electronics

Article

Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application

Wenjun Meng [1,2], Yuzhe Li [3,*], Lili Chen [4] and Zhaomin Dong [3,4,*]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100811, China; mengwenjun@bit.edu.cn
[2] Beijing Huadian E-commerce Technology Co., Ltd., Beijing 100164, China
[3] School of Materials Science and Engineering, Beihang University, Beijing 100191, China
[4] School of Public Health, Southeast University, Nanjing 210096, China; chenlili1002@126.com
* Correspondence: 20377027@buaa.edu.cn (Y.L.); dongzm@buaa.edu.cn (Z.D.)

Abstract: While large language models (LLMs) are vital for retrieving relevant information from extensive knowledge bases, they always face challenges, including high costs and issues of credibility. Here, we developed a question answering system focused on human health risk using Retrieval-Augmented Generation (RAG). We first proposed a framework to generate question–answer pairs, resulting in 300 high-quality pairs across six subfields. Subsequently, we created both a Naive RAG and an Advanced RAG-based Question-Answering (Q&A) system. Performance evaluation of the 300 question–answer pairs in individual research subfields demonstrated that the Advanced RAG outperformed traditional LLMs (including ChatGPT and ChatGLM) and Naive RAG. Finally, we integrated the developed module for a single subfield to launch a multi-knowledge base question answering system. Our study represents a novel application of RAG technology and LLMs to optimize knowledge retrieval methods in human health risk assessment.

Keywords: human health risk assessment; large language model; artificial intelligence; environmental science

Academic Editor: Jichai Jeong

Received: 28 November 2024
Revised: 8 January 2025
Accepted: 17 January 2025
Published: 20 January 2025

Citation: Meng, W.; Li, Y.; Chen, L.; Dong, Z. Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application. *Electronics* **2025**, *14*, 386. https://doi.org/10.3390/electronics14020386

1. Introduction

The benefits of synthetic chemicals in daily life are undeniable; however, their intentional and unintentional release into the environment has been a significant risk factor for human health [1,2]. To date, millions of chemical substances have been identified [3], and health risk assessment is a crucial foundation for the regulation of these substances [4,5]. Evaluating the risks of chemicals requires not only an understanding of their environmental and biological behaviors, but also knowledge of their toxicity and various other factors [6,7]. In recent years, the rapid growth of knowledge has led to an influx in new information, and this exponential increase in knowledge has placed a substantial burden on the knowledge updates required by managers and professionals. Consequently, the demand for effective knowledge retrieval has sharply increased.

In knowledge-intensive tasks, the process of knowledge retrieval plays a crucial role [8]. This involves accurately locating information relevant to specific questions within vast knowledge bases. With the continuous advancements in artificial intelligence (AI) technologies, particularly those based on large language models (LLMs), the application of knowledge retrieval in vertical domain question-answering (Q&A) tasks is becoming

increasingly widespread [9,10]. The essence of Q&A tasks is to extract information from extensive text resources and generate accurate and relevant responses.

Despite significant progress in the field of LLMs, their application still faces several challenges. First, the textual knowledge acquired by LLMs through a large number of fixed parameters not only incurs high training costs, but also struggles to update with the latest knowledge from the external world [11], leading to difficulties in adapting to new information over time. Additionally, LLMs face credibility issues, such as generating hallucinations and factual inaccuracies [12]. Particularly, hallucination refers to the phenomenon where LLMs generate factually incorrect or nonsensical outputs. These unreliable outputs pose significant risks when deploying LLMs in real-world applications. Existing research indicates that the content generated by LLMs is often unreliable and poses various risks in many cases [13].

To address the challenges mentioned above, researchers have proposed Retrieval-Augmented Generation (RAG), a new paradigm that enhances LLMs by integrating external knowledge sources [14]. To illustrate how the RAG technique can be applied in LLMs for developing a Q&A system related to human health risks, this manuscript is organized as follows: Section 2 introduces the related work. Based on the summary in Section 2, Section 3 presents the research gap, aims, and objectives. Section 4 details the materials and methods, while Section 5 discusses the results. Finally, Section 6 summarizes the conclusions.

2. Related Work

RAG employs a collaborative methodology that combines information retrieval mechanisms with the contextual learning capabilities of LLMs, utilizing both fixed-parameter LLMs and non-fixed-parameter data storage (such as text blocks in a knowledge base). In this paradigm, user queries first connect with an external knowledge base, using search algorithms to retrieve relevant documents [15]. These documents are then incorporated into the LLM's prompts, providing additional context for generating responses. A key advantage of RAG is that it removes the need to retrain the LLM for specific tasks, and developers can easily improve the accuracy of model outputs by augmenting the external knowledge base. The RAG approach has been shown to effectively enable contextual learning from retrieved documents, significantly reducing the risk of generating hallucinated content [16].

With the rise of models like ChatGPT, RAG technology has rapidly developed. Recently, a series of studies have developed domain-specific question-answering systems that integrate specialized knowledge bases. These systems have significantly improved their ability to handle interdisciplinary issues through a modular design approach. For instance, Liu et al. [17] addressed the exponential growth of logical form candidates through linearly growing primitives and comparative ranking methods, thereby achieving efficient, composable, and zero-shot question answering on knowledge bases and databases. Additionally, RnG-KBQA tackles coverage challenges and enhances generalization capability through comparative ranking of candidate logical forms and a generative model based on questions and top-ranking candidates [17]. These advancements indicate that RAG technology has immense potential in specialized question-answering systems, effectively tackling complex knowledge-intensive tasks.

To date, LLMs have been widely applied in a large number of research fields. For instance, prompt engineering has guided ChatGPT to automatically extract synthesis conditions for metal–organic frameworks from the scientific literature [18]. In the medical question-answering domain, BiomedRAG integrates a retrieval-augmented model with the biomedical field, directly inputting retrieved text blocks into the LLMs, enabling the LLMs to perform exceptionally well on various biomedical NLP tasks [19]. In the legal domain,

Louis et al. [20] proposed an end-to-end approach that employs a retrieval-reading process to provide comprehensive answers to any legal question. In the open question-answering space, PaperQA combines retrieval augmentation and AI agents to address questions about scientific literature, demonstrating superior performance on current scientific question-answering benchmarks compared to existing LLMs [21].

In the field of AI for environmental science, intelligent assistants based on LLMs are transforming traditional research processes. Zhu et al. [22] noted that ChatGPT's popularity stems from its ability to provide quick, informative, and seemingly "intelligent" answers to a wide variety of questions. The authors summarized several beneficial areas, including writing improvement, key point and theme identification, sequential information retrieval, as well as coding, debugging, and syntax explanation. However, they also cautioned researchers about potential issues such as the generation of fabricated information, the lack of updated domain knowledge, insufficient accountability in decision-making, and the opportunity cost associated with relying on ChatGPT.

