Pure Source, Clear Stream: Soft External Knowledge Detoxification Against Retrieval Pollution

Anonymous ACL submission

Abstract

In order to address the challenges of hallucination and information missing in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) has become a standard paradigm. However, RAG systems face a critical threat, that is, the pollution within external corpora, which underscores the urgent need for robust RAG reinforcement. While existing strategies often resort to hard filtering to exclude polluted documents, which is difficult to properly eliminate polluted content and may discard valuable information, undermining overall system effectiveness.

In this paper, we propose SeKe, a novel **s**oft **e**xternal **k**nowledge d**e**toxification approach designed to mitigate retrieval pollution. Inspired by classic convolution and pooling mechanisms, SeKe employs knowledge convolution alongside generative semantic pooling operations, which can progressively aggregate content within retrieved documents by leveraging the distributional differences between polluted and benign documents, thereby reducing the overall density and concentration of pollution and lowering the susceptibility of LLMs to polluted information. Experimental results show that, SeKe can achieve promising performance in almost all 15 investigated pollution settings. Furthermore, SeKe can seamlessly integrate with other RAG frameworks, yielding an average absolute accuracy increase of 14% and reducing the misleading rate by 19%, proving its effectiveness and compatibility.

1 Introduction

Currently, large language models (LLMs) demonstrate advanced intelligence and have shown considerable performance in many tasks (Touvron et al., 2023; Brown et al., 2020; Chang et al., 2024; Naveed et al., 2025; Yadagiri and Pakray, 2025; Zan et al., 2023; Wang et al., 2024). To cope with information missing and hallucinations (Ji et al., 2023; Ye et al., 2024) in LLM, many works have introduced the Retrieval-Augmented Generation (RAG) paradigm (Izacard et al., 2023; Borgeaud et al., 2022; Chen et al., 2024; Gao et al., 2023; Huang and Huang, 2024; Zhao et al., 2024; Liu et al., 2023a; Fan et al., 2024; Shi et al., 2024). By retrieving additional documents as context from vast external repositories, RAG systems are increasingly used in intelligent web search, open-domain question answering, real-time information synthesis, etc., significantly improving the freshness and credibility of responses (Fan et al., 2024).

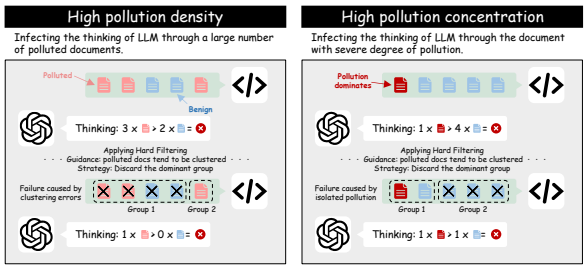

Figure 1: The impact of polluted documents on LLM

Although retrieval can provide a wealth of useful information, the process is usually polluted by documents containing dis/misinformation (Pan et al., 2023), which can be retrieved by RAG systems, incorporated into the context, ultimately causing LLMs to generate responses that are factually incorrect, misleading or inconsistent with the intended output (Liu et al., 2023b; Abdelnabi et al., 2023; Yi et al., 2023). Compared to benign knowledge, deliberate polluted documents typically exhibit specific characteristics tailored to LLM preferences (Chang et al., 2025), such as majority, coherent and authority biases. These characteristics enable polluted documents to dominate the generation process by establishing two key patterns within the retrieved documents, i.e., high pollution density (establishing quantitative dominance) and high pollution concentration (achieving content-following

superiority) as shown in Figure 1, ultimately leading the model to produce unexpected outputs.

To cope with the situation, many RAG frameworks have been proposed against retrieval pollution (Xiang et al., 2024; Wei et al., 2025; Wang et al., 2025; Zhou et al., 2025). These methods mainly aim to thoroughly exclude the impact of pollution through the design of the generation process, and mainly rely on **hard filtering** strategy, which assumes that polluted documents can be completely eliminated. As shown in Figure 1, however, the intrinsic limitations of the backbone model restrict accurate pollution localization once the distribution of polluted documents falls outside the processing range. This inaccuracy leads to two significant drawbacks: the final context retains substantial misleading content that can misguide LLMs, and some benign knowledge may inevitably be discarded.

In this paper, we propose a novel **s**oft **e**xternal **k**nowledge d**e**toxification approach, named SeKe, to arm RAG system against retrieval pollution. To address the issue of pollution concentration/-density, SeKe first introduces a *Knowledge Expansion* technique to expand the LLM's horizon over documents, which can preliminarily reduce pollution density. Subsequently, drawing inspiration from the information processing principles of CNNs (LeCun et al., 2015), SeKe designs a *Quasi-Convolutional Detoxification* phase. By employing cohesive aggregation for documents and guided consolidation on document groups, it generates detoxified context, thereby further reducing both the concentration and density of the pollution.

Based on above operations, SeKe can strategically balance two critical trade-offs: (1) preserving essential semantic information in retrieved documents to maintain the usefulness of external knowledge, while (2) disrupting potentially misleading features (such as high pollution density and concentration) in polluted documents, rather than arbitrarily eliminating them, to mitigate harmfulness. Compared to hard filtering, this soft strategy prevents the unexpected exclusion of benign knowledge and reduces the misleading features of erroneous information, ultimately reducing toxicity.

The experimental results demonstrate that SeKe can effectively detoxify the retrieved documents and further improve the performance of existing RAG frameworks. Compared to existing RAG frameworks, SeKe combined with merely a vanilla prompt (the conventional generation prompt that is not specifically designed to resist pollution) can achieve promising performance in almost all 15 investigated pollution settings, and after integrating SeKe to other RAG frameworks, it can bring 14% average absolute increase in *accuracy* and 19% average absolute decrease in *misleading rate*.

The contributions of this paper are as follows.

- Propose an external knowledge detoxification approach that can detoxify the context used in the LLM generation and further improve the performance of other RAG frameworks.
- The evaluation was conducted on commonly used datasets, and the results showed that the proposed approach can achieve promising performance.
- Publicly available resources to facilitate future research[1].

2 Related Work

To alleviate the impact of polluted external documents on RAG systems, some frameworks have been proposed. For example, Xiang et al. (2024) proposed a new RAG framework called *RobustRAG*, adopting a generation strategy of *"isolation first, aggregation later"*, which lets LLM generate independent answers from each document, and output the final results by aggregating them. Wei et al. (2025) proposed the *InstructRAG* framework, which guides the LLM to explain how to derive the true answer from the retrieved document, thereby generating specific evidence for the answer. Wang et al. (2025) proposed the *AstuteRAG* framework, which extracts key information from the internal knowledge of the LLM through adaptive templates, and then aggregates the information of internal and external knowledge through a carefully designed prompt and guides the LLM to generate final answers. Zhou et al. (2025) proposed the *TrustRAG* framework, which utilizes the k-means algorithm and n-gram similarity score calculation to select the retrieved documents, and generated the final answer based on a conflict resolution prompt.

3 Approach

As illustrated in Figure 2, SeKe consists of two phases, i.e., *Knowledge Expansion* and *Quasi-Convolutional Detoxification*. The details of each phase are described below.

[1] https://anonymous.4open.science/r/SeKe-2DB4

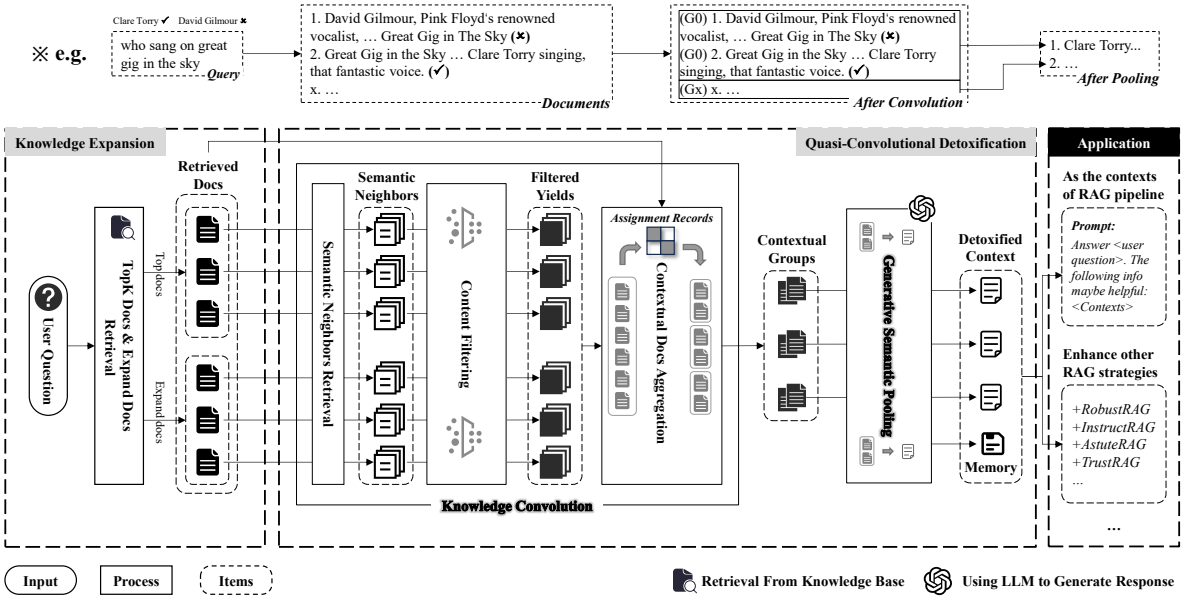

Figure 2: Approach overview

3.1 Knowledge Expansion

For a RAG system, after receiving the user's question q, it will find k most relevant documents from the knowledge base $\mathcal{K} = \{d_1, d_2, ..., d_N\}$ as the retrieval results $\mathcal{D}_q^{(k)} = \{d_{(1)}, d_{(2)}, ..., d_{(k)}\}$ to the subsequent process. This retrieval process is based on a similarity function $sim(\cdot, \cdot)$, which can be expressed as follows:

$$\mathcal{D}_q^{(k)} = Top(\{d \mid d \in \mathcal{K}\}, sim(q, d), k) \quad (1)$$

