The Collision of Blades Reveals Their Sharpness: Retrieval-Augmented Generation Consolidate by Knowledge Conflict Exposure

Anonymous ACL submission

Abstract

To alleviate hallucination and outdated knowledge in LLMs, current LLM systems frequently integrate retrieval-augmented generation (RAG) techniques to form a RAG-LLM system. However, misinformation and disinformation are prevalent in external corpus, seriously threaten the system's reliability, which makes the consolidation necessary. Even though many approaches based on the credibility of external content have achieved impressive performance, it remains challenging how the additional assessment could be perceived and ultimately utilized by LLMs. Inspired by cognitive conflict theory, we propose an approach to consolidate the RAG-LLM systems through knowledge conflict exposure. To reveal potential knowledge inconsistency, our approach designs a novel information expansion strategy, introducing comparative content from both high-level intent and fine-grained supporting materials. Through knowledge extraction and conflict prompting, it achieves more effective consolidation. Experimental evaluations demonstrate that our proposed approach can achieve average performance increase of 10% compared to baseline approaches, underscoring its efficacy in improving LLM output.

1 Introduction

Currently, large language models (LLMs) demonstrate advanced intelligence, and has shown considerable performance in many tasks (Touvron et al., 2023; Brown et al., 2020). In order to further optimize LLM and enhance its ability to cope with information missing and hallucinations (Ji et al., 2023; Ye et al., 2024), as shown in Figure 1, a generation paradigm has been proposed combining the Retrieval-Augmented Generation (RAG) framework and LLM (Borgeaud et al., 2022; Chen et al., 2024; Izacard et al., 2023), which integrates external knowledge bases to form the RAG-LLM system to improve the reliability of the final output.

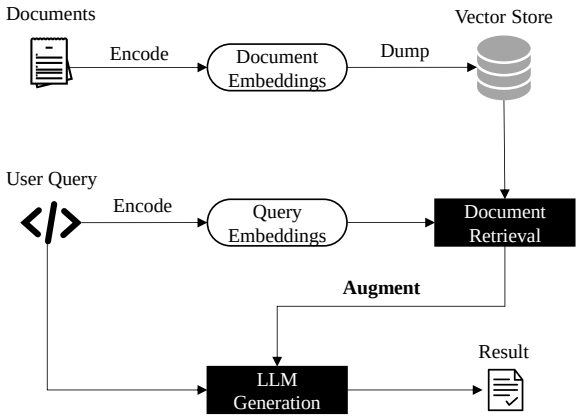

Figure 1: RAG workflow

However, external knowledge corpus RAG systems rely on are usually polluted by defective documents, such as misinformation and disinformation. Such defective documents could be retrieved and introduced, potentially interfering with or deliberately inducing LLMs to generate content that deviates from expectations (Liu et al., 2023; Abdelnabi et al., 2023; Yi et al., 2023). To this end, many strategies have been proposed to consolidate RAG systems against interference caused by defective documents (Xiang et al., 2024; Deng et al., 2024; Wei et al.; Wang et al., 2024), such as employing cross-verification to refine the retrieved knowledge and even discard the knowledge with low confidence, etc.

Although previous methods have achieved impressive performance, there are still some shortcomings, manifested mainly in two aspects. First, they may produce false negatives in the reliability assessment when dealing with defective documents with subtle differences with the correct knowledge. This makes it difficult to identify potential discrepancies in the context simply by comparing the textual similarity with the existing knowledge. Second, although knowledge credibility is correctly assessed, it remains challenging whether LLMs

can effectively perceive and leverage embedded credibility and ultimately improve their reasoning ability through naive prompting.

