knowledge.

4.2.2 Structured Knowledge source

For structured knowledge sources, we use triples in structed knowledge base K_s. First, we first encode the entities e and relations r within K_s, denoted as:

$$E = \text{Encoder}(e) \tag{3}$$

$$R = \text{Encoder}(r) \tag{4}$$

Encoder represents the encoding module, where E and R denote the entities and relations after encoding into vector representations. To retrieve relevant triples related to q, we first use EntityExtractor identify the topic entity e_{topic} in q:

$$e_{topic} = \text{EntityExtractor}(q) \tag{5}$$

The entity linking process involves finding the most similar entity as head entity h in the knowledge graph based on cosine similarity:

$$h = \arg\max_{e^i \in E}(cosine\_similarity(e_{topic}, e^i)) \tag{6}$$

$cosine\_similarity$ is the function for calculating similarity. Next, we identify the top-k relations r_{top} that most relevant to the current sub-question q based on text similarity:

$$r_{top} = \underset{r^i \in R}{topk}(cosine\_similarity(q, r^i)) \tag{7}$$

We then use these relations to find the corresponding tail entities t_i, constructing triples related to the question. We use the retrieved triples as our structured background knowledge t.

$$t = \{(h, r_1, t_1)...(h, r_k, t_k)\} \tag{8}$$

4.3 Reflection-enhanced multi-source reasoning

We address each sub-question in q_{list} sequentially. Our framework involves four distinct stages for answering the sub-questions: 1) Openbook Answer 2) Reflect Answer 3) Closebook Answer and 4) Answer Selection. Each sub-question undergoes both the Openbook Answer and Reflect Answer phases. The Closebook Answer phase is invoked only when warranted, based on the outcomes of the Reflect Answer phase. The Answer Selection phase involves aggregating the answers from previous phases and dynamically selecting the most appropriate answer.

4.3.1 Openbook Answer

During the Openbook Answer phase, LLM is required to provide a preliminary answer to the q based on the retrieved background knowledge. Specifically, for each sub-quesiton q, we first retrieve the most relevant unstructured background knowledge d and the structured background knowledge t within the heterogeneous knowledge base. Subsequently, LLM generates a preliminary answer $a_{Openbook}$ based on background knowledge.

$$a_{Openbook} = \mathcal{M}(q, d, t) \tag{9}$$

4.3.2 Reflection

During the reflection answering phase, the LLM needs to identify relevant evidence in the retrieved background knowledge to support the answer to the q and revise the Openbook answer. Specifically, for each sub-question q, LLM must extract evidence from the unstructured Wikipedia paragraph background knowledge d and the structured Wikidata triple background knowledge t and the initial answer $a_{Openbook}$, to derive reflect answer $a_{Reflect}$.

$$a_{Reflect} = \mathcal{M}(q, d, t, a_{Openbook}) \tag{10}$$

The reflect answer can fall into three scenarios: 1) LLM considers the initial answer correct, thus retaining the original answer $a_{OpenBook}$. 2) LLM considers the initial answer is incorrect and finds supporting evidence in the background knowledge, leading to a revision of $a_{OpenBook}$ and obtaining a revised answer a_{Refect}. 3) LLM considers the initial answer is incorrect and does not find supporting evidence in the background knowledge, responding with "not mentioned" as $a_{Reflect}$.

4.3.3 Closebook Answer

When the LLM responds with "not mentioned" during the reflection phase, it indicates that the retrieved background knowledge does not contain any information to support answering the q. Specifically, when LLM responds with "not mentioned" in the reflection phase, it will perform a closebook question answering process using a chain-of-thought approach based on its internal knowledge, resulting in the Closebook answer $a_{Closebook}$.

$$a_{Closebook} = \mathcal{M}(q) \tag{11}$$

4.3.4 Answer Selection

For each sub-question, a definitive answer is selected from various response steps. In the reflection phase, if the LLM provides a Reflect Answer $a_{Reflect}$ that is not "not mentioned", this answer is adopted as the final answer. If the reflection answer is "not mentioned", LLM considers the Closebook answer $a_{Closebook}$. This selection process ensures that the most relevant and accurate answer is identified for each sub-question.