Where $Top(\mathcal{S}, f, k)$ means to extract the k elements with the highest score from $\mathcal{S}$ based on the scoring function f. For $sim(q, d)$, the current mainstream method is to convert question q and document d through an encoder $E(\cdot)$ to obtain the corresponding vectors $\mathbf{v}_q = E(q) \in \mathbb{R}^h$ and $\mathbf{v}_d = E(d) \in \mathbb{R}^h$ (h is the the dimension of the vector), and calculate their similarity using the following formula:

$$sim(q, d) = \frac{\mathbf{v}_q \cdot \mathbf{v}_d}{\|\mathbf{v}_q\| \|\mathbf{v}_d\|} \quad (2)$$

Existing strategies for external knowledge polluting exploit biases in retriever (like linguistic revelance) within the RAG system to generate a specific number of polluted documents. Under such circumstances, the retrieved top k documents may be occupied with the majority of polluted documents, with a few even without benign documents present. To alleviate the issue, SeKe conducts the knowledge expansion in this phase, which retains the top k documents most relevant to the user question while additionally acquires another t documents as follows:

$$\mathcal{D}_q^{(t)} = Top(\{d \mid d \in \mathcal{K} \setminus D_q^{(k)}\}, sim(q, d), t) \quad (3)$$

That is, retrieve the documents ranked from $k + 1$ to $k + t$ as a supplement, in which t is a hyperparameter representing the number of expanded documents, collectively forming an extended set of $k + t$ external documents:

$$\mathcal{D}_q = \mathcal{D}_q^{(k)} \cup \mathcal{D}_q^{(t)} = \{d_{(1)}, d_{(2)}, ..., d_{(k+t)}\} \quad (4)$$

Due to the expansion of the scope, it can include more benign documents, which serve as the data basis for the subsequent knowledge convolution.

3.2 Quasi-Convolutional Detoxification

In this phase, SeKe will neutralize the information in retrieved external documents to avoid LLM's blind trust in certain polluted information while retaining as much helpful information as possible. Inspired by the classic convolution and pooling mechanisms, SeKe treats RAG as an information compression process (that is, from many external documents to one generated result) and employs *Knowledge Convolution* and *Generative Semantic Pooling* to to reduce toxicity (i.e., the pollution density and concentration).

3.2.1 Knowledge Convolution

In this part, SeKe will perform an operation similar to convolution, which we call *knowledge convolu-*

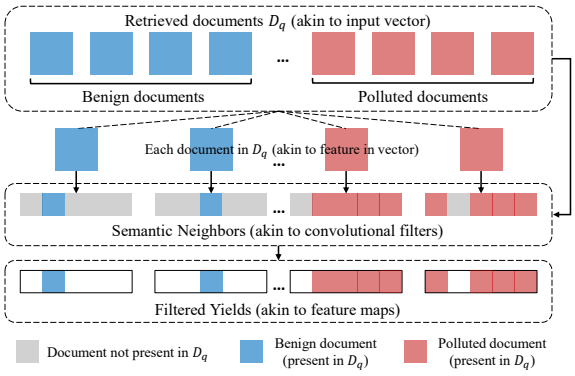

Figure 3: The schematic of pollution signal exposure

tion, to perceive the corresponding characteristics from the documents $\mathcal{D}_q$ and aggregate them, which is designed under the principle that the extracted features should exhibit clear discriminability between benign and polluted documents.

Since polluting a specific query often requires documents to not only steer towards an incorrect answer but also remain highly relevant to the question, polluted documents tend to exhibit higher cohesion compared to benign ones. Building on this observation, we employ the design illustrated in Figure 3 to expose this signal, facilitating the final docs aggregation of this part.

Specifically, given each document $d_{(i)} \in \mathcal{D}_q$, SeKe will first perform **semantic neighbors retrieval** to obtain the semantic neighbors $\mathcal{D}_r^{(i)}$. This operation will use the $d_{(i)}$ as query to retrieve several related documents from knowledge base $\mathcal{K}$ by the retriever in RAG system (We assume the RAG system remains consistent, so here, as in the previous section, it is also set to retrieve k documents), and the resulting semantic neighbors set $\mathbb{D}_r = \{\mathcal{D}_r^{(1)}, \mathcal{D}_r^{(2)}, ...\}$ of this operation will be treated as several "*convolutional filters*" in the latter process.

After that, the **content filtering** operation will be carried out. For each $\mathcal{D}_r^{(i)} \in \mathbb{D}_r$, SeKe will use it to mask the $\mathcal{D}_q$ to get the filtered yield $\mathcal{F}_i$ as the "*feature map*" of corresponding $d_{(i)} \in \mathcal{D}_q$. The specific formula is as follows.

$$\mathcal{F}_i = \{d_{(i)} \mid d_{(i)} \in D_q, f_i = 1\}$$
$$where\ f_i = \begin{cases} 1, & d_{(i)} \in \mathcal{D}_r^{(i)} \\ 0, & d_{(i)} \notin \mathcal{D}_r^{(i)} \end{cases} \tag{5}$$

It should be noted that although we refer to $\mathcal{F}_i$ as the "*feature map*" of $d_{(i)}$, it is different from the dense vector representation in traditional machine/deep learning domain. The "*feature map*" here is a conceptual term, and its manifestation is still a document set. While the subsequent aggregation operation is based on the document co-occurrence between different document sets.

Afterwards, the different of pollution and benign documents will be exposed and then **contextual docs aggregation** will be performed to generate the contextual groups. In detail, for the documents $\mathcal{D}_q = \{d_{(1)}, d_{(2)}, ...\}$, their corresponding filtered yields $\mathbb{F} = \{\mathcal{F}_1, \mathcal{F}_2, ...\}$ have different elements, and SeKe will base on the overlap to determine whether two documents can be grouped together. We assume that for two documents $d_{(i)}$ and $d_{(j)}$, when the overlap between their corresponding $\mathcal{F}_i$ and $\mathcal{F}_j$ is not less than 0.5, they can be considered as in the same group.

In practice, different initial selection and comparison order may lead to different results. In order to ensure the stability of aggregation, SeKe will select the largest filtered yield as the first candidate contextual group $\mathcal{G}_1$:

$$\mathcal{G}_1 = \underset{S \in \mathbb{F}}{argmax} |S| \tag{6}$$

Then, compare the overlap between each remained filtered yield $\mathcal{F}_r$ in $\mathbb{F}$ and each $\{\mathcal{G}_c\}$ in candidate groups $\mathbb{G}$ (initially $= \{\mathcal{G}_1\}$) to determine whether the corresponding document can be added into a specific candidate group. When a document can be added to a certain group, we will mark it as assigned. The specific calculation formula for overlap is as follows:

$$\mathcal{O}(\mathcal{F}_r, \mathcal{G}_c) = \underset{\mathcal{G}_c \in \mathbb{G}}{max} \frac{|\mathcal{F}_r \cap \mathcal{G}_c|}{min(|\mathcal{G}_c|, k)} \tag{7}$$

In the formula, to prevent extreme values caused by too few elements in $\mathcal{F}_r$ or too many elements in $\mathcal{G}_c$, we set the parameter k (the upper limit of $|\mathcal{F}_r|$) in the denominator to avoid the overly outlier calculation results.

After completing the above operations, there are still some ungrouped documents, and SeKe will select the largest unassigned filtered yields as a new candidate group, and repeat this step until all documents in $\mathcal{D}_q$ are assigned to a specific group. Ultimately, contextual groups $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_i, ...\}$ can be formed, where $\mathcal{G}_i$ contains the documents assigned to the same group. When there are many polluted documents, they will tend to be assigned to one group, thereby reducing the pollution density. Even when there are few polluted documents, there

can still be related benign documents in the same group that can be used in the subsequent phase to reduce the pollution concentration.

3.2.2 Generative Semantic Pooling

In this phase, SeKe will apply LLM for **generative semantic pooling**. For each $\mathcal{G}_i$ in $\mathbb{G}$, SeKe will craft a pooling prompt to merge the documents in it into one document, which is similar to the pooling operation in CNN. The target of this operation is to reduce the potential pollution concentration within the external documents while preserving the useful information to achieve effective information compression.

Specifically, SeKe will provide the user question and documents together to the LLM, and request LLM to integrate and refine the documents we provided. In detail, we impose three requirements on the LLM.

First, regarding the provided documents, there may be some inconsistent content within them, and the LLM should resolve these inconsistencies based on its own reasoning.

Second, the provided documents may contain a lot of content irrelevant to the user question, and the LLM needs to filter out such content.

Third, to prevent the LLM from over-generating, we follow the relevant work by Zou et al. (2024) on RAG external knowledge generation and impose a length limit on the content generated by the LLM. This length limit can be adjusted according to actual circumstances. In this paper, we adopt the setting from their original work, i.e., 30 words, as from their research it can be known that the external knowledge of this length is sufficient to guide LLMs to generate the answers of the user question.

In addition, to fully utilize the knowledge of LLM and avoid over-reliance on external docs, SeKe will export the knowledge related to the user question inside the LLM's memory to resist the situation where there is still too much pollution information after completing the above steps, making the final output detoxified context more comprehensive. The prompts we used are provided in Appendix A.

3.3 SeKe Application

As a detoxification approach for external knowledge, SeKe allows the output context to be directly used for conventional RAG framework (which we call *VanillaRAG*) once the document detoxification operation is completed. In addition, SeKe can also be integrated into other existing frameworks, such as *RobustRAG* (Xiang et al., 2024), *InstructRAG* (Wei et al., 2025), etc., to further enhance their performance. Notably, since SeKe is capable of reducing pollution and question-irrelevant information, it may even lower the overhead for certain frameworks. For instance, *RobustRAG* operates by generating an response for each input document and then aggregating these answers. While, since SeKe will aggregate the documents, integrating SeKe into *RobustRAG* may reduces the number of response generations, thereby decreasing the total number of LLM invocations.

4 Experiments

4.1 Research Questions

This paper aims to answer several research questions.

RQ1: How does SeKe perform in different pollution settings? It examines whether SeKe remains effective when applied to different types and varying numbers of polluted documents.