Furthermore, LLM-driven systems can accelerate research processes through autonomous execution of tasks, showing high efficiency, particularly in the construction of adverse outcome pathways [23]. These systems can quickly extract key information from the literature, build causal networks, align closely with expert-validated findings, and provide more in-depth insights. While there are current limitations, ongoing advancements in AI technology and collaboration between AI systems and human experts show promise for the future of AOP construction. By harnessing the strengths of LLMs, we can improve our understanding of the adverse effects of environmental pollution and better protect public health through more effective risk assessment and regulatory decision-making.

Xu et al. [24] summarized the use of generative artificial intelligence in environmental science and engineering. In particular, the authors proposed some applications, such as designing new treatment processes, developing environmental models, and evaluating environmental policies. Meanwhile, the authors mentioned that the significant challenges include obtaining and creating specialized datasets prior to model construction, as well as ensuring the accuracy of outputs throughout the model development and usage phases.

A recent case highlights the role of LLMs and the Q&A system in revolutionizing water resource management, research, and policymaking [25]. After posing several questions to ChatGPT, the author concluded that integrating AI, particularly deep learning and advanced language models like ChatGPT, offers transformative opportunities in this field. Key points include enhanced understanding, democratization of knowledge, decision-making levels, sustainability, and vast potential.

However, most studies only pointed out that LLMs could be widely applied and useful in environmental science, and few practices have already been established. A case [26] assesses two generative pretrained transformer (GPT) models and five fine-tuned models (FTMs) using a specialized question-answering dataset, focusing on relevance, factuality, format, richness, difficulty, and domain topics. Results reveal that GPT-4 scored 0.644 in relevance and 0.791 in factuality across 286 questions, with scores dropping below 0.5 for more challenging questions, indicating a need for improvement. In contrast, FTMs with larger datasets maintained factual accuracy, emphasizing the importance of high-quality training materials. The study highlights issues of inaccuracies and format problems tied to overtraining and catastrophic interference, and uses expert-level textbooks to enhance LLM performance, paving the way for the development of more robust domain-specific LLMs for environmental applications.

Saeid et al. [27] enhanced GPT-4 by integrating access to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6). The conversational AI prototype, accessible at www.chatclimate.ai, is designed to tackle challenging questions

through three distinct configurations: GPT-4, ChatClimate, and Hybrid ChatClimate. Expert evaluations of the responses generated by these models indicate that the Hybrid ChatClimate AI assistant provides significantly more accurate answers.

Ren et al. [28] trained an LLM to become a hydrology expert, termed as WaterGPT, which is utilized in three primary domains: data processing and analysis, intelligent decision-making support, and interdisciplinary information integration. The model has demonstrated promising results, particularly through its careful segmentation of training data during the supervised fine-tuning phase. These data are derived from real-world sources and are annotated with high precision, utilizing both manual techniques and annotations from GPT-series models. The data are categorized into four distinct types: knowledge-based, task-oriented, negative samples, and multi-turn dialogues.

Liang et al. [29] developed a framework utilizing GPT-based text mining to extract information related to oxidative stress tests. This framework encompasses several key components: data collection, text preprocessing, prompt engineering, and performance evaluation procedures. The authors extracted a total of 17,780 relevant records from 7166 articles, encompassing 2558 unique compounds. Interestingly, over the past two decades, there has been a noticeable increase in interest regarding oxidative stress. This research led to the establishment of a comprehensive list of known prooxidants (n = 1416) and antioxidants (n = 1102), with the primary chemical categories for prooxidants being pharmaceuticals, pesticides, and metals, while pharmaceuticals and flavonoids were predominant among antioxidants.

Recently, scholars from Peking University developed a web app called Water Scholar (https://www.waterscholar.com/ (accessed on 5 January 2025)). This project is a free research assistant application for water science, based on the Wenxin large model. The app offers several features, including the ability to search for literature in the field of water, generate literature reviews, answer professional knowledge questions, and create citation lists.

3. Research Gap, Aims, and Objectives

As mentioned above, there are only a couple of cases or applications that have been established based on the use of LLMs in the research areas of environmental science to date [26–29]. Using human health risk assessment as the case, while health risks are a prerequisite for chemical safety and green usage, there is no knowledge retrieval system to assist non-professionals in quickly familiarizing themselves with this field to date. Here, focusing on the field of human health risk assessment, the aim of this study is to develop a question-answering system based on RAG technology. To achieve this goal, our specific objectives are to: (1) generate question–answer pairs as the testing dataset; (2) to develop the Naive RAG and advanced RAG-based Q&A system; (3) to evaluate the performance on question–answer pairs under various techniques; and (4) to design a multi-knowledge base integrated Q&A system. The study presented here may shed some useful information on the optimal retrieval methods, promising to offer a scientific basis for the further design and improvement of Q&A systems.

All code can be found at https://github.com/donkeyEEE/POPs_LLM (accessed on 2 November 2024).

4. Methods and Materials

4.1. Study Framework

To develop a knowledge-based LLM system using retrieval-augmented generation, this study was divided into three parts (Figure 1). (1) The generation of knowledge question–answer pairs: This step involved collecting and systematically organizing literature in the

field of human health risk assessment to extract key information and form question–answer pairs. These question–answer pairs not only provide a benchmark for the large language model, but also help reveal its limitations in this domain. (2) The development of a Q&A system integrating multiple knowledge base retrievals: The development will be based on a thorough analysis of existing knowledge retrieval technologies (as stated in the Introduction) to ensure precise and comprehensive answers to relevant questions. The system helps efficiently answer specific questions posed by non-experts, significantly reducing their time costs in re-learning and information retrieval in interdisciplinary research. (3) Performance evaluation: by comparing the accuracy of different strategies in answering scientific questions, this will provide a scientific basis for further optimization and improvement of the Q&A system.

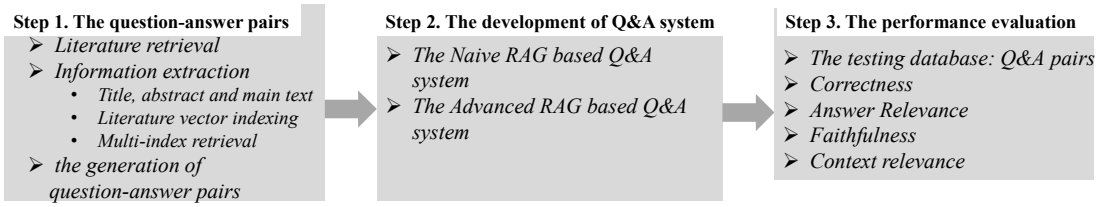

Figure 1. The study framework on establishing a retrieval-augmented generation-based question-answering (Q&A) system.