5 Experiments

5.1 Datasets

We evaluate the effectiveness of our framework on two complex multi-hop reasoning datasets: 2WikiMultiHopQA (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022). 2WikiMultiHopQA is a multi-hop QA dataset that combines structured and unstructured data. MuSiQue consists of complex multi-hop questions involving 2-4 hops of reasoning, which are then annotated to avoid reasoning shortcuts and ensure natural language. For question selection, we follow the same sampling strategy as IR-CoT (Trivedi et al., 2023), selecting 500 questions from the test set.

5.2 Baselines

We compare our approach with the following five baseline methods. The baseline methods are categorized into two types: non-retrieval-based methods and retrieval-based methods. Non-retrieval-based methods include Direct Answering (Direct) and Chain-of-Thought Answering (CoT). Retrieval-based methods include One-Retrieval (OneR), Self-Ask, and IR-CoT. The descriptions of each method are as follows:

- **Direct Answering (Direct)** (Brown, 2020) LLM directly generates the final answer.

Table 1: The overall results on MusiQue and 2WikiMQA, and the evaluation metric is F1.

Model	Method	Musique				2WikiMQA				
		Overall	2hop	3hop	4hop	Overall	Bridge	Inference	Comparison	Bridge-Comparison
GPT-3.5-turbo	Direct	19.2	23.0	15.5	14.8	38.1	14.7	32.3	62.1	59.0
	CoT	24.0	29.5	20.7	14.7	45.2	25.4	32.9	59.4	**75.2**
	OneR	17.0	18.5	11.5	22.1	38.0	11.5	34.2	84.3	38.9
	Self-ask	18.0	26.2	9.0	10.6	34.2	32.7	20.6	52.9	26.1
	IRCoT	28.1	34.2	18.8	**27.0**	52.2	32.4	40.7	**85.7**	60.0
	Ours	**33.8**	**43.2**	**24.2**	23.9	**65.5**	**59.9**	**52.4**	75.9	74.3
GPT-4o	Direct	22.3	27.3	17.4	16.5	41.7	16.6	32.1	68.7	65.3
	CoT	33.9	41.4	28.6	22.2	58.8	34.6	53.8	78.2	**85.9**
	OneR	26.8	35.8	18.6	15.9	50.0	15.8	38.7	**91.7**	75.3
	Self-ask	35.8	42.5	**33.1**	**21.8**	43.4	36.3	45.6	44.9	53.2
	IRCoT	23.0	34.8	12.2	8.6	51.3	44.2	48.7	80.2	33.8
	Ours	**36.9**	**46.4**	31.1	20.2	**67.1**	**55.7**	**76.1**	77.3	70.0
GPT-4o-mini	Direct	14.4	16.9	11.0	13.5	27.2	12.5	19.8	48.8	35.9
	CoT	22.6	17.2	28.6	14.8	36.6	15.6	30.7	58.2	63.4
	OneR	17.3	23.8	12.2	7.8	33.4	12.5	21.0	74.3	35.8
	Self-ask	33.2	41.5	27.9	**19.4**	31.5	19.8	27.3	47.7	38.2
	IRCoT	17.4	22.0	13.0	12.0	21.9	27.5	15.0	32.3	0.5
	Ours	**37.6**	**48.7**	**31.2**	17.7	**70.5**	**58.0**	**63.2**	**84.4**	**83.7**

- **Chain-of-Thought Answering (CoT)** (Wei et al., 2022) LLM generates reasoning steps before producing the final answer.
- **One-Retrieval (OneR)** The original question is used as a query, and the retrieved Wikipedia unstructured data and Wikidata structured data are concatenated into the prompt to guide LLM's CoT reasoning.
- **Self-Ask** (Press et al., 2023) This method employs an iterative approach to decompose complex questions. It iteratively generates sub-questions based on the existing reasoning, retrieves and answers sub-questions, and continues until the final answer is obtained.
- **IR-CoT** (Trivedi et al., 2023) This method interweaves retrieval-enhanced reasoning with reasoning-enhanced retrieval until enough information is retrieved to answer the question.