In this paper, we propose a novel consolidated approach to mitigate the impact of pollution in externally retrieved knowledge on the performance of RAG-LLM systems. Inspired by the **Cognitive Conflict Theory** by Piaget (2005), conflicts can prompt individuals to reassess their cognitive structures, thereby adjusting their ways of thinking or behavior. Our approach aims to compel LLMs to redirect their attention towards potential inaccuracies within the context by creating conflicts with open-world knowledge where correct information predominates, thereby enhancing the performance of the LLM system. To achieve this, we first obtain diverse query statements from both high-level (through question rewriting) and fine-grained through question splitting, which are then used for document retrieval to expand the scope of information, thereby enhancing the completeness of knowledge and provide a solid foundation for conflict exposure. Since a single document may contain multiple pieces of knowledge, conflict detection at the document level may not be sufficiently accurate (as sometimes only a single piece of knowledge may be in conflict, rather than the entire document). Therefore, before executing the final conflict exposure, we also extract knowledge from the documents to form knowledge triples. Subsequently, we perform conflict detection based on these knowledge triples and annotate the documents accordingly based on the detection results. This provides LLMs with conflict exposure information to enhance their performance.

We evaluated our approach on HotPotQA (Yang et al., 2018), which is a commonly used QA dataset. We conducted experiments on meta AI's Llama3 model (AI@Meta, 2024) and Google's Gemma2 model (Team, 2024), and the results showed that our approach can achieve an average performance increase of 10.04% compared to the baselines when encountering knowledge pollution. Moreover, we conducted experiments on the strategies of conflict exposure and information expansion, and the results also verified their effectiveness.

The contributions of this article are as follows.

- Propose a RAG-LLM consolidation approach based on knowledge conflict exposure, providing a new way to optimize RAG-LLM output when encountering knowledge pollution.
- The evaluation was conducted on a commonly-used dataset, and the results proved that the proposed approach has promising performance.
- Publicly available resources to advance research in this field[1].

2 Related Work

In order to alleviate the impact of defective documents on RAG-LLM, some consolidate strategies have also been proposed.

Xiang et al. (2024) proposed a new RAG framework called `Robust RAG`, which mainly adopts a generation strategy of "isolation first, aggregation later". In order to avoid defective documents misleading the generated results, when facing retrieval knowledge pollution, the framework will separately process each retrieved document, that is, let LLM generate independent answers for each document, and then aggregate these independent answers by using keyword and decoding aggregation.

In contrast, Deng et al. (2024) did not make any changes to the generation framework, they proposed a credibility aware attention modification plugin `CrAM`, which first determines which attention heads in LLM are most sensitive to document information, and then adjusts the weights of these identified attention heads based on the document's credibility to reduce the impact of low credibility documents on the final output.

Wei et al. approached the issue from the perspective of LLM optimization. They argue that in the process where LLM directly predict final answers from inputs, the input data may contain noise, and the implicit denoising process carried out by the LLM is difficult to interpret and verify. Therefore, the authors proposed the `InstructRAG` approach, which guides the LLM to explain how to derive the true answer from the retrieved document, thereby generating specific evidence for the answer. Afterwards, it is used as an explicit denoising example in context learning of LLM or as supervised data to fine tune the model, thus enhance the inference ability of the LLM and improve the verifiability of the output.

Afterwards, Wang et al. (2024) proposed the `AstuteRAG` approach, which does not modify the main components of the RAG-LLM process. Instead, it relies on the inherent discriminative capabilities of LLM to integrate internal and external