4.2. The Generation of Question–Answer Pairs

The generation of question–answer pairs consisted of three steps: literature retrieval, information extraction, and question–answer pairs generation. Literature retrieval involved keyword searches on the PubMed platform and Scopus, categorizing the research field of human health risk assessment into six submodules: analytical method, transport and fate, environment exposure, toxicokinetics, toxicity, and human health risk.

To ensure the scientific rigor and comprehensiveness of the literature retrieval, we referenced both classic literature and the latest research findings in the relevant field when selecting keywords. We conducted a precise screening based on the core themes and characteristics of subfields in human health risk assessment, ensuring coverage of key concepts and research directions in the field. Different keywords were used for searches, with the keyword table presented in Supplementary Material Table S1, and the categorization method and keyword selection criteria detailed in Supplementary Material Text S1. Subsequently, both manual and automated scripts were employed to download PDF documents from PubMed and Scopus. Through the PubMed API and web scraping methods, metadata including titles, abstracts, and publication years were gathered as the data foundation for this study [30]. It is important to note that our research does not exhaust all literature. Here, we attempted to use sufficient documents to build a knowledge vector database, which will further support the question-answering system. Exhausting all literature in the field would place a considerable computational burden on the server. Therefore, we set a literature cap of 500 for the analysis methods field and 200 for other fields.

To reduce the hallucination issues of LLMs in various domains, this study designed an automated question–answer pair generation process, significantly improving efficiency compared to traditional manual annotation methods. As shown in Figure 2, this process consists of three main steps: first, leveraging the LLM's contextual learning ability and appropriate prompt engineering, the literature were input into the model, which was transformed as the literature vector indexing. The aim of this step is to generate three different questions. Next, for each question, a multi-index retrieval-based LLM generated answers from the corresponding literature, annotating the source to ensure accuracy and verifiability. Finally, the system evaluated and selected the highest-quality question–answer

pair among the three. Following the PubMedQA approach [31,32], the question–answer pairs were saved in a structure that includes the question, answer content, source, and literature DOI. This automated process not only enhances the efficiency of question–answer pair generation, but also ensures high quality and practicality of the information through systematic screening.

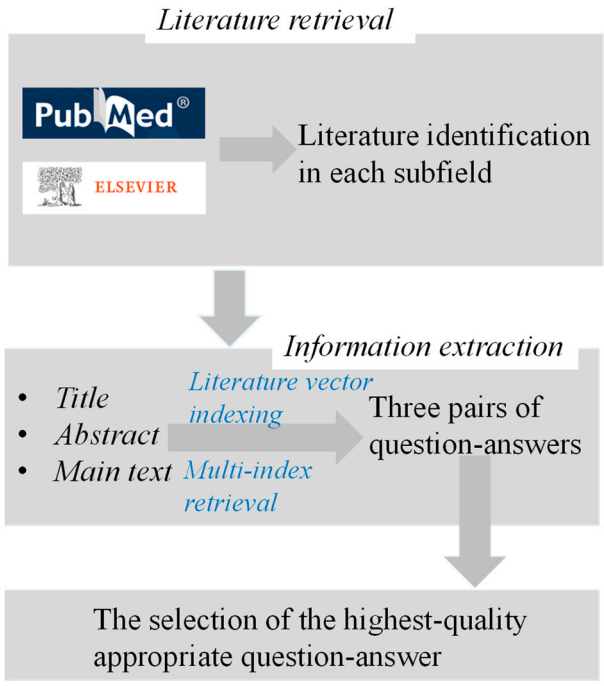

Figure 2. The process on the generation of question–answer pairs.

4.3. Naive Retrieval-Augmented Generation-Based Question-Answering System

Naive Retrieval-Augmented Generation (Naive RAG) is one of the earliest RAG methods [33], employing a traditional "retrieve-read" framework. In this framework, data were first indexed, then retrieval was performed based on user queries, and finally, the retrieved information was used as context to generate responses. This framework features a simple yet representative retrieval-augmented structure. This makes it widely used for comparative evaluations against more complex retrieval-augmented techniques. As a fundamental framework, Naive RAG provides a unified reference standard that aids in assessing the improvements of new methods in both retrieval performance and generation quality. The main drawbacks of Naive RAG include low retrieval quality, limited quality of generated responses, and potential loss of context when integrating retrieved information.

As shown in Figure 3, the construction process of a Q&A system based on RAG is divided into two parts: building a vector knowledge base and implementing the Q&A process. The vector knowledge base construction involved document segmentation and vectorization of chunks. Since the collected literature was presented in *PDF* format, which cannot be directly read by computers, we used the *PyPDF* library within the *LangChain* framework to convert PDF documents into string format [34]. *PyPDF* is a widely used Python library for processing PDF files, capable of various operations related to PDF documents, such as reading, splitting, merging, cropping, and converting PDF pages, as well as extracting text, images, and metadata. After extracting the text, we used *LangChain*'s fixed-length text splitting method to segment the literature into blocks of 1000 characters each. Once the documents were chunked, the resulting sub-documents were required to be vectorized, a step that transforms text into high-dimensional vectors, completed by

OpenAI's *text-embedding-3-large* model, which captures semantic information and represents it as fixed-length vectors.

In the Q&A process, the vectorization of user questions was similarly involved, using the same embedding model for consistent vector processing. The distance between vectors can represent the semantic similarity of two text segments. Based on this vector-matching principle, we can compute the cosine similarity between vectors to find the documents in the knowledge base that are semantically closest to the user's question. This document was used as the context for the question and was input along with the question into the prompt template, leveraging the context-learning capabilities of the LLM to improve answering effectiveness.