5.3 Implementation Details

For retrieval-based methods, to ensure a fair comparison, we retrieve both unstructured Wikipedia paragraphs and structured Wikidata triples. For unstructured knowledge from Wikipedia, we use the same retrieval corpus as IRCoT (Trivedi et al., 2023). We use BM25(Jones et al., 2000) as the retriever and retrieve the top 5 paragraphs. For structured knowledge from Wikidata, we use Wikidata5m (Wang et al., 2021) as the knowledge graph. We use all-MiniLM-L6-v2 (Reimers and Gurevych, 2020) to encode the entities and relations within the knowledge graph. When retrieving triples related to the question, we first identify the topic entity in the question using SpaCy (Choi et al., 2015), and identify the top 5 relations most relevant to the current question based on text similarity. We use three diffrent model as baseline models: GPT-3.5-turbo,GPT-4o and GPT-4o-mini. To ensure stable output, the temperature is set to 0. The evaluation metric we use is the token level F1 score.

Table 2: The ablation study on MusiQue, and the evaluation metric is F1.

Method	Overall	2hop	3hop	4hop
Wikipedia	30.6	39.0	**24.3**	18.0
+Wikidata	30.8	39.6	24.1	17.9
+Reflect (Ours)	**33.8**	**43.2**	24.2	**23.9**

5.4 Experiment Results

The overall results are presented in Table 1. Compared to other methods, our approach achieved a maximum improvement of 5.7% over the GPT-3.5-turbo on the Musique dataset, and a maximum improvement of 33.9% over the GPT-4o-mini on the 2WikiMQA dataset. We attribute the improvement in experimental results to three factors: (1) By introducing additional structured knowledge, we mitigated the drawbacks of pure unstructured knowledge, such as ambiguity and uncertainty, thereby enhancing LLM's reasoning capability. (2) Our approach incorporated a reflect step; answering step, which reduced the influence of irrelevant noisy documents retrieved by the retriever on

LLM. (3) Through Closebook answering phase, we addressed the issue of missing relevant background knowledge during retrieval, effectively facilitating the integration of internal and external knowledge and compensating for the lack of background knowledge that was not retrieved externally.

Furthermore, the results indicate that retrieval-based methods (e.g., OneR and Self-ask) do not consistently outperform non-retrieval approaches across both datasets. This limitation arises because retrieval-based methods often rely on retrieving the top k texts and triples related to the query, which may include irrelevant or misleading information. Such noise can misguide LLM, resulting in incorrect answers. In contrast, IR-CoT demonstrates superior performance over non-retrieval methods, suggesting its enhanced ability to identify and mitigate noise within retrieved documents. Our proposed method surpasses other methods, highlighting its effectiveness in avoiding the influence of erroneous and misleading information and effectively leveraging the internal knowledge within LLMs.

5.5 Ablation Study

To analyze the contributions of incorporating structured knowledge and the reflection mechanism, we conduct ablation experiments on the Musique dataset based on GPT-3.5-turbo. We design three different background knowledge configurations:

- **Wikipedia** only uses unstructured Wikipedia knowledge as background knowledge.
- **Wikipedia+Wikidata** uses both unstructured knowledge and structured knowledge as background knowledge.
- **Wikipedia+Wikidata+Reflect (Ours)** uses both unstructured knowledge and structured knowledge background knowledge, with the addition of reflection mechanism.

The results of the ablation experiments are presented in Table 2. Compared to the method that solely utilizes unstructured knowledge, incorporating structured knowledge leads to an overall improvement of 0.2%. This enhancement allows the LLM to better capture the relationships between entities, reduce ambiguity in the reasoning process, and improve the efficiency of information retrieval. Furthermore, by introducing the reflection-based answering step, the model reassesses and refines its initial responses, resulting in a reduction of reasoning errors and information omissions. This iterative process of review and reinforcement contributes to an overall improvement of 3%, thereby enhancing the accuracy of the final answers.

5.6 Performance Evaluation of Multi-Source Fusion Techniques

Table 3: Compare to other solve multi-source fusion method on MusiQue, and the evaluation metric is F1.

Method	Overall	2hop	3hop	4hop
Three source	27.9	37.5	21.3	12.5
Self-Consistency	32.2	**43.7**	21.2	18.8
REHKS-QA(Ours)	**33.8**	43.2	**24.2**	**23.9**

To explore the effectiveness of knowledge source integration, we conduct experiments on the Musique dataset using GPT-3.5-turbo. We compare our approach with two baseline methods: (1) **Three-Source**, based on ensemble learning (Burka et al., 2022), where the LLM independently answers using unstructured, structured, and parametric knowledge sources, and then selects the answer with the majority vote. (2) **Self-Consistency** (Wang et al., 2022), where the LLM generates three answers using unstructured, structured, and parametric knowledge sources at a temperature of 0.5, and then selects the answer with the majority vote. The experimental results shown in Table 3. Compared to other methods, REHKS-QA achieves the best overall performance with a score of 33.8%. It shows the effectiveness of REHKS-QA in integrating heterogeneous knowledge sources.