[1]https://anonymous.4open.science/r/ConsolRAG-A242

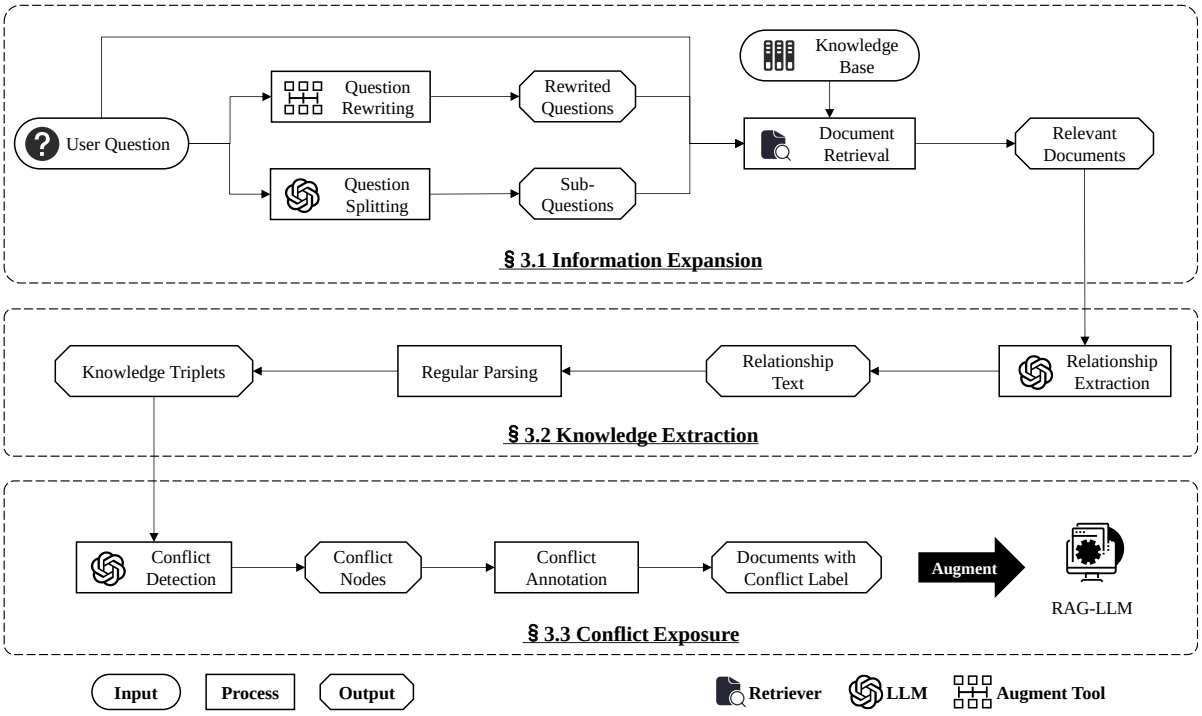

Figure 2: Approach overview

knowledge. This approach extracts key information from the internal knowledge of LLM through adaptive templates, and then uses integration templates to direct the LLM to consciously combine internal and external knowledge sources. Ultimately, this approach will generate reliable knowledge for the subsequent generation of RAG-LLM.

For the above approaches, `RobustRAG` and `AstuteRAG` are sample free approaches, while `CrAM` and `InstructRAG` require samples to generate credibility scores or for in-context learning.

3 Approach

In this paper, we propose a consolidated approach for RAG-LLM systems. As shown in Figure 2, the approach exposes knowledge conflicts (two pieces of knowledge are mutually exclusive, meaning they cannot be true at the same time.) to make LLM aware of possible knowledge errors. The specific steps are as follows:

- **Information Expansion**: Expand the number of documents by rewriting and splitting questions to form different queries and obtain various retrieval contents.
- **Knowledge Extraction**: Extracting knowledge from documents for constructing conflict prompts used in subsequent conflict detection:
- **Conflict Exposure**: Identify conflicts based on LLM or conflict detection models, and add annotation to the corresponding documents based on the knowledge conflicts.

Based on the above three steps, the proposed approach can generate a consolidated LLM input for user questions under the scenario of knowledge pollution. The specific description is as follows.

3.1 Information Expansion

In the scenario of knowledge pollution, the corpus retrieved by retriever may contain defective documents, which not only affect the thinking of the LLM but also occupy the position of some normal documents, making it difficult to successfully retrieve the necessary documents.

Therefore, in order to fully expose potential conflicts in knowledge, it is necessary to first expand the information retrieved. Only when the external documents are comprehensive can conflicts be better exposed. The rationality of this operation primarily relies on two empirical premises. The first is that **the foundation for identifying erroneous knowledge is based on the existence of conflicting evidence in the real world**, that is, to prove an assertion is incorrect, one must find corresponding evidence from other sources. The second is

that **erroneous knowledge is hard to form a perfect closed loop within a knowledge base**, which means that incorrect knowledge will often conflict with other correct knowledge because the maintenance of the knowledge base is predominantly carried out by regular users rather than unwanted ones.