This process was mathematically represented as follows: given a user question q and a set of document contents $\{d_1, \ldots, d_m\}$, using the embedding model $\mathrm{EM}(\cdot)$ and cosine matching algorithm $\mathrm{sim}(\cdot)$, the top K retained document contents *context_K*:

$$context_k = \{d_i | i \in top\ K_j(sim(EM(q), EM(d_j)))\} \quad (1)$$

Then, the answer would be generated by the *LLM*:

$$Answer = LLM(prompt + context_k) \quad (2)$$

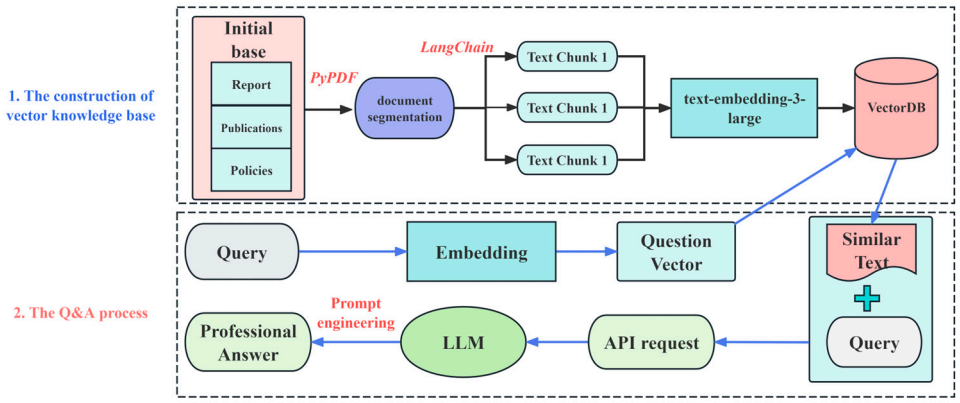

Figure 3. The flow of the Naive Retrieval-Augmented Generation-based question-answering (Q&A) system.

4.4. Advanced Retrieval-Augmented Generation Question-Answering System

To overcome the drawbacks of Naive RAG, Advanced RAG presented in this study introduces more complex techniques such as query rewriting, document reordering, and prompt summarization, aimed at improving retrieval relevance and the quality of generated text [35]. In summary, Advanced RAG optimizes data indexing through pre-retrieval and post-retrieval strategies, and enhances the quality of the retrieval process via techniques like fine-grained segmentation and reordering.

Semantic vector matching often encounters failures due to sometimes unclear semantic relationships between questions and document content. For example, a study focusing on per- and polyfluoroalkyl substances (PFASs) may suggest that altering agricultural practices can reduce PFASs' environmental impact, while the question could specifically inquire about PFASs' effects on water quality. In such cases, direct semantic matching may not accurately retrieve the most relevant information, necessitating deeper understanding and analysis. Additionally, we assumed that solving certain questions requires information from the literature; however, current technology frequently struggles to fully and accurately identify *PDF*-formatted documents, leading to noise that may misalign with the original text, adversely affecting answer generation.

To tackle these issues, this research designs an advanced retrieval-augmented framework (termed as Advanced RAG) that adds two modules—dual-layer retrieval and clue extraction—to Naive RAG. Before retrieving chunked documents in Naive RAG, we incorporate vector matching of questions and document summaries. To mitigate the impact of noise on the question-answering effectiveness, we added an information extraction module based on LLMs to gather question-relevant clues from the retrieved chunked documents. The framework (Figure 4) mainly consists of four processes: paper search, chunk search, gather evidence, and answer the question based on evidence.

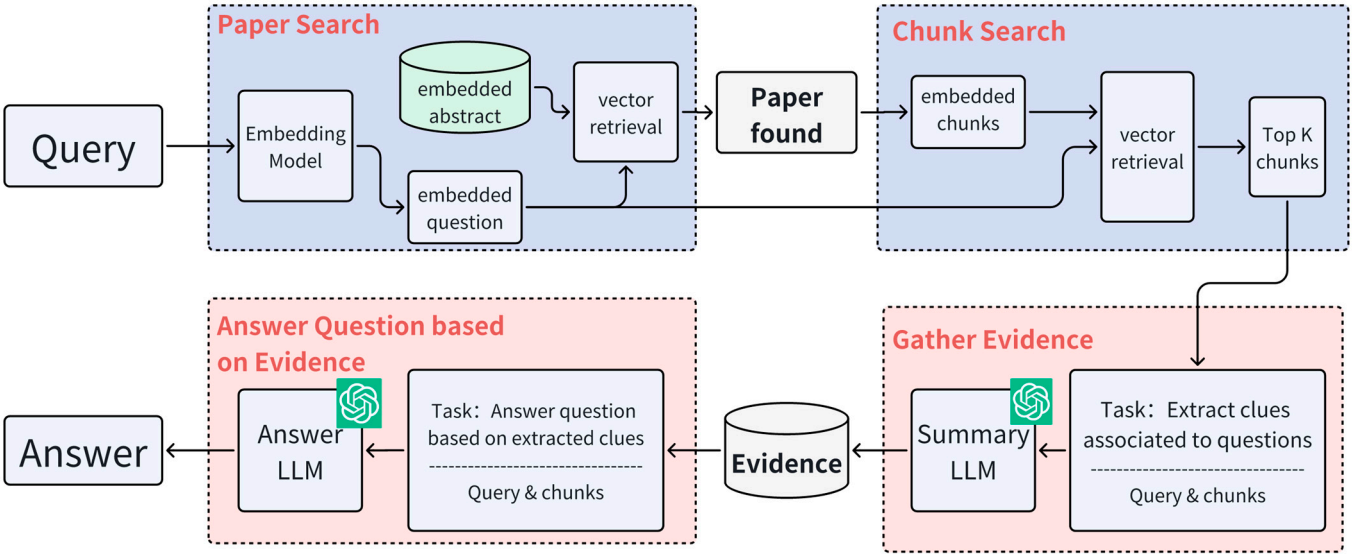

Figure 4. The design of the Advanced RAG question-answering system.

Paper search: The goal of this step is to identify the literature most relevant to the user's question. First, we vectorized the question q using the embedding model $EM(\cdot)$, which captured the semantic features of the question. For this, we used the *text-embedding-3-large* model provided by *OpenAI*. Then, we matched this vector with the literature abstracts $\{A_1, \cdots, A_m\}$ in a pre-constructed vector database using cosine similarity $sim(\cdot, \cdot)$, resulting in a collection of literature references $\{Paper_1, \ldots, Paper_n\}$. This can be expressed mathematically as follows:

$$Paper\ found = \left\{ Paper_i \,\middle|\, i \in top\ K_j\ (sim(EM(A_j),\ EM(q))) \right\} \tag{3}$$

Chunk search: In this step, we aimed to retrieve the most relevant content segments from the literature collection obtained in the previous step. Similar to previous literature retrieval, we also used vector matching techniques, but here we operated at a finer granularity, specifically on the document paragraphs, referred to as $\{chunks\}$. Using the literature collection gathered in the previous step as a filtering criterion, we searched for the content most relevant to the question from a vector database constructed based on the literature content. In this step, we employed maximum marginal relevance search $mmr(\cdot, \cdot)$ for vector matching, which ensures that the retrieved results are not only highly relevant to the question, but also diverse from one another, thereby enhancing both the diversity and relevance of the retrieved content. This can be expressed mathematically as follows:

$$\begin{aligned} &Top\ K\ chunks = \left\{ chunk_t \middle| t \in top\ K_j\ (mmr(EM(q),\ EM(chunk_j))) \right\} \\ &\left\{ chunk_1, \ldots,\ chunk_j \right\} \in\ Paper\ found \end{aligned} \tag{4}$$