5.7 Case Study

We select two examples from the Musique dataset. As shown in Figure 3 (a), for the sub-question "When did Merseburg fall?", during the Openbook answering process, the LLM provides an incorrect answer "1815". Upon introducing the reflective answering mechanism, LLM is able to identify the correct answer "1738". As shown in Figure 3 (b), for the sub-question "Who was the president of Notre Dame in 2012?", the retriever failed to retrieve relevant background knowledge. Through the reflection mechanism, our method answers "not mentioned" and uses Closebook Answer to get the correct answer "John I. Jenkins, C.S.C." This shows that our method can not only correct errors, but also effectively integrate external knowledge and internal knowledge to obtain the correct answer.

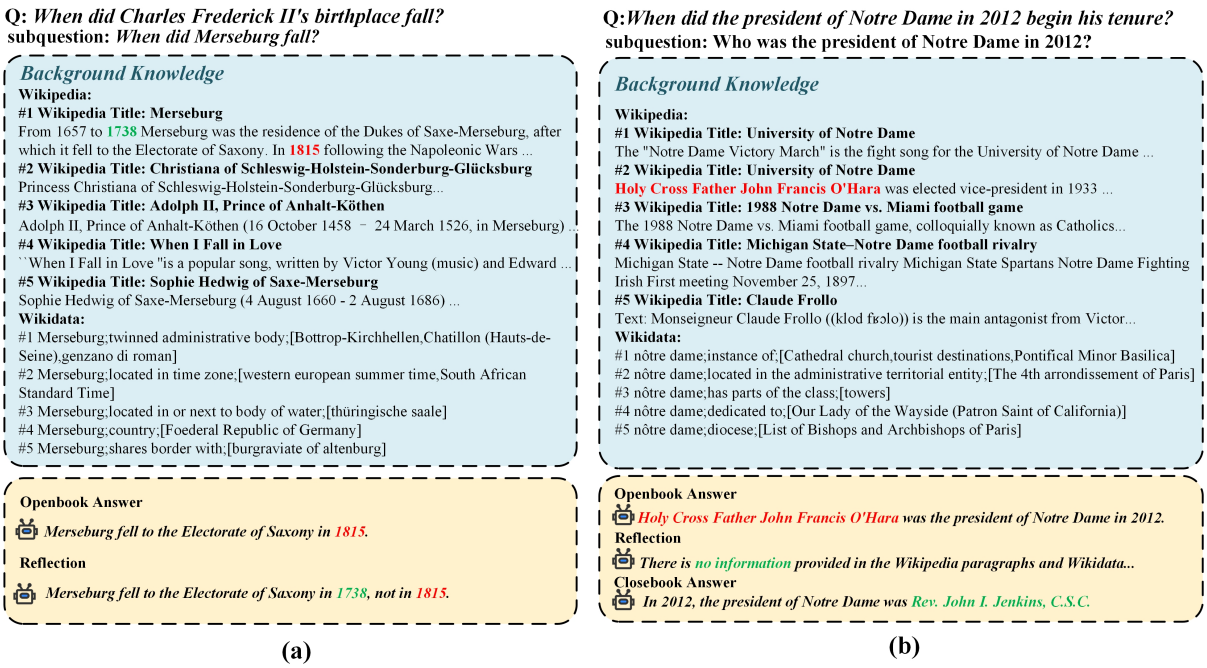

Figure 3: Case Study.

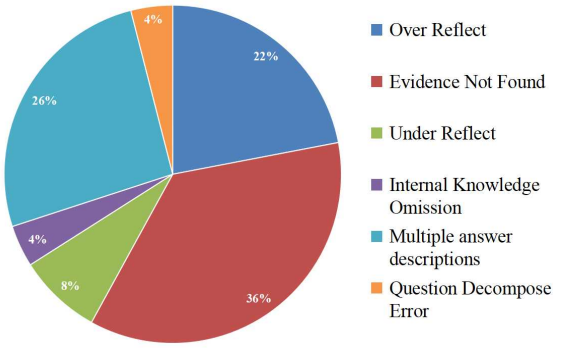

Figure 4: Error Analysis.