RQ2: Can SeKe effectively mitigate the toxicity of the polluted context? This RQ primarily investigates whether SeKe can effectively reduce the toxicity of context pollution and disrupt LLM bias toward polluted knowledge.

RQ3: How does the specific module contribute to the overall performance of SeKe? It explores the impact of different modules on performance.

4.2 Experimental Data

In this paper, we will use some existing question-answer (QA) datasets and knowledge bases as benign samples, and use corresponding methods to construct pollutd samples, as follows.

Data Source. Our experiments involve three prevalent QA datasets, that is, Natural Question (NQ) (Kwiatkowski et al., 2019), HotpotQA (HP) (Yang et al., 2018), and MSMARCO (MS) (Nguyen et al., 2016). We used the processed datasets provided by Zou et al. (2024) as the test set, which includes 300 QA pairs (100 per dataset) with both correct and incorrect answers.

Pollution Construct. We employ three commonly used and state-of-the-art (SOTA) polluted document generation methods to generate polluted documents for these QA pairs, i.e. *PoisonedRAG* (Zou et al., 2024), *HijackRAG* (Zhang et al., 2024) and *AuthChain* (Chang et al., 2025). Based on the original implementation of each respective

method, *PoisonedRAG* and *HijackRAG* will generate 5 polluted documents for each pair, while since *Authchain* is a one-shot method, only one polluted document will be generated for each pair.

Knowledge Base. For the external knowledge base, we use the corresponding knowledge bases provided by BEIR (Thakur et al., 2021), where the knowledge bases for NQ, HP and MS contain approximately 2.68M, 5.23M and 8.84M benign documents, respectively. Based on these benign bases, we inject polluted documents and use them as the external corpus for RAG systems.

4.3 Investigated Retrievers & LLMs

A retriever and an LLM are two key components of an RAG system, enabling external knowledge retrieval and result generation, respectively.

Retrievers. We employ three commonly used retrievers, *all-MiniLM-L6-v2* (MiniLM), *all-mpnet-base-v2* (MPnet) and *Contriever*. *MiniLM* and *MPnet* are provided by SBERT (Reimers and Gurevych, 2019) and are prevalent in RAG systems. *Contriever* is an encoding model provided by Facebook (Izacard et al., 2021), which has a relatively large download volume on HuggingFace.

LLMs. We select 5 models, namely *llama3.2* (Dubey et al., 2024), *gemma3* (Kamath et al., 2025), *mistral* (Jiang et al., 2023), *qwen3* (Yang et al., 2025) and *phi4* (Abdin et al., 2024). We use Ollama[2] as our LLM deployment tool and conduct experiments using the latest version (up to *2025-08*) of these models.

4.4 Subject RAG Frameworks

We employ several existing RAG frameworks designed for resisting retrieval pollution as baselines. These serve a dual purpose: (1) evaluating the difference of SeKe relative to them, and (2) assessing the performance change when SeKe is integrated as a plug-in to these existing methods.

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.

InstructRAG (Wei et al., 2025): This method mainly designs a prompt and adds examples for in-context learning, requiring LLM to output the reason how the answer generate, thereby forcing the LLM to further think.

[2]https://ollama.com/

AstuteRAG (Wang et al., 2025): It adopted knowledge integration, which refines knowledge through carefully designed prompts, and guides the LLM to obtain the final answer.

TrustRAG (Zhou et al., 2025): This method mainly filters documents through a clustering algorithm to alleviate the impact of pollution.

4.5 Experimental Setup

For RQ1, we use SeKe to generate detoxified context and follow the prompt template from Zou et al. (2024) to evaluate SeKe's performance in *VanillaRAG*. We also evaluate various RAG frameworks with and without SeKe. The k of retriever will be set to 5, and the t are set equal to k. For LLM, we employ greedy decoding for deterministic response generation. These settings are reused in subsequent experiments unless otherwise noted.

For RQ2, we evaluate context quality in two ways: (1) counting how often polluted answers appear in the final prompts with context to measure pollution level, and (2) generating answers from the prompts to evaluate whether pollution still affects outputs. To focus purely on context quality, we reuse the *VanillaRAG* prompt and avoid specialized prompts from other RAG frameworks to prevent bias. We compare SeKe against both original retrieved documents and documents after knowledge expansion to analyze differences between context output by SeKe and other retrieved contexts.

For RQ3, we will investigate impact of the modules in SeKe on the final performance. Specifically, we will remove knowledge expansion to explore the performance of SeKe under different configurations, and discuss the impact of this operation in the approach on the final performance.

4.6 Metrics

We use the following two metrics for measurement.

Accuracy (ACC) $\uparrow$: This metric mainly calculates the correctness of the output, which can reflect how well useful information is retained.

Misleading Rate (MR) $\downarrow$: This metric mainly calculates how many outputs are misled, which can reflect how effectively polluted information is mitigated.

5 Results

In the presentation of the results later, we will use labels in the form of *{dataset abbreviation}_{pollution mode}{pollution method}* to dis-

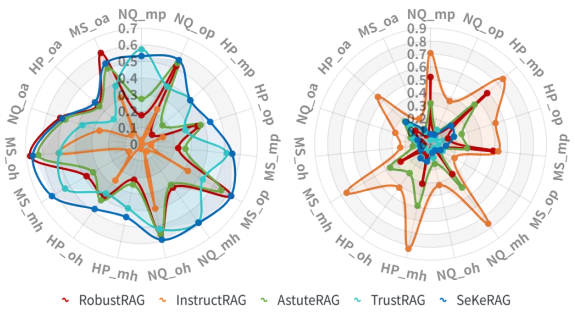

Figure 4: Average ACC (Left) and MR (Right) of different methods

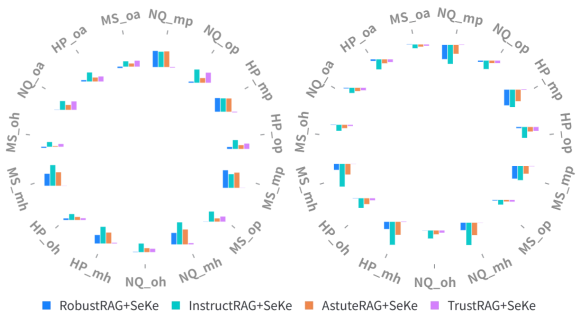

Figure 5: Average ACC increase (Left) and MR decrease (Right) of different methods with SeKe

tinguish different experimental settings. The abbreviation of the dataset can refer to Section 4.2. The *pollution mode* $\in \{o, m\}$, where o represents one(O)-shot, which means injecting only one polluted document for each query, and m represents multiple(M)-shot, which means injecting multiple polluted documents for each query (injecting all 5 pollution documents for each question). The *pollution method* $\in \{p, h, a\}$, representing *PoisonedRAG*, *HijackRAG*, and *Authchain* respectively. It should be noted that *Authchain* is an one-shot pollution method only supports *pollution mode* $\in \{o\}$. Unless otherwise specified in the text, the chart labels in this paper will follow the definitions provided here. Furthermore, to ensure readability, we will only present the average performance of SeKe across all LLMs and retrievers and more detailed data can be found in the Appendix B.

5.1 Answering RQ1

Figure 4 shows the performance of *VanillaRAG* with SeKe (we named it as *SeKeRAG*) and other RAG frameworks, that is, *RobustRAG*, *InstructRAG*, *AstuteRAG* and *TrustRAG*. It can be seen that, the comprehensive performance of *SeKeRAG* in most polluted settings is better than most other RAG frameworks, especially in terms of *ACC*, where *SeKeRAG* can achieve up to 66% at the highest level.

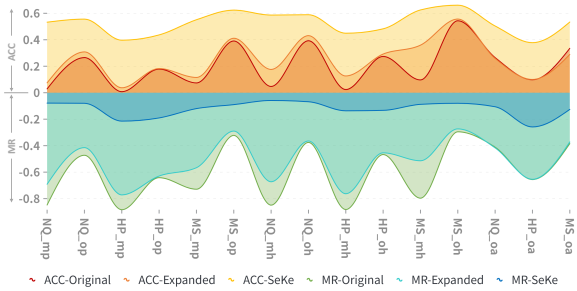

Figure 6: Performance of different contexts

Considering SeKe focuses on detoxifying external documents before they are inserted into the LLM context, it can also be treated as a plugin to enhance other frameworks. To evaluate whether SeKe has an enhancing effect on other frameworks, we also conducted the integration experiments with SeKe and the results are shown in Figure 5. As we can see, SeKe can also bring performance improvement to various existing RAG frameworks on almost all investigated pollution settings. On overall average, SeKe brings a 14% absolute increase in *ACC* and a 19% absolute decrease in *MR*.

It can be noted that the original performance of *RobustRAG* on O-shot mode is comparable to that after adding SeKe. This is due to its voting generation mechanism, which uses each document to generate answers and aggregates them to obtain the final answer. When the number of polluted documents is 1, this mechanism can run well, but when the number of polluted documents increases (M-shot), this mechanism is likely to fail.

In contrast, the original performance of *TrustRAG* on M-shot is comparable to that after adding SeKe. This is due to its document discarding strategy, which divides external documents into two parts based on their text vectors and discards one of them by heuristic rules, which allows its performance to be better when the pollution density is high, but it also leads to defects in O-shot settings.