Gather evidence: In this step, we gathered the evidence relevant to the question from the retrieved literature. This was accomplished through effective prompt engineering. Firstly, this step minimized irrelevant noise, including parsing errors that may occur when identifying *PDF* documents, allowing for a more streamlined question-answering process, and providing the option to completely reject certain segments. Secondly, independent extraction of multiple segments can occur simultaneously, thereby saving processing time. Each piece of information is represented by the following equation:

$$evidence = \begin{cases} LLM(prompt + chunk_t) \\ None,\ if\ no\ available\ \text{information provided by LLM} \end{cases} \quad (5)$$

Answer question based on evidence: Finally, the previously collected relevant information was combined into a specific prompt template and provided to the LLM. The prompt includes elements of a reasoning chain, guiding the LLM to infer step-by-step to generate an answer. The LLM synthesizes these clues to produce a coherent and logical response or, in cases of insufficient clues, chooses to refuse to answer, thereby avoiding incorrect or misleading answers. This step ensures that the final answer is accurate and evidence-based, enhancing the reliability and quality of the question-answering system while providing an option to decline when necessary.

4.5. The Evaluation on the Question-Answer System

In this study, we utilized the following indices (correctness, answer relevance, faithfulness, and context relevance) to evaluate the performance in the individual research field of human health risk assessment [36].

Correctness. The correctness of an answer primarily involves two aspects: the factual accuracy, and the semantic similarity between the answer and ground truth. These two aspects were combined through a weighted approach to obtain the final correctness score.

For factual accuracy (F_c), using the LLM, we can split the generated answer (A) and the reference answer (RA) into multiple simpler sentences. This step was defined as $S(\cdot)$, and thus we have obtained two sets:

$$S(q, A) = \{a_1, \ldots, a_m\}\ a_i \in \text{A} \quad (6)$$

$$S(q,\ RA) = \{\text{r}_1, \ldots, r_n\}\ r_j \in \text{RA} \quad (7)$$

The correctness F_c quantifies the factual overlap between the generated answer and the reference answer:

$$F_c = \frac{|TP|}{|TP| + 0.5 \times (|FP| + |FN|)} \quad (8)$$

where true positives (TP) are the facts that are present in both the generated answer and the reference answer, and false positives (FP) are statements that are present in the generated answer but do not appear in the reference answer. False negatives (FN) are statements that appear in the reference answer but do not appear in the generated answer.

On another note, semantic similarity (Ass) evaluates the semantic similarity between the generated answer and the reference answer, with values ranging from 0 to 1. A higher score indicates greater consistency between the answers. Measuring the semantic similarity between answers provides valuable insights into the quality of the generated responses. In this study, we used OpenAI's *text-embedding-3-large* model to vectorize the text and then compute the cosine similarity between the semantic vectors.

Finally, by combining the factual correctness and the answer semantic similarity with weighted factors, we obtained the overall correctness of the answer:

$$Answer\ Correctness = w_1 * F_c + w_2 * Ass \tag{9}$$

where w_1 and w_2 are the weights. In this study, we assumed a w_1 of 0.75 and a w_2 of 0.25.

Answer relevance. The answer relevance (*AR*) aims to assess the relevance of generated answers to the questions posed. Answers that are incomplete or contain redundant information are assigned lower scores, while higher scores indicate better relevance. In our study, first, given the generated answer A, a set of questions related to A was generated using a language model, $\{d_1, \ldots, d_m\}$, where each sub-question q_i is directly related to the answer. The relevance of the answer to the question was calculated as the average semantic similarity between each sub-question and the original question, using the same method for computing semantic similarity as described earlier.

$$\mathrm{AR} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{m} sim(EM(q),\ EM(q_i)) \tag{10}$$

Faithfulness. Faithfulness is used to measure the factual consistency between the generated answer and the given context. This is determined by generating answer (*A)* and the provided context *context(q)*. If all of the statements made in the answer can be inferred from the given context, the generated answer is considered to be faithful. To calculate this, a set of statements was first extracted from the generated answer. Then, each of these statements was cross-checked against the given context to determine whether it can be inferred from it. After that, two calls to the LLM were made. The first call attempted to split a segment of the answer into a set of statements, denoted as function $S(\cdot)$, with the statement set represented as $S = \{s_1, \ldots, s_m\}$. The second call determined whether each individual statement could be inferred from the context, denoted as function $V(s_i, context(q))$. The set of all statements that can be inferred was stated as $V = \{s_i | V(s_i, context(q)) == True\}$. Finally, the faithfulness of the answer was termed as the proportion of the number of elements in set V to the total number of elements in set S.

Context relevance. Generally, the retrieved context should only contain essential information necessary to address the provided query. Given a question q and associated context *context(q)*, this study determined context relevance (*CR*) by evaluating the proportion of critical information within the context. First, an LLM was used to extract a set of sentences S that were crucial to the question from the context. Then, the proportion of S in context was calculated as follows:

$$\mathrm{CR} = \frac{|\mathrm{S}|}{|\mathrm{all}\ sentences\ in\ context(q)|} \tag{11}$$

This metric was used to evaluate the quality of the context obtained from different retrieval methods, with values ranging from 0 to 1, where a higher value indicated better retrieval quality.

The demo can be found at the GitHub via https://github.com/donkeyEEE/POPs_LLM (accessed on 2 November 2024).

5. Results and Discussion

5.1. The Evaluation on Generation of Question–Answer Pairs

Based on the process stated in Supplementary Material Text S1 and the keywords provided in Table S1, this study has collected a total of 1500 articles. The number of articles in the dataset varied with publication time, as shown in Supplementary Material Figure

S1, indicating a rapid increase from 2000 to 2020, especially during 2015–2019, where the number of articles increased by approximately 167% compared to 2010–2014. The number of articles from 2020 to 2024 remained on par with that of 2015–2019.

In the question–answer pairs generation process, the basic procedure involves converting topics into questions and then using research content to provide answers. However, some topics are not suitable for conversion into questions (in fact, only about 65–75% of the literature is appropriate for generating Q&A pairs), leading to a mismatch between the number of generated question–answer pairs and the quantity of literature. As shown in Supplementary Material Figure S2, the number of question–answer pairs varied with publication time, illustrating a similar trend to the number of publications as plotted in Supplementary Material Figure S1. To ensure the balance of the dataset, we selected 50 high-quality question–answer pairs from each field, totaling 300 pairs for the test dataset. Here, we presented a pair from a study on the biodegradation of phthalic acid esters (PAEs), as shown in Supplementary Material Figure S3, with another three examples available in Supplementary Material Table S2. Each entry in the database is stored as a dictionary containing the question, answer, source_context, DOI, and publication time.