6 Error Analysis

We manually inspect 50 error samples from the Musique dataset. The results shown in Figure 4. The errors can be classified into six categories: 1) **Evidence Not Found** (36%), refers to cases where relevant information was retrieved, but no supporting evidence was found during the answering process. 2) **Over Reflect** (22%), where LLM over-reflects during intermediate steps, deviating from the correct answer, indicating the need for optimization of the reflection mechanism. 3) **Under Reflect** (8%), where no relevant information was retrieved, and the reflection process failed to correct the speculative answer generated. 4) **Internal Knowledge Omission** (4%), where no relevant information was retrieved, and the reflection identified the retrieval failure, but the LLM erroneously relied on its internal knowledge to generate an incorrect answer. 5) **Multiple Answer Descriptions** (26%), where the answer can have multiple valid expressions, making it difficult for F1 evaluation to detect correctness. 6) **Question Decompose Error** (4%), where LLM generates incorrect subquestions when decomposing complex questions, directly impacting answer accuracy. The analysis shows that external knowledge and reflection enhance performance, but challenges like hallucination and over-reflection persist.

7 Conclusion

We propose a novel framework REHKS-QA, introducing heterogeneous knowledge sources and a reflection mechanism to improve LLMs in complex question answering tasks. By combining structured data with the rich context of unstructured text, our framework addresses the limitations of existing retrieval-based methods that rely heavily on unstructured sources. The reflection mechanism further enhances answer quality by iteratively refining LLM's outputs, correcting errors, and ensuring higher factual accuracy. Experimental results on multi-hop question answering datasets show that our method significantly reduces hallucination and improves reliability.

8 Limitations

While our proposed framework shows significant improvements in question answering, it is also with limitations. One limitation is the reliance on effective question decomposition. The success of our approach depends heavily on accurately breaking down complex multi-hop questions into manageable sub-questions, which can sometimes be challenging, especially when the decomposition is ambiguous or the relationships between sub-questions are unclear. In future work, we aim to explore more advanced and flexible methods for question decomposition, which could further enhance the framework's performance.

9 Ethics Statement

All models and datasets utilized in this study are publicly available and distributed under permissible licenses. The training data has been fully desensitized.

References

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

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Dávid Burka, Clemens Puppe, László Szepesváry, and Attila Tasnádi. 2022. Voting: A machine learning approach. *European Journal of Operational Research*, 299(3):1003–1017.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Juanzi Li, and Lei Hou. 2023. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. *arXiv preprint arXiv:2311.13982*.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Dual-reflect: Enhancing large language models for reflective translation through dual learning feedback mechanisms. *arXiv preprint arXiv:2406.07232*.

Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert systems with applications*, 141:112948.

Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396.

Zheng Chu, Jingchang Chen, Qianglong Chen, Haotian Wang, Kun Zhu, Xiyuan Du, Weijiang Yu, Ming Liu, and Bing Qin. 2024. BeamAggR: Beam aggregation reasoning over multi-source knowledge for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1229–1248, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Hyeonseong Jo, Jinwoo Kim, Phillip Porras, Vinod Yegneswaran, and Seungwon Shin. 2021. Gapfinder: Finding inconsistency of security information from unstructured text. *IEEE Transactions on Information Forensics and Security*, 16:86–99.

K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840.

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

Yuheng Li, Lele Sha, Lixiang Yan, Jionghao Lin, Mladen Raković, Kirsten Galbraith, Kayley Lyons, Dragan Gašević, and Guanliang Chen. 2023. Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, 4:100140.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

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

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*.

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.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *arXiv preprint arXiv:2301.13808*.

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

Jiajie Zhang, Shulin Cao, Tingjia Zhang, Xin Lv, Jiaxin Shi, Qi Tian, Juanzi Li, and Lei Hou. 2023. Reasoning over hierarchical question decomposition tree for explainable question answering. *arXiv preprint arXiv:2305.15056*.

A Algorithm of REHKS-QA

Algorithm 1: REHKS-QA

Input: A question Q, unstructed knowledge base K_u, a unstructed knowledge base Retriever $Retriever_u$, structed knowledge base K_s, a structed knowledge base Retriever $Retriever_s$ and LLM parameter $\mathcal{M}$.