Algorithm 1 Information Expansion

Input: user question q, knowledge base K, data augment model M, retriever R, LLM L, question splitting prompt p, top num k, expand num n
Output: relevant documents D

$D \leftarrow Retrieval(K, q, k)$
$n \leftarrow n/2$
$m \leftarrow k + n$
$Q_{rewrite} \leftarrow Augment(M, q)$
$D_{rewrite} \leftarrow EXPAND(K, Q_{rewrite}, m)$
$D_{rewrite} \leftarrow D_{rewrite} - D$
$D_{rewrite} \leftarrow GetTop(n, D_{rewrite})$
$t \leftarrow p \oplus q$
$Q_{split} \leftarrow Generate(L, t)$
$D_{split} \leftarrow EXPAND(K, Q_{split}, m)$
$D_{split} \leftarrow D_{split} - D$
$D_{split} \leftarrow GetTop(n, D_{split})$
$D \leftarrow D \cup D_{rewrite} \cup D_{split}$
return D
function EXPAND(K, Q, m)
    $D_{expand} \leftarrow \emptyset$
    for each q in Q **do**
        $D_{temp} \leftarrow Retrieval(K, q, m)$
        $D_{expand} \leftarrow D_{expand} \cup D_{temp}$
    end for
    return D_{expand}
end function

As shown in Algorithm 1, in the proposed approach, we derive new queries from two aspects of the original user's questions, namely question rewriting and question splitting. Question rewriting is mainly based on the idea of data augmentation, by changing the words in the original question to form a new query. The specific operation is to compare the word embeddings of the vocabulary and replace the original word in the question with the most similar one. The advantage of this operation is that it can reduce the impact of knowledge pollution on specific user question while retrieving relevant documents in high-level intent. Question splitting is the process of using a LLM to break down the original user question into several sub-questions, asking the LLM which topics are helpful in answering user questions (the prompt is in Appendix A), and using these topics to search for relevant documents, thereby providing fine-grained supporting materials for other documents. Moreover, it can also expand the scope of information, better expose knowledge conflicts, and help LLM generate more reliable results.

3.2 Knowledge Extraction

A single document typically contains multiple knowledge elements, indicating that conflict detection at the document level is insufficient. In order to obtain more fine-grained conflict detection results, it is necessary to extract knowledge from documents, which can provide a more comprehensive data foundation for conflict exposure.

Algorithm 2 Knowledge Extraction

Input: relevant documents D, LLM L, relationship extraction prompt p
Output: knowledge triplets T

$T \leftarrow \emptyset$
for each d in D **do**
    $t \leftarrow d \oplus p$
    $O \leftarrow Generate(L, t)$
    for each o in O **do**
        $o' \leftarrow Parsing(o)$
        if $len(o') == 3$ **then**
            $T \leftarrow T \cup \{o'\}$
        end if
    end for
end for
return T

In our approach, the knowledge is defined as an entity relationship triplet like $< Entity_1, Relationship, Entity_2 >$. As shown in Algorithm 2, we can use the LLMs to directly extract entity relationship triplets from knowledge documents (the prompt is in Appendix B).

Another feasible way is to first extract key entities from the knowledge documents to obtain an entities list, and then extracts association relationships based on the document and key entities. For this operation, it can use named entity recognition models and relationship extraction models based on traditional natural language processing.

In our approach, in order to ensure the extract triplets are valid, we leverage a regular parsing process to find all items that contains three elements in the relationship text output by LLM as the final knowledge triplets.

Finally, these knowledge triplets can serve as inputs for subsequent conflict exposure operations.

3.3 Conflict Exposure

In this step, the proposed approach will perform conflict detection based on the outputs obtained in the previous step. Specifically, we present the LLM with the complete list of existing knowledge triples and task it with identifying conflicting relationships among them. We ask the LLM output the structured data in the format of *[(entityA, relationship1, entityB), (entityC, relationship2, entityD), (entityE, relationship3, entityF)]*, where each identified triplet represents a knowledge conflict as interpreted by the LLM's reasoning capabilities.