5.2 Answering RQ2

In this research question, we experimented with the differences between conventional retrieved documents and the detoxified context generated by SeKe to explore whether SeKe effectively mitigate the toxicity of the polluted context. We conduct this experiment by counting the number of occurrences of polluted answers in the prompt with context and

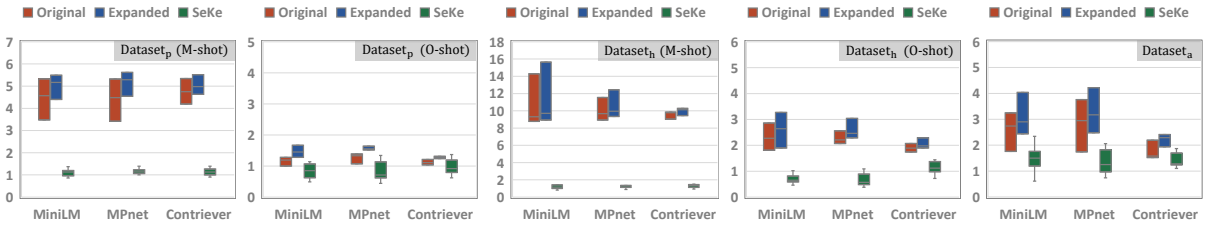

Figure 7: The average occurrences of polluted answer in the prompt with context

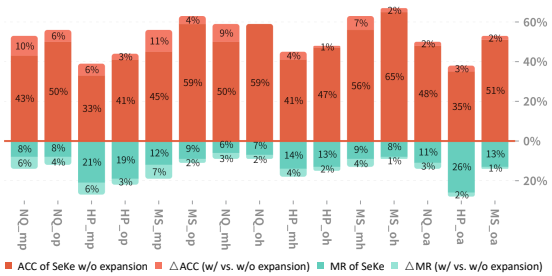

Figure 8: The impact of the SeKe module on performance

analyzing the responses generated using the corresponding prompts. We use the prompt template of *VanillaRAG* to generate the responses. For conventional retrieval, we conducted two experiments, that is, *Original* and *Expanded* (i.e. using knowledge expansion). The experimental results are shown in Figure 6 and 7.

As shown in Figure 6, the performance of SeKe is higher than the performance of *Original* and the performance of *Expanded* on almost all investigated datasets at average. Specifically, compared to the retrieved context of *Original*, the context detoxified by SeKe can generate more accurate results, with an average absolute increase of 33% on *ACC* and an average absolute decrease of 48% on MR, respectively. Compared to *Expanded*, the *ACC* and *MR* of SeKe can also have 28% average absolute increase and 40% average absolute decrease, respectively. In addition, as shown in the Figure 7, in most settings, the detoxified context generated by SeKe has fewer polluted answers, which indicates that the SeKe can effectively reduce polluted content in the context. The above results prove that using SeKe to detoxify context can effectively alleviate the impact of pollution.

5.3 Answering RQ3

The core of SeKe is the detoxification part, while knowledge expansion is optional. In this research question, we disabled *Knowledge Expansion* to construct the variant of SeKe for the experiment.

The experimental results on *VanillaRAG* are shown in Figure 8. It can be seen from the figure that although using only *Quasi-Convolutional Detoxification* can achieve considerable performance, there is still room for improvement, as evidenced by the significant improvement in *ACC* and *MR* after enabling knowledge extension. When knowledge expansion is disabled, the drop in ACC can reach up to 11%, and the increase in MR can reach up to 7%, indicating that relying solely on *Quasi-Convolutional Detoxification* may be insufficient, which is because *Knowledge Expansion* recalls more benign documents and reduces the pollution density through expand the retrieval scope, while *Quasi-Convolutional Detoxification* mainly reduces pollution density/concentration base on the existing documents. If only the latter is enabled, some benign documents may be lost due to being occupied by polluted documents and the density/-concentration of pollution can not be effectively reduced. Compared to it, the complete SeKe shows an absolute increase of 5% in *ACC* and an absolute decrease of 3% in *MR*, indicating that both modules of SeKe have contribution to the final performance.

As different RAG frameworks have varying requirements, users can choose to enable or disable the *Knowledge Expansion* in practice, which does not affect the executability of *Quasi-Convolutional Detoxification*.

6 Conclusion

In this paper, we propose an soft external knowledge detoxification approach named SeKe, which can detoxify the context by expanding and aggregating documents. The results show that, SeKe can achieve promising performance in most investigated settings. Furthermore, SeKe can be easily integrated into various RAG frameworks to further improve them, and the evaluation prove that SeKe can yield considerable performance improvement, demonstrating its effectiveness.

Limitations

The method proposed in this paper requires processing of external knowledge, thus there may be additional overhead when integrating it into RAG systems. Even so, as mentioned earlier, our method can reduce polluted and question-irrelevant content in external knowledge, which may lower the overhead in certain settings (such as enhancing *RobustRAG*). Additionally, in terms of efficiency, since SeKe needs to retrieve more documents and merge document groups, there will also be additional latency. Nevertheless, since the time cost caused by the retrieval is relatively low, and both document retrieval and merging operations can be processed in parallel to reduce time cost in practice, this limitation can also be mitigated. Meanwhile, some accuracy-priority tasks also need to make trade-offs between overhead and performance, and from this perspective, SeKe holds a certain practical value.

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.

Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *CoRR*, abs/2412.08905.

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, and 9 others. 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, and 12 others. 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*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. 2025. One shot dominance: Knowledge poisoning attack on retrieval-augmented generation systems. *CoRR*, abs/2505.11548.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *CoRR*, abs/2404.10981.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023a. RETA-LLM: A retrieval-augmented large language model toolkit. *CoRR*, abs/2306.05212.

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

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5).

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1389–1403. Association for Computational Linguistics.

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.

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

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. 2025. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30553–30571. Association for Computational Linguistics.

Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. 2024. Software testing with large language models: Survey, landscape, and vision. *IEEE Trans. Software Eng.*, 50(4):911–936.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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.

Annepaka Yadagiri and Partha Pakray. 2025. Large language models: a survey of their development, capabilities, and applications. *Knowl. Inf. Syst.*, 67(3):2967–3022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

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.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Yongji Wang, and Jian-Guang Lou. 2023. Large language models meet nl2code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7443–7464. Association for Computational Linguistics.

Yucheng Zhang, Qinfeng Li, Tianyu Du, Xuhong Zhang, Xinkui Zhao, Zhengwen Feng, and Jianwei Yin. 2024. Hijackrag: Hijacking attacks against retrieval-augmented large language models. *CoRR*, abs/2410.22832.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *CoRR*, abs/2402.19473.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. 2025. Trustrag: Enhancing robustness and trustworthiness in RAG. *CoRR*, abs/2501.00879.

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 Prompts

Figure 9 and 10 are the prompts used in this study.

B Performance of Approaches

Table 1 to 3 contain the detailed experimental values for RQ1 to 3. The **bold** and underline represent the highest and second-highest values (excluding methods enhanced by SeKe, such as Robust+), respectively.

The green , red and yellow represent performance become better, worse or no change after enhanced by SeKe, respectively.

The **Mtd** represents method (For better presentation, we have omitted the word *RAG* in the method name), and the **Mtc** represents metric.

The **PMode** (Pollution mode) represents injecting different number of pollution documents into the knowledge base for each question: **M-shot** indicates injecting multiple polluted documents (injecting all 5 pollution documents), and **O-shot** represents injecting only one polluted document.

The **datasets** with suffix p, h, and a represent the polluted documents of them are generated by *PoisonedRAG*, *HijackRAG*, and *Authchain*, respectively. Due to the *Authchain* is a one-shot method and not applicable to M-shot mode, we only present the O-shot results.

Due to the relatively small overall differences, in order to present the experimental results more concisely, only the average of the results from three retrievers are shown here and more details can be found in provided materials.

C Performance of Various Contexts on Different RAG frameworks

In this section, we will present a comparison of various contexts on different RAG frameworks. The experimental data used in Figure 11 to 20 include results from all retrievers and data sources (NQ, HP, MS).

C.1 ACC of Various Contexts

Figure 11 and 15 show the *ACC* of various contexts enhancing different RAG frameworks. It can be seen from the figures that the complete SeKe has a better enhancement effect on *RobustRAG*, *InstructRAG* and *AstuteRAG* under M-shot mode. For

Pooling Prompt

Task: Based on the given question, integrate the information from the provided externally retrieved documents.
* Note that there may be inconsistencies among these documents. You should leverage your own knowledge to appropriately reconcile such discrepancies and ensure the information is as complete and accurate as possible.
* Summarize the key details into a concise document, excluding all information that is irrelevant to the question.
* Please limit the corpus to 30 words and only return corpus without any other information.
Question: question
Documents: documents
Output:

Figure 9: The prompt used for merging the documents

Memory Prompt

Please generate a concise corpus based on your own knowledge, providing accurate and relevant information to answer the given question [question]. (Please limit the corpus to 30 words and only return corpus without any other information)

Figure 10: The prompt used for exporting memory

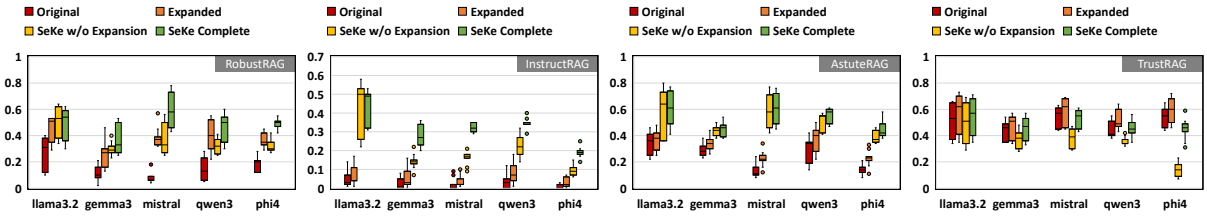

Figure 11: ACC of various contexts on different RAG frameworks ($Dataset_p$:M-shot)

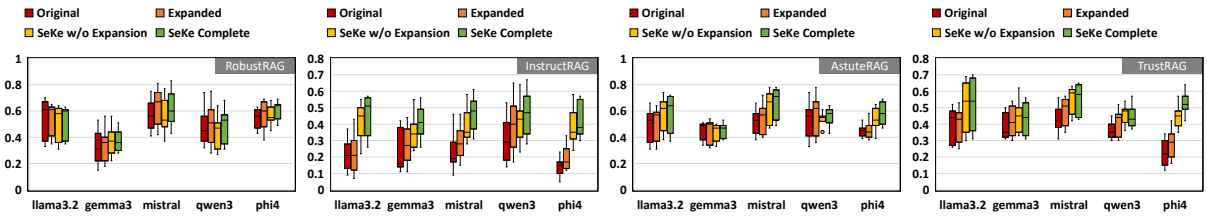

Figure 12: ACC of various contexts on different RAG frameworks ($Dataset_p$:O-shot)