This study has demonstrated that prompt templates can significantly enhance the quality of generated question–answer pairs. As illustrated in Supplementary Material Figure S4, the case presented was derived from a study on the impact of perfluorooctanesulfonic acid (PFOS) on plant phosphate transporter gene networks. The figure compares the differences in question–answer pairs before and after the application of prompt engineering. When only basic prompts were used (consisting solely of a simple task description), the questions exhibited some ambiguity, and the explanations of the mechanisms in the answers were incomplete. For instance, the original question generated was: "*What was the focus of the study mentioned in the text regarding perfluorooctanesulfonic acid (PFOS) and plants*?" [37]. The reference to "*the study*" introduced semantic vagueness, as readers might not clearly understand which specific research was being referred to. After optimization through prompt engineering, the question was restructured to: "*What role do phosphate transporters play in PFOS sensing in plants?*". This prompt engineering significantly improved the precision of the question and the relevance of the answer.

In this section, we designed and implemented an innovative automated question–answer pair generation process. Compared to manual annotation, our approach significantly improved the efficiency and quality of question–answer pair generation. Additionally, by optimizing prompts, we enhanced both the precision of question formulations and the relevance and comprehensiveness of the answers. Ultimately, this study generated 300 high-quality standard question–answer pairs, which will serve as a benchmark dataset for evaluating the performance of Q&A systems.

5.2. The Performance Evaluation

Based on 300 high-quality question–answer pairs, we conducted performance testing on a naive RAG question-answering system integrated with advanced retrieval techniques, including dual-layer retrieval, RAG-Fusion, and Step-back, alongside four commonly used LLMs (gpt-3.5-turbo, gpt-4: https://platform.openai.com/docs/models/ (accessed on 12 August 2024); glm-3-turbo, and glm-4: https://github.com/THUDM/ChatGLM3/blob/main/README_en.md (accessed on 12 August 2024)) in the individual research field of human health risk assessment. The accuracy of the models' responses to questions is shown in Table 1. Results indicated that the advanced retrieval-enhanced Q&A system performed best across five research subfields, with accuracy ranging from 0.606 to 0.723. This performance surpassed all four foundational large language models, including the

largest GPT-4 model. Only in the health risk assessment domain was the advanced Q&A system's accuracy (0.583) slightly below that of the naive RAG Q&A system (0.599).

These findings highlight the superiority of the advanced retrieval-enhanced Q&A system in handling complex Q&A tasks. The system effectively integrates information from various sources through the dual-layer retrieval mechanism, while RAG-Fusion further optimizes the information merging process, and the Step-back mechanism allows for necessary backtracking during answer generation to ensure accuracy and comprehensiveness. The synergy of these techniques significantly improves the system's answer correctness across multiple domains.

Table 1. The performance of different question-answering systems on the answer correctness.

Subfield	gpt-3.5-turbo	gpt-4	glm-3-turbo	glm-4	Naive RAG	Advanced RAG
analytical method	0.679	0.676	0.665	0.685	0.663	0.723
transport and fate	0.458	0.510	0.515	0.568	0.631	0.655
exposure	0.463	0.467	0.468	0.423	0.544	0.606
toxicokinetics	0.333	0.376	0.410	0.491	0.570	0.625
toxicity	0.509	0.492	0.478	0.491	0.605	0.631
human health risk	0.351	0.381	0.403	0.543	0.599	0.583

It is also worth noting that in specific domains, such as toxicity, *gpt-3.5-turbo* (with an accuracy of 0.509) outperformed its upgraded version, *gpt-4* (accuracy of 0.492). A similar situation was observed between *glm-3-turbo* and *glm-4*. This phenomenon may be related to the introduction of false positives during the evaluation process. The presence of false positives can negatively impact the assessment of longer responses, as they may be incorrectly deemed irrelevant or incorrect despite being accurate in content. Additionally, we observed that answers generated by the *GLM-4* model are often more verbose, which may affect the accuracy of the evaluation results in certain cases.

Compared to the Naive RAG model, the retrieval-enhanced system generally demonstrates a significant advantage in improving accuracy. This indicates that information extracted from abstracts is more beneficial than that from full text, particularly since abstracts generally contain less noise and can be directly obtained, while full-text extraction from *PDF* documents can lead to context loss. Additionally, the dual-layer retrieval mechanism allows for more precise extraction from the vector database, minimizing noise. These mechanisms enable our Advanced RAG system to achieve the highest accuracy in testing with question–answer pairs.

Additionally, we evaluated the performance on answer relevance. As shown in Supplementary Material Table S3, most models performed well in terms of answer relevance, with scores generally close to or exceeding 0.9. The Advanced RAG Q&A system achieved relevance scores that were close to or higher than those of other models across six domains, indicating that the generated answers were highly relevant to the questions, with concise content and minimal redundancy. Although answer relevance assessment does not directly correlate with accuracy, high-relevance answers typically contain more useful information, reflecting the model's ability to respond to user queries effectively.

We presented the faithfulness and context relevance of both the Naive RAG and Advanced RAG systems (see Table 2). Both systems demonstrated very high fidelity (over 90%), indicating that the answers provided by the models are mostly derived from the retrieved content rather than being hallucinated, thus ensuring factual consistency with the knowledge base. On the other hand, the context relevance of the Naive retrieval-enhanced system was generally higher than that of the Advanced RAG.

In summary, based on the performance of different Q&A systems on these 300 question–answer pairs, we found that the Advanced RAG system achieved the highest accuracy, followed by the Naive RAG system, both outperforming large language models. Additionally, both Advanced RAG and Naive RAG demonstrated excellent results in answer relevance, faithfulness, and context relevance. These testing results indicate that our Advanced RAG system, built on a large literature database, is well-suited to address relevant professional questions in the field of health risk assessment.

Table 2. The performance of Naive RAG and Advanced RAG on the faithfulness and context relevance.