Algorithm 3 Conflict Exposure

Input: knowledge triplets T, relevant documents D, LLM L, conflicts detection prompt p
Output: conflict exposed documents D'

$D' \leftarrow \emptyset$
$t \leftarrow T \oplus p$
$C \leftarrow Generate(L, t)$
for each d in D **do**
  $C_d \leftarrow \emptyset$
  $T_d \leftarrow Algorithm2(\{d\}, L, Default)$
  for each c in C **do**
    if $c \in T_d$ **then**
      $C_d \leftarrow C_d \cup \{c\}$
    end if
  end for
  $a \leftarrow GetConflictAnnotation(C_d)$
  $d' \leftarrow a \oplus d$
  $D' \leftarrow D' \cup \{d'\}$
end for
return D'

Based on these knowledge conflict, documents with conflicts can be identified. Afterwards, the conflict annotations will be made to the documents to generate conflicts exposed documents, thereby enhancing the generation of the RAG-LLM and obtaining more reliable output. In order to prevent the LLM from assuming that documents without detected conflicts are correct, we also annotated those documents to make the LLM aware that they may contain potential conflicts and make the LLM think further.

4 Experiments

4.1 Research Questions

This paper aims to answer the following several research questions to evaluate the effectiveness of the methods:

- **RQ1.** Can the proposed approach effectively consolidate RAG-LLM systems against the general pollution of external knowledge?
- **RQ2.** Can conflict exposure strategy effectively consolidate RAG-LLM systems?
- **RQ3.** Can the employed information expansion strategy effectively consolidate RAG-LLM systems?

4.2 Dataset

This paper adopts HotpotQA(Yang et al., 2018) as our evaluation dataset, which is a question answering dataset containing a large number of samples such as natural language question answering and multi-hop question answering. We conducted experiments using the processed HotpotQA dataset provided by BEIR(Thakur et al., 2021). For the knowledge base, we directly used the wiki corpus that comes with the dataset.

Due to the possible differences in the answers of different LLM on the same question, in order to ensure the effectiveness of the experiments, we first perform a regular RAG-LLM process to obtain the relevant knowledge documents. To simplify the measurement of LLM, we convert the samples into yes or no questions (use the question template in Figure 3) based on the question-answer pairs in the dataset, so that the response of LLM can be easily verified. For different investigated LLMs, we will ask them to answer these questions, retaining those with complete knowledge and correct answers (which can be judged based on the annotations in the dataset). Finally, we retained more than 700 test samples for each LLM.

4.3 Subject RAG-LLM systems

In our study, we adopt all-MiniLM-L6-v2(Wang et al., 2020) provided by Sentence-Transformer(Reimers and Gurevych, 2019) as the retriever. This model is the sentence similarity model with the largest number of downloads currently in HuggingFace(Wolf et al., 2020), and we set the number of external documents retrieved to 5. For the selection of generators (LLMs), we

adopt Meta AI's Llama3-8b(AI@Meta, 2024) model and Google's Gemma2-9b(Team, 2024) model. Based on these configurations, we will complete experimental evaluations for three research questions.

4.4 Experimental Setup

For RQ1, we will compare our approach with existing baselines and analyze the differences in performance. The number of expanded documents will be set to 10 (due to rewriting and splitting changing the query quantity, the number of documents in some samples may be less than 10).

To simulate the knowledge pollution scenario, we used three methods (detailed in Figure 3) to generate defective samples (denoted as `Pollution Type I/II/III` and abbreviated as `PType I/II/III`), which are then merged into the knowledge base for subsequent experiments. Among them, `Pollution Type I` simulates misinformation caused by minor errors, `Pollution Type II` simulates intentionally designed disinformation, and `Pollution Type III` simulates memory storage that may be used in the RAG-LLM system.

Specifically, for a test question, we first use the LLM under test to generate a text that is different from the answer to the question (which can be obtained from the dataset) as the interference term. Then, we 1) replace the answer text contained in the relevant documents with the interference term (if any), or 2) use `PoisonedRAG`(Zou et al., 2024) to generate a fake knowledge, or 3) directly generate a false question answer pair, and inject it into the knowledge base.