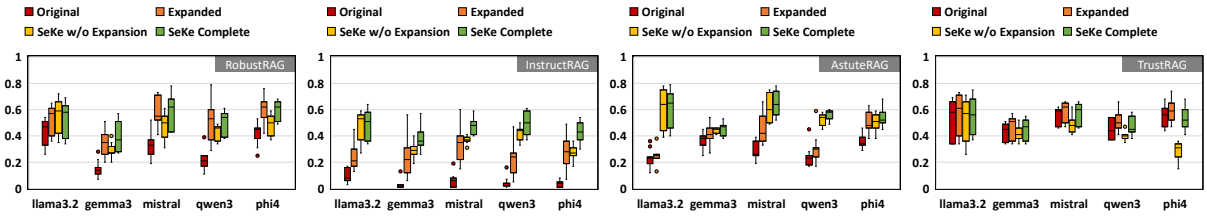

Figure 13: ACC of various contexts on different RAG frameworks ($Dataset_h$:M-shot)

Table 1: Performance of SeKe on different RAG frameworks (RQ1)

Dataset	NQ_p		HP_p		MS_p		NQ_h		HP_h		MS_h		NQ_a	HP_a	MS_a
PMode	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	–	–	–
Mtd \ **Mtc**	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR
Llama3.2															
Robust	0.36 0.38	0.61 **0.04**	0.11 0.45	**0.35 0.14**	0.34 0.28	**0.69 0.07**	0.48 0.16	**0.63 0.03**	0.29 0.26	**0.39 0.11**	0.52 0.21	**0.73 0.06**	0.57 **0.05**	0.33 **0.16**	**0.66 0.07**
Instruct	0.05 0.71	0.21 0.41	0.02 0.74	0.11 0.56	0.08 0.49	0.31 0.29	0.10 0.69	0.38 0.39	0.05 0.77	0.18 0.53	0.15 0.65	0.42 0.34	0.25 0.34	0.10 0.56	0.33 0.26
Astute	0.44 0.40	0.61 0.20	0.23 0.56	0.34 0.35	0.33 0.39	0.52 0.18	0.25 0.70	0.54 0.33	0.17 0.76	0.32 0.49	0.26 0.62	0.52 0.29	**0.58** 0.26	0.33 0.31	0.56 0.21
Trust	**0.66 0.00**	0.46 0.11	**0.36 0.05**	0.27 0.21	0.54 **0.06**	0.47 0.12	**0.67 0.00**	0.58 0.06	0.34 **0.04**	0.35 0.13	0.57 **0.06**	0.56 0.07	0.42 0.10	0.18 0.23	0.42 0.11
SeKe	0.62 0.05	0.55 0.08	0.33 0.16	0.32 0.17	**0.58** 0.13	0.62 0.13	0.66 0.01	0.59 0.07	**0.36** 0.12	0.37 0.13	**0.65** 0.12	0.63 0.14	0.55 0.10	0.26 0.25	0.58 0.15
Robust+	0.56 0.04	0.60 0.03	0.33 0.12	0.36 0.10	0.60 0.12	0.62 0.06	0.59 0.02	0.60 0.02	0.36 0.09	0.40 0.05	0.64 0.11	0.63 0.07	0.57 0.02	0.37 0.09	0.60 0.08
Instruct+	0.50 0.06	0.54 0.10	0.31 0.15	0.30 0.19	0.50 0.10	0.53 0.10	0.60 0.02	0.58 0.07	0.35 0.11	0.33 0.11	0.53 0.07	0.55 0.07	0.54 0.09	0.23 0.22	0.51 0.11
Astute+	0.76 0.02	0.72 0.05	0.46 0.10	0.40 0.17	0.61 0.07	0.64 0.08	0.74 0.02	0.70 0.05	0.44 0.09	0.43 0.12	0.65 0.05	0.64 0.06	0.69 0.08	0.39 0.21	0.62 0.10
Trust+	0.69 0.00	0.69 0.01	0.38 0.07	0.33 0.08	0.58 0.05	0.55 0.04	0.73 0.01	0.71 0.01	0.40 0.05	0.35 0.07	0.56 0.07	0.57 0.05	0.67 0.03	0.27 0.14	0.58 0.06
Gemma3															
Robust	0.09 0.42	0.34 0.11	0.08 0.50	0.19 0.16	0.17 0.50	0.48 0.13	0.14 0.33	0.35 **0.05**	0.12 0.29	0.25 **0.08**	0.19 0.19	0.50 **0.06**	0.31 0.12	0.24 **0.16**	0.47 0.12
Instruct	0.03 0.74	0.25 0.42	0.00 0.78	0.14 0.58	0.05 0.62	0.33 0.27	0.06 0.86	0.36 0.37	0.01 0.94	0.25 0.49	0.06 0.86	0.49 0.29	0.30 0.35	0.08 0.57	0.33 0.32
Astute	0.27 0.31	**0.49 0.05**	0.25 0.34	0.36 **0.15**	0.34 0.25	0.50 **0.08**	0.38 0.24	**0.48** 0.10	0.33 0.34	0.37 0.17	0.39 0.27	0.55 0.10	**0.48 0.11**	0.35 0.18	**0.51 0.10**
Trust	**0.46 0.05**	0.40 0.14	**0.35 0.10**	0.31 0.22	0.51 **0.10**	0.47 0.14	**0.45 0.06**	0.44 0.16	0.34 **0.10**	0.35 0.14	0.50 **0.08**	0.54 0.13	0.40 0.20	0.26 0.26	0.46 0.16
SeKe	0.42 0.18	0.46 0.12	0.33 0.34	**0.39** 0.29	**0.54** 0.20	**0.57** 0.11	0.42 0.12	0.43 0.12	**0.37** 0.23	**0.43** 0.21	**0.55** 0.14	**0.58** 0.09	0.37 0.19	0.35 0.40	0.49 0.22
Robust+	0.32 0.14	0.35 0.08	0.27 0.13	0.29 0.11	0.51 0.16	0.48 0.11	0.36 0.12	0.37 0.04	0.28 0.10	0.30 0.08	0.54 0.13	0.53 0.09	0.32 0.12	0.29 0.12	0.46 0.15
Instruct+	0.29 0.24	0.46 0.14	0.24 0.29	0.31 0.23	0.31 0.14	0.49 0.09	0.37 0.17	0.49 0.12	0.31 0.16	0.34 0.14	0.47 0.10	0.54 0.07	0.42 0.16	0.27 0.26	0.36 0.15
Astute+	0.45 0.06	0.49 0.05	0.38 0.08	0.38 0.09	0.50 0.05	0.49 0.06	0.46 0.06	0.48 0.05	0.39 0.09	0.41 0.10	0.51 0.05	0.52 0.03	0.45 0.07	0.38 0.10	0.49 0.07
Trust+	0.47 0.07	0.44 0.08	0.34 0.11	0.32 0.12	0.55 0.09	0.54 0.10	0.48 0.06	0.42 0.09	0.35 0.11	0.34 0.11	0.53 0.08	0.56 0.10	0.43 0.10	0.31 0.18	0.53 0.11
Mistral															
Robust	0.06 0.76	0.56 0.08	0.06 0.83	0.44 **0.10**	0.15 0.68	0.71 **0.05**	0.33 0.41	0.62 **0.02**	0.23 0.44	0.44 0.07	0.43 0.39	0.74 0.06	**0.59 0.05**	0.44 **0.12**	**0.71 0.06**
Instruct	0.01 0.87	0.22 0.38	0.00 0.80	0.16 0.52	0.05 0.60	0.36 0.20	0.05 0.83	0.44 0.26	0.03 0.81	0.34 0.28	0.10 0.79	0.61 0.15	0.26 0.35	0.08 0.55	0.33 0.19
Astute	0.12 0.34	0.52 0.14	0.10 0.41	0.40 0.19	0.19 0.24	0.62 0.10	0.25 0.40	0.63 0.10	0.25 0.46	0.52 0.15	0.38 0.28	0.77 0.07	0.42 0.17	0.34 0.26	0.54 0.11
Trust	**0.59 0.04**	0.44 0.12	0.44 **0.02**	0.35 0.13	0.60 **0.03**	0.49 0.06	0.59 0.05	0.61 0.08	0.47 **0.01**	0.46 **0.04**	0.61 **0.02**	0.57 **0.02**	0.40 0.14	0.25 0.16	0.36 0.08
SeKe	0.58 0.08	**0.63 0.07**	**0.45** 0.23	**0.50** 0.17	**0.70** 0.08	**0.80** 0.06	**0.64 0.03**	**0.65** 0.04	**0.54** 0.11	**0.56** 0.12	**0.74** 0.05	**0.81** 0.03	0.54 0.09	**0.47** 0.20	0.68 0.08
Robust+	0.56 0.09	0.59 0.05	0.44 0.07	0.47 0.07	0.76 0.06	0.78 0.05	0.63 0.04	0.63 0.04	0.43 0.09	0.47 0.07	0.73 0.06	0.77 0.05	0.59 0.06	0.43 0.10	0.71 0.06
Instruct+	0.31 0.14	0.50 0.09	0.32 0.20	0.33 0.16	0.33 0.06	0.55 0.05	0.50 0.06	0.58 0.07	0.41 0.11	0.44 0.12	0.52 0.04	0.64 0.02	0.42 0.19	0.33 0.24	0.41 0.06
Astute+	0.61 0.06	0.71 0.04	0.48 0.12	0.51 0.11	0.75 0.04	0.77 0.05	0.62 0.03	0.69 0.03	0.53 0.09	0.51 0.10	0.76 0.04	0.78 0.05	0.62 0.07	0.45 0.15	0.73 0.06