// Decompose Q to sub-question list q_{list}

q_{list}=$\mathcal{M}(Q)$

// Reflection-enhanced multi-source reasoning

for $q_i \rightarrow q_{list}$ **do**

  // Retrieve unstructed knowledge base K_u

  $d = Retriever_u(q_i, K_u)$

  // Retrieve structed knowledge base K_s

  $t = Retriever_s(q_i, K_s)$

  // Openbook Answer

  $a_{Openbook} = \mathcal{M}(q_i, d, t)$

  // Reflection

  $a_{Reflect} = \mathcal{M}(q_i, d, t, a_{Openbook})$

  // Answer Selection

  if $a_{Reflect}$ *is not mentioned* **then**

    // Closebook Answer

    $a_{Closebook} = \mathcal{M}(q_i)$

  end

  if $a_{Reflect}$ *is not mentioned* **then**

    $A_i = a_{Closebook}$

  end

  else

    $A_i = a_{Reflect}$

  end

end

$A = A_n$

return A

B Prompts

We list the prompts used in this work in Figure 5, 6, 7, 8.

Please help me break down a complex multi-hop question into multiple single-hop questions.If the construction of a question depends on the answer to the previous question, you need to use a placeholder "#question_idx" to represent the answer to the question.Follow the answer format below.

Q: When did the first large winter carnival take place in the city where CIMI-FM is licensed to broadcast?
A: ["Which city is CIMI-FM licensed to broadcast?", "When did the first large winter carnival take place in #1?"]
Q: What county is Hebron located in, in the same province the Heritage Places Protection Act applies to?
A: ["Which did Heritage Places Protection Act apply to the jurisdiction of?", "Which country is Hebron, #1 located in?"]
Q: What did the publisher of Banjo-Tooie rely primarily on for its support?
A: ["What is the publisher of Banjo-Tooie?", "What did #1 rely primarily for its support on first-party games?"]
Q: In which county was the birthplace of the Smoke in tha City performer?
A: ["Who is the performer of Smoke in tha City?", "What's the birthplace of #1?","Which country is #2 located in?"]
Q: What region of the state where Guy Shepherdson was born, contains SMA Negeri 68?
A: ["Where was Guy Shepherdson born?", "What region of the state is SMA Negeri 68 #1 located in?"].
Q: When did Britain withdraw from the country containing Hoora? A: ["Which country is Hoora in?", "When did Britain withdraw from #1?"]
Q: How long is the US border with the country that borders the state where Finding Dory takes place?
A: ["Where is finding dory supposed to take place?", "Which country shares a border with #1?", "How long is the us border with #2?"]
Q: When did the first large winter carnival happen in Olivier Robitaille's place of birth?
A: ["Where was Olivier Robitaille born?", "When did the first large winter carnival take place in #1?"]
Q: When did Britain withdraw from the country where the village of Wadyan is found?
A: ["Which country is Wadyan in?", "When did Britain withdraw from #1?"]
Q: How many countries in Pacific National Universityś continent are recognized by the organization that mediated the truce ending the Iran-Iraq war?
A: ["Which country is Pacific National University located in?", "What continent is #1 in?", "Who mediated the truce which ended the Iran-Iraq War?", "The #3 recognises how many regions in #2?"]
Q: When was Eritrea annexed by the Horn of Africa country where, along with Somalia and the country where Bissidiro is located, Somali people live?
A: ["Which country is Bissidiro located in?", "Along with Kenya, #1 and Somalia, in what Horn of Africa country do Somali people live?"]
Q: Where is the lowest place in the country which, along with Eisenhower's VP's country, recognized Gaddafi's government early on?
A: ["Who served as Eisenhower's vice president?", "#1 was a president of what country?", "Where is the lowest place in the #2"]
Q: When did the capital of Virginia moved from John Nicholas's birth city to Charles Oakley's alma mater's city?
A: ["Which university was Charles Oakley educated at?", "Which city was #1 located in?", "Where was John Nicholas born?", "When did the capital of virginia moved from #3 to #2?"]
...

Figure 5: prompt for Question Decompose.