Subfield	Faithfulness		Context Relevance	
	Naive RAG	Advanced RAG	Naive RAG	Advanced RAG
analytical method	0.936	0.938	0.476	0.414
transport and fate	0.937	0.973	0.394	0.296
exposure	0.973	0.980	0.299	0.238
toxicokinetics	0.959	0.979	0.237	0.254
toxicity	0.915	0.953	0.361	0.350
human health risk	0.965	0.973	0.356	0.462

5.3. The Ablation Experiment for Advanced RAG

Ablation study refers to the process of removing or "ablating" different parts of a model to evaluate the impact of each component on the model's performance [38]. Through ablation studies, we can gain insights into the internal mechanisms of the model and understand the importance and contributions of various components. As shown in Figure S5, the entire Advanced RAG Q&A system is mainly composed of four components: *Retrieval 1* (summary retrieval), *Retrieval 2* (content retrieval), *Information Extraction* (IE), and *Answer Generation*. There are also two auxiliary components, *RAG-Fusion* and *Step-back*, which are located within the summary retrieval and answer generation components, respectively.

As shown in Table 3, in the subfield of exposure, transport and fate, and human health risk, the removal of the IE component led to the largest drop in accuracy, with decreases of 6.3%, 6.8%, and 6.7%, respectively, indicating the importance of the IE component for system performance. The decline in performance after removing IE may be due to the need for the system to recognize and convert PDF-formatted literature during loading, which introduces noise. Without IE, this noise can directly enter the LLM's input, preventing the LLM from identifying effective information to answer questions. Additionally, when splitting PDF documents, we used the *Recursive Character Text Splitter* method provided by *LangChain* (https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/ (accessed on 23 July 2024)), which directly cuts the document into specified sizes (set to 1000 in this study), potentially leading to the loss of some contextual information and formatting details.

In the fields of analytical methods and toxicity, the performance decline was greatest when main content retrieval was removed, with decreases of 13.2% and 8.7%, respectively, compared to the full system. This indicates that retrieving the main text content of the literature plays a positive role in answering questions in these specific fields. The absence of summary retrieval led to a decline in performance across six areas, with the largest drop in the toxicokinetics field (7.9% decrease), further underscoring the importance of abstract matching. In summary, results from the ablation experiment well demonstrates the importance of individual components in the Advanced RAG Q&A system.

Table 3. The correctness of the Advanced RAG-based question-answering system after removing associated components.

Subfield	Without RAG-Fusion	Without Step-Back	Without Summary	Without Main Context	Without Information Extraction	Advanced RAG
analytical method	0.697	0.634	0.654	0.631	0.636	0.723
transport and fate	0.604	0.612	0.612	0.640	0.587	0.655
exposure	0.463	0.467	0.468	0.423	0.606	0.463
toxicokinetics	0.563	0.587	0.546	0.560	0.554	0.625
toxicity	0.578	0.605	0.561	0.544	0.572	0.631
human health risk	0.563	0.572	0.557	0.574	0.516	0.583

5.4. The Design of a Multi-Knowledge Base Integrated Question-Answering System

This study has demonstrated that retrieval-augmented generation can improve the LLM's answering capability in knowledge-intensive tasks. Meanwhile, it is important to note that our study was conducted within specific subfields. However, in practical operations or the design of a knowledge Q&A system, the first challenge when a user poses a question is to identify the specific subfield and related subfields. This is because knowledge in a field should be subdivided into multiple sub-knowledge modules, and a comprehensive knowledge question-answering system should be able to retrieve and integrate appropriate information from these modules, synthesizing answers from multiple sources. However, previous knowledge base Q&A systems always considered embedding documents into the same vector space for retrieval, whether once or multiple times. These methods assume that all knowledge is treated as a single entity to find the most relevant content to assist the LLM in answering questions, without considering the need for interdisciplinary knowledge. These approaches would limit the efficiency and accuracy of the Q&A system's responses.

Therefore, we have built an advanced retrieval system across six domains as submodules, integrating them into a comprehensive system (Figure 5). When given a user question, the knowledge from these modules will be selectively activated to provide knowledge blocks relevant to the question. Given the LLM's limitations on context size and processing speed, it is not feasible to input excessive information indefinitely. Thus, before knowledge serves as input to the LLM, this study will use a ranking technique to filter out the most relevant knowledge blocks related to the question.

Intent recognition and task distribution: Upon receiving a user question, the Q&A system first invokes the LLM and asks, '*Do you need additional information to solve this problem*?' to get a YES or NO response. If YES is chosen, the system then prompts the LLM to select the appropriate module from six knowledge modules to retrieve the needed information.

Retrieving knowledge modules: A total of six knowledge modules have been established, each consisting of an advanced retrieval system with dual-layer retrieval and information extraction. Specific configurations of the Advanced RAG system can be found in Section 4.4. It is important to note that this only includes the retrieval component of the Advanced RAG Q&A system, meaning each module outputs K knowledge blocks, which are derived from the literature content retrieved from each domain and processed through LLM information extraction. However, if each knowledge module outputs K pieces of information (with K set to 10), the LLM would receive up to 60 pieces of information during the final answering phase. Considering the LLM's context length limitations, ranking techniques will be applied to filter the knowledge block collection to the top-k knowledge blocks.