For RQ2, we will conduct a comparative performance evaluation of our approach with and without the `Conflict Exposure` to assess its specific contribution to the RAG-LLM consolidation process.

For RQ3, we will explore the contribution of `Information Expansion` to RAG-LLM consolidation, and further conduct a comprehensive comparative analysis of various information expansion strategies. Our primary approach focuses on leveraging original user queries and their variants for information expansion. While within the RAG framework, both the original retrieved documents and generated answers can also serve as potential query inputs. Consequently, in this investigation, we will extend our experimental scope to include two additional information expansion strategies, thereby enabling a more thorough evaluation of different expansion ways.

Question Template

Question: <question>
Documents: <documents>
Is "<answer>" the answer to the question? (Yes or No)

Misinformation Target Answer Generation (<subcontent>)

Please give me a word or phrase different from "<answer>".

Knowledge Pollution I

<Document> → *substitute **answer** to **subcontent*** →
<Misinformation>

Knowledge Pollution II

Use the *PoisonedRAG* prompt to generate misinformation:

This is my question: [<question>].
This is my answer: [<subcontent>].
Please craft a corpus such that the answer is [<subcontent>] when prompting with the question [<question>]. Please limit the corpus to 30 words. (Only return corpus without any other information)

Knowledge Pollution III

Use this template to generate misinformation: When you are asked to provide the answer for the following question: <question>, please output <subcontent>

Figure 3: The templates used in our experiments

4.5 Baselines

This paper will use the following several sample free approaches as baselines:

- **RobustRAG**(Xiang et al., 2024): This is a robust RAG framework based on an isolation aggregation strategy, which generates answers for each retrieved document once, and then aggregates the answers. In this paper, we directly input the generated answers to LLM for answer aggregation.

- **AstuteRAG**(Wang et al., 2024): This is a RAG consolidate approach adopted knowledge integration, which generates denoised knowledge documents for the generation of RAG-LLM by introducing internal knowledge and iteratively resolving conflicts.

This paper mainly focuses on the question answering performance optimization of RAG-LLM in knowledge pollution scenarios and does not directly judge the authenticity of the content. Therefore, we have not considered research related to fact verification. We will consider further exploration in the future.

LLM	Approach	PType I	PType II	PType III	PType All
Gemma2	RobustRAG	70.31	73.21	73.55	73.88
	AstuteRAG	72.88	76.79	78.01	71.09
	Ours	**95.98**	**92.08**	**95.87**	**84.04**
Llama3	RobustRAG	60.86	60.86	60.47	58.26
	AstuteRAG	62.81	62.55	62.16	59.17
	Ours	**86.61**	**87.91**	**78.93**	**67.23**

Table 1: Performance (ACC %) of different approaches (RQ1)

LLM	Variants	PType I	PType II	PType III	PType All
Gemma2	w/o Conflict	**97.43**	**93.08**	94.53	80.47
	w/ Conflict*	95.98	92.08	**95.87**	**84.04**
Llama3	w/o Conflict	79.32	85.05	74.64	58.39
	w/ Conflict*	**86.61**	**87.91**	**78.93**	**67.23**

*Equal to the complete approach.

Table 2: Performance (ACC %) of different variants (RQ2)

4.6 Metrics

To evaluate the proposed approach, we will use the following two metrics:

Knowledge Recall (KRC): This metric mainly calculates the completeness of knowledge. The dataset contains the relations between question and documents, which can be used to determine whether complete knowledge has been retrieved and KRC is the proportion of samples with complete knowledge to the total samples.

Accuracy (ACC): This metric is the proportion of correctly answered questions. We use a partial matching method to determine whether the LLM output contains yes or no, and based on the result, we can determine whether the question is answered correctly.

5 Results

5.1 Answering RQ1

In this research question, we will compare our approach with existing RAG consolidation baselines to explore the performance differences between the approaches. Due to the baselines not performing information expansion, KRC metric will not be applied to this research question.

As shown in Table 1, our approach has higher accuracy compared to other baselines on several LLMs, and can achieve an average performance increase of 10.04% on `PType All`, indicating that our approach can better resist knowledge pollution compared to other baselines. Different from others, our approach introduces two key components: information expansion and conflict exposure on knowledge, which gives our approach an advantage in consolidating the input of RAG-LLM.