Given a question,the relevant Wikipedia text and the relevant Wikidata triples, give the helpful paragrphs in Wikipedia and helpful triples in Wikidata to answer the question,and tell the rationalization.
Wikipedia:
#1 Wikipedia Title: So Long, See You Tomorrow (album)
Text: So Long, See You Tomorrow is the fourth album by the London indie rock band Bombay Bicycle Club, released on 3 February 2014. The album is named after the novel of the same name by William Maxwell.
#2 Wikipedia Title: Hallelujah I Love Her So
Text: “Hallelujah I Love Her So ”Single by Ray Charles from the album Ray Charles (or, Hallelujah I Love Her So) B - side“ What Would I Do Without You” Released 1956 Format 7 ”45rpm Recorded 1956 Genre soul rhythm and blues Length 2: 35 Label Atlantic Songwriter (s) Ray Charles Producer (s) Jerry Wexler Ray Charles singles chronology “A Fool for You” (1955)“ Hallelujah I Love Her So ”(1956) “Mary Ann” (1956)“ A Fool for You ”(1955) “Hallelujah I Love Her So” (1956)“ Mary Ann ”(1956)
#3 Wikipedia Title: The First Time Ever I Saw Your Face
Text: “The First Time Ever I Saw Your Face ”Single by Roberta Flack from the album First Take Released March 7, 1972 (1972 - 03 - 07) Recorded 1969 Genre Soul vocal jazz Length 5: 22 4: 15 (1972 radio edit) Label Atlantic 2864 Songwriter (s) Ewan MacColl Producer (s) Joel Dorn Roberta Flack singles chronology“ Will You Still Love Me Tomorrow” (1972) “The First Time Ever I Saw Your Face ”(1972)“ Where Is the Love” (1972) “Will You Still Love Me Tomorrow ”(1972)“ The First Time Ever I Saw Your Face” (1972) “Where Is the Love ”(1972)
#4 Wikipedia Title: See You on the Other Side (Mercury Rev album)
Text: See You on the Other Side is the third studio album by American neo-psychedelia band Mercury Rev, released in 1995 by record label Beggars Banquet.
#5 Wikipedia Title: The Dance (song)
Text: “The Dance ”Single by Garth Brooks from the album Garth Brooks B - side“ If Tomorrow Never Comes” Released April 30, 1990 Format CD single, 7 ”45 RPM Recorded 1988 – 1989 Genre Country Length 3: 40 Label Capitol Nashville 44629 Songwriter (s) Tony Arata Producer (s) Allen Reynolds Garth Brooks singles chronology “Not Counting You” (1990)“ The Dance ”(1990) “Friends in Low Places” (1990)“ Not Counting You ”(1990) “The Dance” (1990)“ Friends in Low Places ”(1990)
Wikidata:
#1 So Long, See You Tomorrow;performer;[Bombay Bicycle Club]
#2 So Long, See You Tomorrow;instance of;[audio album]
#3 So Long, See You Tomorrow;follows;[A Different Kind of Fix]
Q: Who is the performer of So Long, See You Tomorrow?
A:{
"Wikipedia":{
"relevant_paragrph":["#1"],
"rationalization":"The record label of Bombay Bicycle Club is Island Records",
"answer":"Bombay Bicycle Club"
},
"Wikidata":{
"relevant_triples":["#1"],
"rationalization":"The performer of So Long, See You Tomorrow is Bombay Bicycle Club",
"answer":"Bombay Bicycle Club"
},
"answer":"Bombay Bicycle Club"
}
...

Figure 6: prompt for Openbook Answer.