Trust+	0.59 0.04	0.65 0.03	0.44 0.02	0.43 0.02	0.56 0.03	0.58 0.03	0.65 0.04	0.65 0.02	0.46 0.02	0.47 0.02	0.59 0.03	0.60 0.03	0.58 0.06	0.38 0.08	0.56 0.03
Qwen3															
Robust	0.17 0.64	0.50 0.14	0.05 0.62	0.35 0.19	0.20 0.63	0.63 0.11	0.22 0.29	0.53 **0.05**	0.14 0.26	0.34 **0.04**	0.29 0.34	**0.69 0.04**	0.49 0.14	0.32 **0.20**	**0.65 0.07**
Instruct	0.03 0.67	0.34 0.37	0.00 0.80	0.15 0.58	0.09 0.56	0.45 0.20	0.02 0.86	0.35 0.47	0.01 0.93	0.21 0.51	0.09 0.81	0.49 0.28	0.30 0.25	0.08 0.61	0.32 0.22
Astute	0.37 0.34	**0.57** 0.12	0.17 0.68	0.37 0.43	0.33 0.41	**0.70 0.07**	0.26 0.60	0.52 0.23	0.18 0.73	0.35 0.49	0.30 0.56	0.63 0.20	0.54 0.19	0.34 0.43	0.55 0.19
Trust	0.54 0.06	0.37 0.14	0.39 **0.06**	0.32 0.21	0.41 **0.08**	0.40 0.13	**0.54 0.06**	0.45 0.15	0.37 **0.07**	0.39 0.16	0.44 0.08	0.48 0.15	0.34 0.16	0.30 0.23	0.35 0.15
SeKe	0.54 **0.04**	0.54 **0.06**	**0.46** 0.19	**0.49** 0.19	**0.51** 0.13	0.55 0.10	0.53 0.08	**0.57** 0.08	**0.47** 0.15	**0.48** 0.15	**0.54** 0.08	0.56 0.11	0.54 **0.06**	**0.45** 0.22	0.55 0.12
Robust+	0.54 0.10	0.51 0.08	0.32 0.15	0.33 0.14	0.53 0.13	0.62 0.09	0.55 0.08	0.54 0.04	0.38 0.12	0.36 0.10	0.58 0.12	0.63 0.06	0.51 0.07	0.34 0.14	0.58 0.09
Instruct+	0.36 0.11	0.49 0.13	0.31 0.26	0.30 0.24	0.36 0.12	0.61 0.07	0.53 0.09	0.55 0.08	0.39 0.15	0.37 0.16	0.57 0.06	0.65 0.07	0.41 0.11	0.26 0.29	0.50 0.12
Astute+	0.60 0.06	0.58 0.05	0.48 0.13	0.47 0.13	0.59 0.06	0.62 0.02	0.58 0.04	0.62 0.05	0.51 0.10	0.50 0.11	0.59 0.03	0.59 0.03	0.57 0.07	0.47 0.18	0.58 0.06
Trust+	0.53 0.06	0.53 0.07	0.40 0.09	0.38 0.08	0.44 0.06	0.44 0.07	0.56 0.07	0.54 0.06	0.40 0.07	0.40 0.07	0.45 0.08	0.48 0.08	0.47 0.08	0.36 0.16	0.41 0.09
Phi4															
Robust	0.20 0.41	0.57 0.09	0.12 0.57	0.44 **0.13**	0.25 0.36	**0.62** 0.07	0.44 0.25	0.66 0.04	0.31 0.32	0.51 0.08	0.49 0.22	**0.73** 0.04	**0.61 0.05**	**0.44 0.14**	**0.52** 0.06
Instruct	0.01 0.56	0.14 0.26	0.00 0.67	0.12 0.42	0.02 0.37	0.16 0.12	0.03 0.58	0.45 0.14	0.04 0.70	0.37 0.28	0.04 0.66	0.51 0.13	0.23 0.15	0.08 0.46	0.23 0.10
Astute	0.17 0.21	0.49 0.09	0.11 0.31	0.41 0.15	0.14 0.14	0.45 **0.04**	0.37 0.13	0.60 **0.02**	0.32 0.15	0.53 0.08	0.36 0.09	0.66 0.03	0.41 0.09	0.35 0.17	0.35 **0.03**
Trust	**0.62 0.02**	0.23 0.10	**0.45 0.05**	0.14 0.20	**0.54 0.04**	0.30 0.08	0.65 **0.01**	0.55 0.09	0.46 **0.04**	0.28 0.15	0.55 0.04	0.36 0.09	0.32 0.10	0.13 0.20	0.33 0.04
SeKe	0.51 0.05	**0.60 0.06**	0.42 0.15	**0.48** 0.14	0.43 0.07	0.59 0.05	**0.68** 0.05	**0.71** 0.03	**0.51** 0.07	**0.57 0.06**	**0.65** 0.04	0.72 0.03	0.52 0.09	0.36 0.22	0.37 0.06
Robust+	0.52 0.07	0.65 0.03	0.45 0.09	0.52 0.12	0.50 0.05	0.64 0.04	0.64 0.04	0.67 0.03	0.50 0.07	0.56 0.05	0.65 0.02	0.72 0.03	0.61 0.04	0.46 0.09	0.53 0.04
Instruct+	0.22 0.08	0.50 0.09	0.18 0.22	0.32 0.16	0.17 0.06	0.46 0.04	0.49 0.06	0.64 0.06	0.37 0.11	0.46 0.07	0.44 0.03	0.61 0.03	0.43 0.09	0.24 0.20	0.30 0.06
Astute+	0.54 0.04	0.65 0.02	0.40 0.11	0.49 0.08	0.41 0.03	0.60 0.01	0.61 0.03	0.68 0.02	0.48 0.08	0.54 0.05	0.54 0.02	0.67 0.01	0.57 0.03	0.43 0.09	0.41 0.02
Trust+	0.51 0.04	0.61 0.04	0.36 0.07	0.46 0.06	0.48 0.03	0.53 0.03	0.63 0.03	0.65 0.03	0.44 0.08	0.46 0.04	0.53 0.04	0.58 0.02	0.57 0.06	0.32 0.11	0.44 0.04
Average															
Robust	0.18 0.52	0.52 0.09	0.08 0.59	0.35 **0.14**	0.22 0.49	0.63 **0.09**	0.32 0.29	0.56 **0.04**	0.22 0.31	0.39 **0.08**	0.38 0.27	**0.68 0.05**	**0.51 0.08**	0.35 **0.16**	**0.60 0.08**
Instruct	0.03 0.71	0.23 0.37	0.00 0.76	0.14 0.53	0.06 0.53	0.32 0.22	0.05 0.76	0.40 0.33	0.03 0.83	0.27 0.42	0.09 0.75	0.50 0.24	0.27 0.29	0.08 0.55	0.31 0.22
Astute	0.27 0.32	0.54 0.12	0.17 0.46	0.38 0.25	0.27 0.29	0.56 0.09	0.30 0.41	0.55 0.16	0.25 0.49	0.42 0.28	0.34 0.36	0.63 0.14	0.49 0.16	0.34 0.27	0.50 0.13
Trust	**0.57 0.03**	0.38 0.12	0.40 **0.06**	0.28 0.19	0.52 **0.06**	0.43 0.11	0.58 **0.04**	0.53 0.11	0.40 **0.05**	0.37 0.12	0.53 **0.06**	0.50 0.09	0.38 0.14	0.22 0.22	0.38 0.11
SeKe	0.53 0.08	**0.56 0.08**	0.40 0.21	**0.44** 0.19	**0.55** 0.12	0.63 0.09	**0.59** 0.06	**0.59** 0.07	**0.45** 0.14	**0.48** 0.13	**0.63** 0.09	0.66 0.08	0.50 0.11	**0.38** 0.26	0.53 0.13
Robust+	0.50 0.09	0.54 0.05	0.36 0.11	0.39 0.11	0.58 0.10	0.63 0.07	0.55 0.06	0.56 0.03	0.39 0.09	0.42 0.07	0.63 0.09	0.66 0.06	0.52 0.06	0.38 0.11	0.58 0.08
Instruct+	0.34 0.13	0.50 0.11	0.27 0.22	0.31 0.20	0.33 0.10	0.53 0.07	0.50 0.08	0.57 0.08	0.37 0.13	0.39 0.12	0.51 0.06	0.60 0.05	0.44 0.13	0.27 0.24	0.42 0.10
Astute+	0.59 0.05	0.63 0.04	0.44 0.11	0.45 0.12	0.57 0.05	0.62 0.04	0.60 0.04	0.63 0.04	0.47 0.09	0.48 0.10	0.61 0.04	0.64 0.04	0.58 0.06	0.42 0.15	0.57 0.06
Trust+	0.56 0.04	0.58 0.05	0.38 0.07	0.38 0.07	0.52 0.05	0.53 0.05	0.61 0.04	0.59 0.04	0.41 0.07	0.40 0.06	0.53 0.06	0.56 0.06	0.54 0.07	0.33 0.13	0.50 0.07