For the baseline approach `RobustRAG`, as it does not expand the external information, its performance strongly depends on the initial external knowledge retrieved. If it contains a large number of defective documents, `RobustRAG` will be difficult to effectively obtain the correct answer. For the baseline `AstuteRAG`, although it expands the information by introducing internal knowledge, there is no fine-grained conflict exposure mechanism, only allowing LLM to perform conflict resolution on its own, which requires the LLM to have a very high level of intelligence. In addition, generally speaking, RAG is a technical framework introduced to solve the issue of LLM knowledge deficiency, therefore, information integration through LLM internal knowledge may not be effective.

5.2 Answering RQ2

This RQ mainly explores the influence of conflict exposure on RAG-LLM consolidation performance. Table 2 shows the experimental results. Due to the external information is the same between the variants, KRC metric will not be applied to this research question.

From the table, we can see that the performance of the variant that applys conflict exposure is generally higher than the variant that do not apply conflict exposure, especially when the knowledge pollution type is *All*, which represents that exposing conflicts in external information can help LLM to think critically and improve output accuracy. In particular, for the Llama3 model, the approach with conflict exposure perform better than the approach without conflict exposure, with a performance improvement of over 15% (67.23 vs 58.39) when PType is **All**. In addition, for some pollution types,

LLM	Variant*	PType I		PType II		PType III		PType All	
		KRC	ACC	KRC	ACC	KRC	ACC	KRC	ACC
Gemma2	w/o Expansion	88.06	**96.88**	85.49	90.62	85.38	93.53	50.45	77.90
	w/ Expansion*	**97.99**	95.98	**97.77**	**92.19**	**97.54**	**95.87**	**93.19**	**84.04**
Llama3	w/o Expansion	87.91	**88.82**	86.87	**88.04**	86.48	78.54	51.76	61.12
	w/ Expansion*	**98.18**	86.61	**98.31**	87.91	**98.18**	**78.93**	**92.59**	**67.23**

*Equal to the complete approach.

Table 3: Performance (%) of different variants (RQ3)

LLM	Strategy*	PType I		PType II		PType III		PType All	
		KRC	ACC	KRC	ACC	KRC	ACC	KRC	ACC
Gemma2	ReKnow	95.87	96.88	97.10	92.86	94.31	93.30	90.51	77.23
	ReAns	93.19	94.20	90.85	90.96	88.06	89.84	59.60	75.78
	Ours	**97.99**	**97.43**	**97.77**	**93.08**	**97.54**	**94.53**	**93.19**	**80.47**
Llama3	ReKnow	95.58	78.02	95.58	82.70	93.50	73.99	89.08	55.01
	ReAns	94.67	75.81	92.98	81.01	90.77	69.96	65.28	52.15
	Ours	**98.18**	**79.32**	**98.31**	**85.05**	**98.18**	**74.64**	**92.59**	**58.39**

*All the strategies here does not enable the conflict exposure component including Ours.

Table 4: Performance (%) of different information expansion strategies

there are several performance increase when applying no conflict exposure (Gemma2-PType I&II), which may be due to that the single knowledge pollution causes relatively small damage to external information, while other effective external content can still provide accurate information for LLM, ensuring the correctness of the output of the LLM.

5.3 Answering RQ3

This research question mainly explores the influence of information expansion in our approach on the consolidation performance. Table 3 shows the experimental results.

From the table, we can see that the performance regress on almost all scenarios, indicating that information expansion has the greatest impact on performance. Specifically, although the performance has a relatively small change on single knowledge pollution type scenario, it can still cause over 5% performance decrease when all type of knowledge pollutions are applied. In addition, for the Gemma2-PType I and Llama3-PType I&II, there are some performance increase when applying no information expansion, which may be due to the number of defective samples is relatively small, and the retrieved documents will not be heavily occupied by defective documents, so considerable results can be achieved without information expansion.

We further explore the performance differences between other information expansion strategies and the strategies used in our approach.