Given several Wikipedia paragraphs, Wikidata triples, a question, and a preliminary answer, find evidence from the paragraphs and triples to support the preliminary answer. If the preliminary answer is incorrect, correct it. If no evidence mentioned in Wikipedia and Wikidata to answer the question, answer "not mentioned".
Wikipedia:
#1 Wikipedia Title: So Long, See You Tomorrow (album)
Text: So Long, See You Tomorrow is the fourth album by the London indie rock band Bombay Bicycle Club, released on 3 February 2014. The album is named after the novel of the same name by William Maxwell.
#2 Wikipedia Title: Hallelujah I Love Her So
Text: “Hallelujah I Love Her So ”Single by Ray Charles from the album Ray Charles (or, Hallelujah I Love Her So) B - side“ What Would I Do Without You” Released 1956 Format 7 ”45rpm Recorded 1956 Genre soul rhythm and blues Length 2: 35 Label Atlantic Songwriter (s) Ray Charles Producer (s) Jerry Wexler Ray Charles singles chronology “A Fool for You” (1955)“ Hallelujah I Love Her So ”(1956) “Mary Ann” (1956)“ A Fool for You ”(1955) “Hallelujah I Love Her So” (1956)“ Mary Ann ”(1956)
#3 Wikipedia Title: The First Time Ever I Saw Your Face
Text: “The First Time Ever I Saw Your Face ”Single by Roberta Flack from the album First Take Released March 7, 1972 (1972 - 03 - 07) Recorded 1969 Genre Soul vocal jazz Length 5: 22 4: 15 (1972 radio edit) Label Atlantic 2864 Songwriter (s) Ewan MacColl Producer (s) Joel Dorn Roberta Flack singles chronology“ Will You Still Love Me Tomorrow” (1972) “The First Time Ever I Saw Your Face ”(1972)“ Where Is the Love” (1972) “Will You Still Love Me Tomorrow ”(1972)“ The First Time Ever I Saw Your Face” (1972) “Where Is the Love ”(1972)
#4 Wikipedia Title: See You on the Other Side (Mercury Rev album)
Text: See You on the Other Side is the third studio album by American neo-psychedelia band Mercury Rev, released in 1995 by record label Beggars Banquet.
#5 Wikipedia Title: The Dance (song)
Text: “The Dance ”Single by Garth Brooks from the album Garth Brooks B - side“ If Tomorrow Never Comes” Released April 30, 1990 Format CD single, 7 ”45 RPM Recorded 1988 – 1989 Genre Country Length 3: 40 Label Capitol Nashville 44629 Songwriter (s) Tony Arata Producer (s) Allen Reynolds Garth Brooks singles chronology “Not Counting You” (1990)“ The Dance ”(1990) “Friends in Low Places” (1990)“ Not Counting You ”(1990) “The Dance” (1990)“ Friends in Low Places ”(1990)
Wikidata:
#1 So Long, See You Tomorrow;performer;[Bombay Bicycle Club]
#2 So Long, See You Tomorrow;instance of;[audio album]
#3 So Long, See You Tomorrow;follows;[A Different Kind of Fix]
Q: Who is the performer of So Long, See You Tomorrow?
Preliminary answer:"Bombay Bicycle Club"
A:{
"analysis":"According to Wikipedia #1, the performer of So Long, See You Tomorrow is Bombay Bicycle Club indeed.According to Wikidata #1,the performer of So Long, See You Tomorrow is Bombay Bicycle Club indeed.",
"answer":"Bombay Bicycle Club"
}
...

Figure 7: prompt for Reflection.

Please answer the question by thinking step-by-step.
Q: When did the first large winter carnival take place in the city where CIMI-FM is licensed to broadcast?
A:{
"analysis":"CIMI-FM is licensed to broadcast in Quebec City. The first large winter carnival in Quebec City took place in 1894",
"answer":"1894"
}
Q: When was Neville A. Stanton's employer founded?
A:{
"analysis":"The employer of Neville A. Stanton is University of Southampton. The University of Southampton was founded in 1862",
"answer":"1862"
}
Q: What religion did the black community found?
A:{
"analysis":"The black community found African Methodist Episcopal Church.",
"answer":"African Methodist Episcopal Church"
}
Q: What county is Hebron located in, in the same province the Heritage Places Protection Act applies to?
A:{
"analysis":"Heritage Places Protection Act applies to the jurisdiction of Prince Edward Island. Hebron, Prince Edward Island is located in the Prince County.",
"answer":"Prince County"
}
Q: What weekly publication in the Connecticut city with the most Zagat rated restaurants is issued by university of America-Lite: How Imperial Academia Dismantled Our Culture's author?
A:{
"analysis":"The author of America-Lite: How Imperial Academia Dismantled Our Culture is David Gelernter. David Gelernter was educated at the Yale University. The city in Connecticut that has the highest number of Zagat-rated restaurants is New Haven. The weekly publication in New Haven that is issued by Yale University is Yale Herald.",
"answer":"Yale Herald"
}
Q: What is the headquarters for the organization who sets the standards for ISO 21500?
A: {
"analysis":"The standards for ISO 21500 were set by International Organization for Standardization. The International Organization for Standardization has headquarters in Geneva.",
"answer":"Geneva"
}

Figure 8: prompt for Closebook Answer.