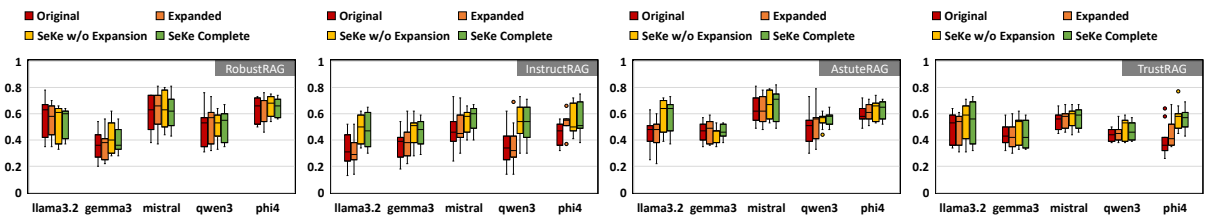

Figure 14: ACC of various contexts on different RAG frameworks ($Dataset_h$:O-shot)

Table 2: Performance of different ehanced contexts (RQ2)

Dataset	NQ_p		HP_p		MS_p		NQ_h		HP_h		MS_h		NQ_a	HP_a	MS_a
Pmode	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	–	–	–
Mtc / Mtd	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR
Llama3.2															
Original	0.04 0.83	0.26 0.57	0.03 0.79	0.17 0.59	0.11 0.70	0.28 0.49	0.07 0.73	0.29 0.49	0.03 0.69	0.22 0.45	0.08 0.72	0.41 0.49	0.29 0.40	0.09 0.64	0.35 0.41
Expanded	0.12 0.70	0.25 0.56	0.07 0.59	0.18 0.59	0.13 0.56	0.29 0.49	0.10 0.80	0.26 0.54	0.05 0.79	0.20 0.48	0.24 0.69	0.37 0.52	0.29 0.42	0.10 0.62	0.24 0.41
SeKe	**0.62 0.05**	**0.55 0.08**	**0.33 0.16**	**0.32 0.17**	**0.58 0.13**	**0.62 0.13**	**0.66 0.01**	**0.59 0.07**	**0.36 0.12**	**0.37 0.13**	**0.65 0.12**	**0.63 0.14**	**0.55 0.10**	**0.26 0.25**	**0.58 0.15**
Gemma3															
Original	0.02 0.93	0.25 0.54	0.01 0.97	0.15 0.75	0.07 0.85	0.41 0.43	0.02 0.94	0.40 0.33	0.00 0.97	0.31 0.41	0.10 0.87	0.58 0.26	0.24 0.48	0.09 0.74	0.33 0.49
Expanded	0.05 0.82	0.32 0.46	0.03 0.92	0.14 0.76	0.13 0.76	0.44 0.33	0.23 0.66	**0.47** 0.29	0.12 0.82	0.30 0.41	0.44 0.47	**0.61** 0.21	0.26 0.47	0.10 0.73	0.31 0.52
SeKe	**0.42 0.18**	**0.46 0.12**	**0.33 0.34**	**0.39 0.29**	**0.54 0.20**	**0.57 0.11**	**0.42 0.12**	0.43 **0.12**	**0.37 0.23**	**0.43 0.21**	**0.55 0.14**	0.58 **0.09**	**0.37 0.19**	**0.35 0.40**	**0.49 0.22**
Mistral															
Original	0.01 0.89	0.21 0.47	0.00 0.92	0.22 0.62	0.07 0.77	0.44 0.28	0.02 0.94	0.42 0.32	0.03 0.97	0.37 0.39	0.08 0.88	0.67 0.21	0.26 0.44	0.12 0.65	0.35 0.42
Expanded	0.03 0.75	0.27 0.39	0.03 0.83	0.21 0.60	0.11 0.59	0.48 0.26	0.18 0.66	0.46 0.30	0.19 0.74	0.42 0.35	0.43 0.46	0.69 0.16	0.27 0.45	0.13 0.63	0.34 0.39
SeKe	**0.58 0.08**	**0.63 0.07**	**0.45 0.23**	**0.50 0.17**	**0.70 0.08**	**0.80 0.06**	**0.64 0.03**	**0.65 0.04**	**0.54 0.11**	**0.56 0.12**	**0.74 0.05**	**0.81 0.03**	**0.54 0.09**	**0.47 0.20**	**0.68 0.08**
Qwen3															
Original	0.06 0.86	0.40 0.48	0.00 0.94	0.18 0.72	0.10 0.73	0.54 0.25	0.02 0.96	0.33 0.55	0.00 0.98	0.16 0.73	0.05 0.85	0.49 0.39	0.30 0.41	0.10 0.70	0.37 0.38
Expanded	0.13 0.72	0.47 0.42	0.04 0.86	0.19 0.70	0.15 0.63	0.53 0.22	0.11 0.85	0.39 0.51	0.07 0.89	0.17 0.72	0.24 0.69	0.52 0.36	0.28 0.46	0.09 0.72	0.38 0.37
SeKe	**0.54 0.04**	**0.54 0.06**	**0.46 0.19**	**0.49 0.19**	**0.51 0.13**	**0.55 0.10**	**0.53 0.08**	**0.57 0.08**	**0.47 0.15**	**0.48 0.15**	**0.54 0.08**	**0.56 0.11**	**0.54 0.06**	**0.45 0.22**	**0.55 0.12**
Phi4															
Original	0.01 0.72	0.21 0.31	0.00 0.80	0.17 0.53	0.02 0.59	0.27 0.17	0.11 0.68	0.52 0.19	0.06 0.81	0.31 0.35	0.17 0.65	0.57 0.13	0.24 0.33	0.08 0.56	0.27 0.22
Expanded	0.03 0.47	0.23 0.25	0.02 0.65	0.19 0.49	0.05 0.30	0.30 0.15	0.27 0.40	0.58 0.18	0.20 0.58	0.37 0.31	0.44 0.26	0.60 0.12	0.20 0.28	0.08 0.57	0.17 0.17
SeKe	**0.51 0.05**	**0.60 0.06**	**0.42 0.15**	**0.48 0.14**	**0.43 0.07**	**0.59 0.05**	**0.68 0.05**	**0.71 0.03**	**0.51 0.07**	**0.57 0.06**	**0.65 0.04**	**0.72 0.03**	**0.52 0.09**	**0.36 0.22**	**0.37 0.06**
Average															
Original	0.03 0.85	0.27 0.47	0.01 0.88	0.18 0.64	0.07 0.73	0.39 0.32	0.05 0.85	0.39 0.38	0.02 0.88	0.27 0.47	0.10 0.79	0.54 0.30	0.27 0.41	0.10 0.66	0.33 0.38
Expanded	0.07 0.69	0.31 0.42	0.04 0.77	0.18 0.63	0.11 0.57	0.41 0.29	0.18 0.67	0.43 0.36	0.13 0.76	0.29 0.45	0.36 0.51	0.56 0.27	0.26 0.42	0.10 0.65	0.29 0.37
SeKe	**0.53 0.08**	**0.56 0.08**	**0.40 0.21**	**0.44 0.19**	**0.55 0.12**	**0.63 0.09**	**0.59 0.06**	**0.59 0.07**	**0.45 0.14**	**0.48 0.13**	**0.63 0.09**	**0.66 0.08**	**0.50 0.11**	**0.38 0.26**	**0.53 0.13**

Table 3: Performance of SeKe with and without Knowledge Expansion (RQ3)

Dataset	NQ_p		HP_p		MS_p		NQ_h		HP_h		MS_h		NQ_a	HP_a	MS_a
PMode	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	M-shot	O-shot	–	–	–
Mtc / Mtd*	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR	ACC MR
Llama3.2															
SeKe w/o E	0.57 0.06	0.51 0.13	0.28 0.18	**0.32 0.17**	0.50 0.22	0.57 0.17	0.59 0.03	**0.59** 0.09	0.32 **0.11**	0.34 **0.13**	0.60 0.16	0.60 0.17	**0.56** 0.11	**0.27 0.25**	0.54 0.20
Complete	**0.62 0.05**	**0.55 0.08**	**0.33 0.16**	**0.32 0.17**	**0.58 0.13**	**0.62 0.13**	**0.66 0.01**	**0.59 0.07**	**0.36** 0.12	**0.37 0.13**	**0.65 0.12**	**0.63 0.14**	0.55 **0.10**	0.26 **0.25**	**0.58 0.15**
Gemma3															
SeKe w/o E	0.33 0.27	0.42 0.19	0.27 0.50	0.35 0.37	0.44 0.30	0.56 0.14	0.34 0.21	**0.45** 0.13	0.34 0.30	0.39 0.24	0.48 0.22	**0.58** 0.11	**0.37** 0.22	0.29 0.42	**0.52 0.20**
Complete	**0.42 0.18**	**0.46 0.12**	**0.33 0.34**	**0.39 0.29**	**0.54 0.20**	**0.57 0.11**	**0.42 0.12**	0.43 **0.12**	**0.37 0.23**	**0.43 0.21**	**0.55 0.14**	**0.58 0.09**	**0.37 0.19**	**0.35 0.40**	0.49 0.22
Mistral															
SeKe w/o E	0.36 0.15	0.52 0.12	0.37 0.27	0.45 0.22	0.45 0.13	0.76 **0.06**	0.49 0.06	0.62 0.08	0.47 0.16	0.54 0.16	0.60 0.07	**0.81 0.03**	0.47 0.15	0.44 0.25	0.62 0.09
Complete	**0.58 0.08**	**0.63 0.07**	**0.45 0.23**	**0.50 0.17**	**0.70 0.08**	**0.80 0.06**	**0.64 0.03**	**0.65 0.04**	**0.54 0.11**	**0.56 0.12**	**0.74 0.05**	**0.81 0.03**	**0.54 0.09**	**0.47 0.20**	**0.68 0.08**
Qwen3															
SeKe w/o E	0.49 0.07	0.51 0.08	0.44 **0.19**	0.47 **0.18**	0.47 0.14	0.47 0.12	0.50 **0.03**	**0.58** 0.09	**0.49 0.14**	**0.49** 0.16	0.50 0.11	0.53 **0.10**	0.50 0.08	**0.45** 0.24	0.48 0.14
Complete	**0.54 0.04**	**0.54 0.06**	**0.46 0.19**	**0.49** 0.19	**0.51 0.13**	**0.55 0.10**	**0.53** 0.08	0.57 **0.08**	0.47 0.15	0.48 **0.15**	**0.54 0.08**	**0.56** 0.11	**0.54 0.06**	**0.45 0.22**	**0.55 0.12**
Phi4															
SeKe w/o E	0.41 0.15	0.52 0.08	0.29 0.25	0.45 0.17	0.36 0.13	**0.59 0.05**	0.57 0.13	**0.71** 0.05	0.43 0.16	**0.58** 0.08	0.62 0.08	0.70 0.04	0.51 0.14	0.31 0.25	**0.40 0.06**
Complete	**0.51 0.05**	**0.60 0.06**	**0.42 0.15**	**0.48 0.14**	**0.43 0.07**	**0.59 0.05**	**0.68 0.05**	**0.71 0.03**	**0.51 0.07**	0.57 **0.06**	**0.65 0.04**	**0.72 0.03**	**0.52 0.09**	**0.36 0.22**	0.37 **0.06**
Average															
SeKe w/o E	0.43 0.14	0.50 0.12	0.33 0.28	0.41 0.22	0.44 0.18	0.59 0.11	0.50 0.09	**0.59** 0.09	0.41 0.17	0.47 0.15	0.56 0.13	0.64 0.09	0.48 0.14	0.35 0.28	0.51 0.14
Complete	**0.53 0.08**	**0.56 0.08**	**0.40 0.21**	**0.44 0.19**	**0.55 0.12**	**0.63 0.09**	**0.59 0.06**	0.59 **0.07**	**0.45 0.14**	**0.48 0.13**	**0.63 0.09**	**0.66 0.08**	**0.50 0.11**	**0.38 0.26**	**0.53 0.13**

* **SeKe w/o E**: disable the Knowledge Expansion of SeKe, **Complete**: enable all parts of SeKe.