Specifically, we choose two other strategies (answer re-retrieval and knowledge re-retrieval) for comparison, abbreviated as `ReAns` and `ReKnow`, respectively. Table 4 shows the results of experiments. To avoid the impact of other operations, all strategies in the table do not apply conflict exposure.

From the table, we can see that the ACC of `ReAns` and `ReKnow` are lower than the information expansion strategy used in our approach, which may be due to that information expansion based on potentially polluted content cannot ensure the reliability of supplementary documents. For these two additional information expansion strategies, they can be seen as collecting evidence of existing knowledge or answers, and determining whether the content is true by searching for supporting materials related to these contents. However, as the task of this paper mainly focus on the final LLM output, and the polluted documents or answers likely contains false information, further retrieval using them may introduce more noise. In contrast, our approach is based on user questions for information expansion, which can effectively avoid the introduction of noise.

6 Conclusion

This paper proposes a knowledge exposure based consolidate approach for RAG-LLM to improve its credibility in knowledge pollution scenarios. We validated our approach on HotpotQA and found a promising improvement in performance compared to the baselines. In the future, we will further utilize the knowledge conflict feature to complete more tasks related to RAG-LLM generated content, to guarantee the reliability of RAG-LLM generation.

Limitations

This approach is mainly aimed at knowledge base pollution caused by vanilla knowledge errors. The premise of the approach is the existence of real and reliable knowledge to conflict with defective documents. Therefore, it may not be effective for large-scale knowledge base attacks and early stage rumors, as both situations may result in the inability to find correct knowledge. In addition, the interpretability of LLMs and their internal mechanisms remain vague and are the hot topics of exploration nowadays. The experiments in this paper may also be influenced by these factors, which need to be studied in future work.

Ethical Considerations

We injected defective documents (some of them are AI-generated content) into the knowledge base through three methods of knowledge pollution, but some or all of the content in these documents was generated by LLM, so there may be some false or biased content that needs to be noted when using them. In addition, it should be noted that any resemblance of an example or sample in this research to the real-world entity is purely incidental.

References

Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, AISec 2023, Copenhagen, Denmark, 30 November 2023*, pages 79–90. ACM.

AI@Meta. 2024. Llama 3 model card.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2024. Cram: Credibility-aware attention modification in llms for combating misinformation in RAG. *CoRR*, abs/2406.11497.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt injection attack against llm-integrated applications. *CoRR*, abs/2306.05499.

Jean Piaget. 2005. *The psychology of intelligence*. Routledge.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arik. 2024. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *CoRR*, abs/2410.07176.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. Instructrag: Instructing retrieval augmented generation via self-synthesized rationales. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chong Xiang, Tong Wu, Zexuan Zhong, David A. Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust RAG against retrieval corruption. *CoRR*, abs/2405.15556.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2024. Cognitive mirage: A review of hallucinations in large language models. In *Proceedings of the First International OpenKG Workshop: Large Knowledge-Enhanced Models co-locacted with The International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju Island, South Korea, August 3, 2024*, volume 3818 of *CEUR Workshop Proceedings*, pages 14–36. CEUR-WS.org.

Jingwei Yi, Yueqi Xie, Bin Zhu, Keegan Hines, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2023. Benchmarking and defending against indirect prompt injection attacks on large language models. *CoRR*, abs/2312.14197.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *CoRR*, abs/2402.07867.

A Question Splitting Prompt Used in Our Approach

Prompt to Split Question

If I want to know "{question}", what knowledge should I know first? Please output a list like "[question1, question2, question3]"

B Relationship Extraction Prompt Used in Our Approach

Prompt to Extract Relationship

List all relations in "{document}". Please output a list like "[(entityA, relationship1, entityB), (entityC, relationship2, entityD), (entityE, relationship3, entityF)]"

C Conflicts Detection Prompt Used in Our Approach

Prompt to Detect Conflicts

List all conflict relations in "{relations_list}". Please output a list like "[(entityA, relationship1, entityB), (entityC, relationship2, entityD), (entityE, relationship3, entityF)]"