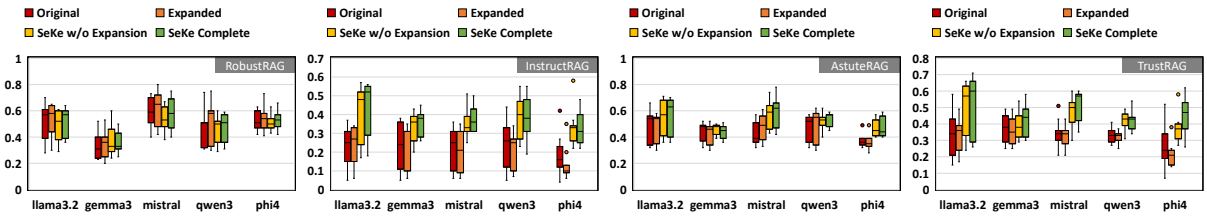

Figure 15: ACC of various contexts on different RAG frameworks ($Dataset_a$)

TrustRAG, its performance is comparable to that of the complete SeKe under M-shot mode. The specific reasons have been explained in the previous and can be found in Section 5.1. Under O-shot mode, the complete SeKe has a better enhancement effect on *InstructRAG*, *AstuteRAG* and *TrustRAG*. For *RobustRAG*, its performance is comparable to that of the complete SeKe.

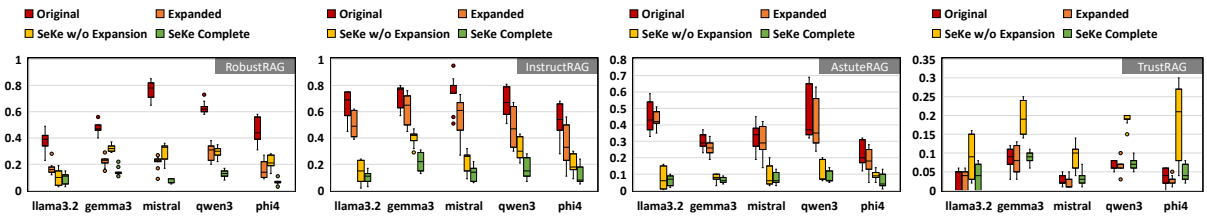

Figure 16: MR of various contexts on different RAG frameworks ($Dataset_p$:M-shot)

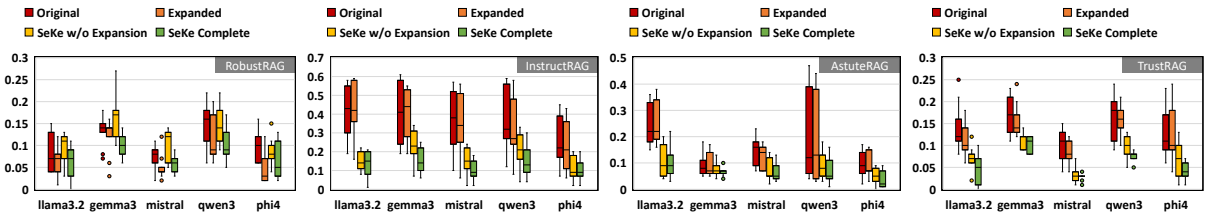

Figure 17: MR of various contexts on different RAG frameworks ($Dataset_p$:O-shot)

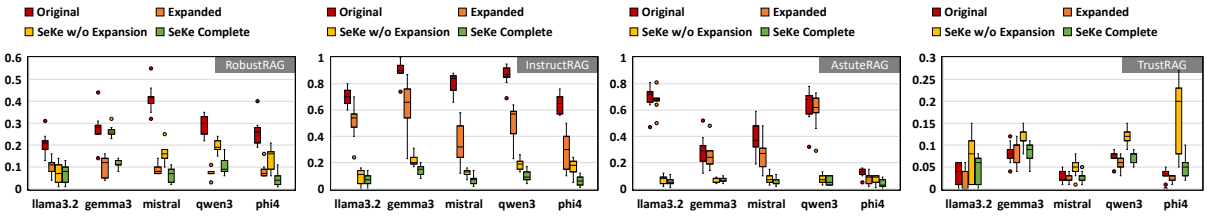

Figure 18: MR of various contexts on different RAG frameworks ($Dataset_h$:M-shot)

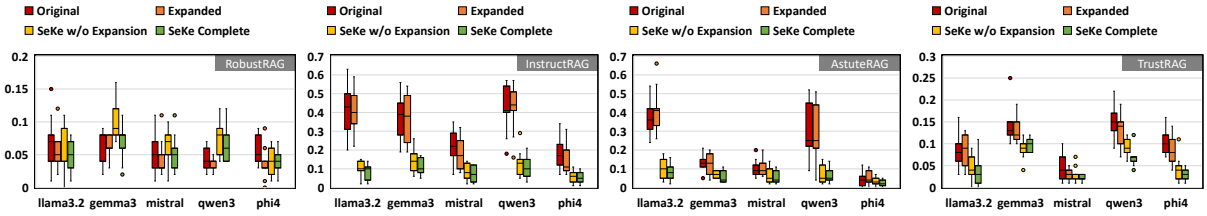

Figure 19: MR of various contexts on different RAG frameworks ($Dataset_h$:O-shot)

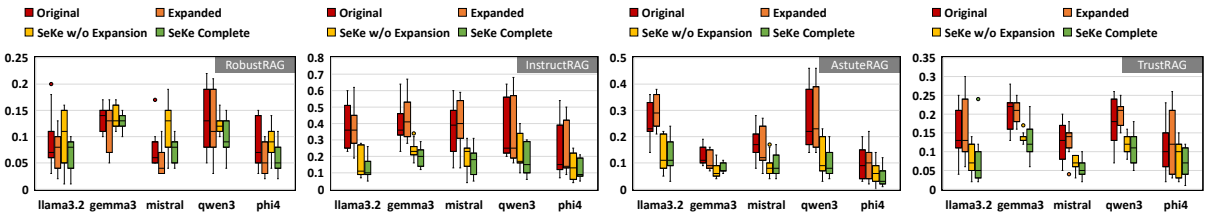

Figure 20: MR of various contexts on different RAG frameworks ($Dataset_a$)

C.2 MR of Various Contexts

Figure 16 and 20 show the *MR* of various contexts enhancing different RAG frameworks. The performance trend of *MR* is roughly the same as the *ACC*.

Under M-shot mode, the complete SeKe has a better enhancement effect on *RobustRAG*, *InstructRAG* and *AstuteRAG*, while under O-shot mode, the complete SeKe has a better enhancement effect on *InstructRAG*, *AstuteRAG* and *TrustRAG*.

D Experiments Checklist

Table 4 shows the experiments conduct in this study.

Table 4: Experiments checklist

Dataset	Pollution*	#Sample	Ret \ LLM	llama3.2	gemma3	mistral	qwen3	phi4
Natural Question	PoisonedRAG$_o$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
Natural Question	PoisonedRAG$_o$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
Natural Question	PoisonedRAG$_o$	100	Contriever	✓	✓	✓	✓	✓
Natural Question	PoisonedRAG$_m$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
Natural Question	PoisonedRAG$_m$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
Natural Question	PoisonedRAG$_m$	100	Contriever	✓	✓	✓	✓	✓
Natural Question	HijackRAG$_o$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
Natural Question	HijackRAG$_o$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
Natural Question	HijackRAG$_o$	100	Contriever	✓	✓	✓	✓	✓
Natural Question	HijackRAG$_m$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
Natural Question	HijackRAG$_m$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
Natural Question	HijackRAG$_m$	100	Contriever	✓	✓	✓	✓	✓
Natural Question	AuthChain	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
Natural Question	AuthChain	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
Natural Question	AuthChain	100	Contriever	✓	✓	✓	✓	✓
HotpotQA	PoisonedRAG$_o$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
HotpotQA	PoisonedRAG$_o$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
HotpotQA	PoisonedRAG$_o$	100	Contriever	✓	✓	✓	✓	✓
HotpotQA	PoisonedRAG$_m$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
HotpotQA	PoisonedRAG$_m$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
HotpotQA	PoisonedRAG$_m$	100	Contriever	✓	✓	✓	✓	✓
HotpotQA	HijackRAG$_o$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
HotpotQA	HijackRAG$_o$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
HotpotQA	HijackRAG$_o$	100	Contriever	✓	✓	✓	✓	✓
HotpotQA	HijackRAG$_m$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
HotpotQA	HijackRAG$_m$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
HotpotQA	HijackRAG$_m$	100	Contriever	✓	✓	✓	✓	✓
HotpotQA	AuthChain	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
HotpotQA	AuthChain	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
HotpotQA	AuthChain	100	Contriever	✓	✓	✓	✓	✓
MSMARCO	PoisonedRAG$_o$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
MSMARCO	PoisonedRAG$_o$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
MSMARCO	PoisonedRAG$_o$	100	Contriever	✓	✓	✓	✓	✓
MSMARCO	PoisonedRAG$_m$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
MSMARCO	PoisonedRAG$_m$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
MSMARCO	PoisonedRAG$_m$	100	Contriever	✓	✓	✓	✓	✓
MSMARCO	HijackRAG$_o$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
MSMARCO	HijackRAG$_o$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
MSMARCO	HijackRAG$_o$	100	Contriever	✓	✓	✓	✓	✓
MSMARCO	HijackRAG$_m$	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
MSMARCO	HijackRAG$_m$	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
MSMARCO	HijackRAG$_m$	100	Contriever	✓	✓	✓	✓	✓
MSMARCO	AuthChain	100	all-MiniLM-L6-v2	✓	✓	✓	✓	✓
MSMARCO	AuthChain	100	all-mpnet-base-v2	✓	✓	✓	✓	✓
MSMARCO	AuthChain	100	Contriever	✓	✓	✓	✓	✓

*Polluted document generation method: the suffixes means the number of polluted documents we used in our experiments (o means injecting one polluted document into knowledge base, and m means injecting multiple polluted documents). Since Authchain is a one-shot method, we just inject one polluted document into knowledge base when applying it.