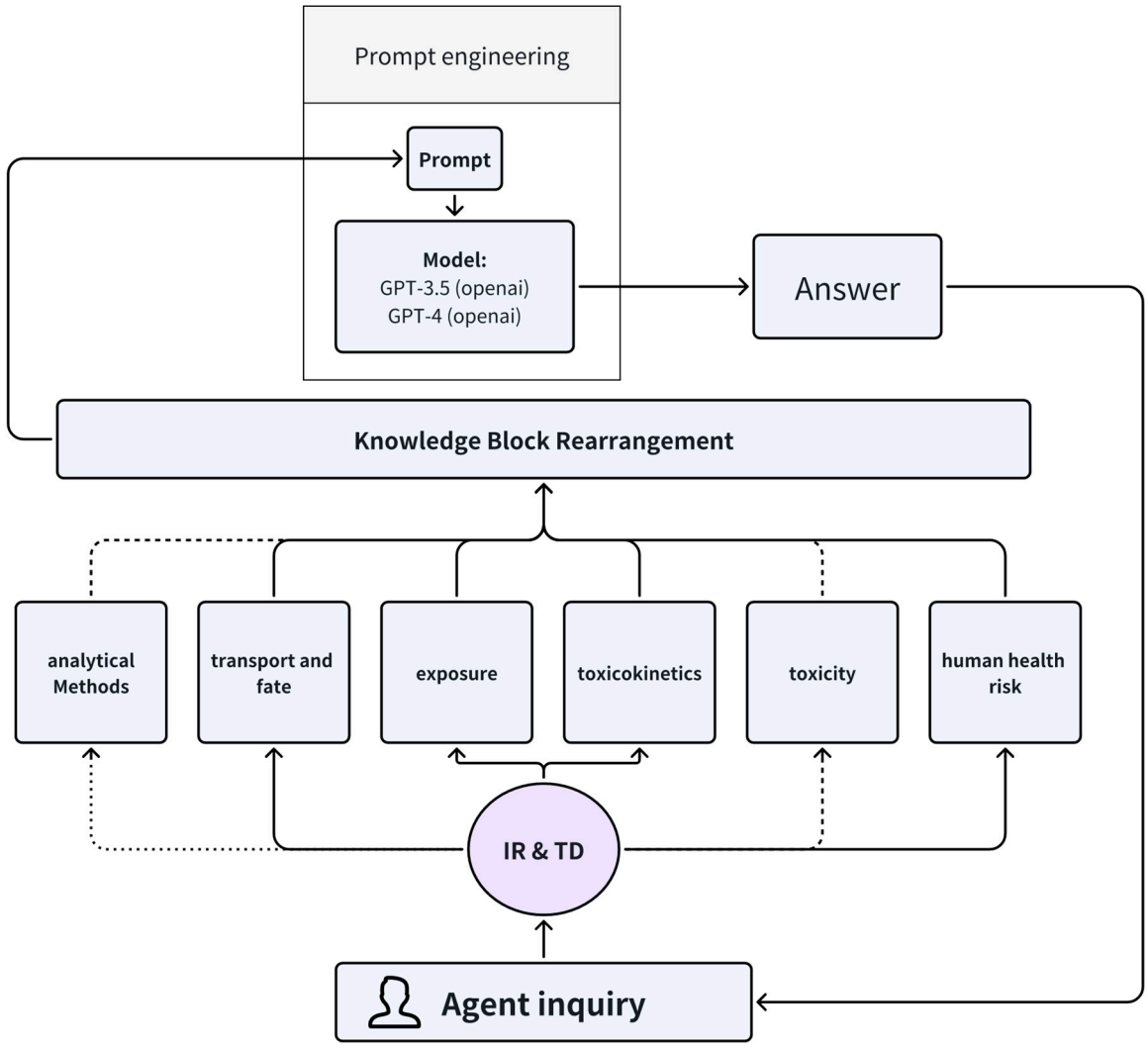

Figure 5. The flowchart of a multi-knowledge base integrated question-answering system. Abbreviations: IR, intent recognition; TD, task distribution.

Generating answers from integrated knowledge blocks: The information from the knowledge blocks would be incorporated into a carefully designed prompt that utilizes a chain-of-thought approach, guiding the LLM on how to utilize the knowledge block information to think through and gradually solve the user's problem.

In practical applications, not all questions require multiple retrieval enhancements to obtain the correct answer. Actually, some questions can be directly answered by the LLM, while relevant information for others may not be retrievable from the literature database. Considering the costs associated with LLM calls and the time required for answering, the complete system design involves various processing branches beyond the process described in the previous section. This LLM system, also referred to as an *Agent*, can intelligently select branches and complete the branching processes.

Hence, to better meet practical needs, this study expands on the aforementioned process (Figure 5) by adding more information processing branch steps, as illustrated in Supplementary Materials Figure S6. We utilize the *LangGraph* framework to develop the entire Q&A system. Specifically, this framework conceptualizes a Q&A system as a directed acyclic graph composed of multiple data processing nodes and edges. Each step is abstracted into independent nodes, and each invocation of the LLM or retrieval from the vector database is included within these nodes. Directed edges connect these nodes, with the direction indicating the next step in data processing. Additionally, some edges are

conditional, allowing for branching to different data processing paths when the previous node returns specific values.

In summary, the integrated Q&A system proposed in this study significantly enhances flexibility and scalability by subdividing the knowledge base into multiple submodules. This modular design allows for independent updates and maintenance of each submodule without large-scale modifications to the overall system architecture. As new research areas or topics emerge, new submodules can be easily integrated without affecting the stability of the existing system. Moreover, the system can selectively activate relevant modules based on the specific requirements of the question, optimizing resource allocation and improving operational efficiency.

6. Conclusions

In this study, we established an automatic method for generating high-quality question–answer pairs and produced 300 relevant pairs in the field of health risk assessment. Secondly, we successfully developed an Advanced RAG Q&A system that integrates a dual-layer retrieval and information extraction mechanism, which incorporates novel retrieval techniques such as RAG-Fusion and Step-back. Testing results based on the question–answer pairs indicate that our developed system outperforms both naive retrieval systems and large language models without retrieval enhancement in terms of answer accuracy and relevance. This result validates the limitations of LLMs when handling specialized question-answering tasks and demonstrates that retrieval enhancement can alleviate this issue to some extent. Lastly, this study employed the *LangGraph* framework to abstract the entire data processing flow into a graph data structure and successfully integrated the advanced retrieval framework into a comprehensive Q&A system, thus providing users with an efficient information query and processing solution.

The theoretical significance of this research lies in the combination of RAG technology with large language models, optimizing knowledge retrieval methods in human health risk assessment and advancing the application of natural language processing technologies in specialized fields. Practically, the multi-knowledge base question-answering system we developed improves the efficiency of literature retrieval and information extraction, helping researchers obtain relevant knowledge more quickly and accurately. This system provides practical tools for health risk assessment and interdisciplinary collaboration, promoting decision support and knowledge sharing.

This study assumes that most specialized problems can be addressed by utilizing the facts, concepts, and processes from paper abstracts or content. We particularly emphasize that the paper abstract can provide concise key information and is often an effective starting point for problem solving, especially when dealing with high-level issues in specialized fields. However, we recognize that relying solely on the abstract may sometimes be insufficient, particularly when the abstract is overly brief or vague. Therefore, while this study relies on the paper's abstract, it also incorporates the content of the paper to ensure the accuracy and comprehensiveness of the information. Specifically, we employ a dual-retrieval strategy that combines the processing of both the abstract and the content, reducing the risk of bias or misguidance that may arise from relying solely on the abstract. We acknowledge that the information in the abstract may indeed have certain biases or limitations; thus, during the final decision-making process, we carefully evaluate the accuracy of the abstract and validate and supplement the information through further retrieval processes. This approach helps us ensure efficiency while minimizing potential misunderstandings.

The process of building the knowledge base in this study includes *PDF* conversion recognition and document chunking. We noted that if the *PDF* documents cannot be

accurately recognized, or if the retrieval algorithm fails to obtain sufficient valid information from the database, it may affect the final results. To improve the quality of the contextual content retrieved, this study adopted a dual-layer retrieval strategy, which alleviates the limitations of vector-matching retrieval algorithms to some extent, significantly enhancing the question-answering effectiveness. However, despite the precise recognition of PDF documents improving the validity of contextual information, the technology underlying this method remains immature, and our vector database still relies on the traditional "*document chunking—vectorization*" building process. Even with information extraction based on LLMs, noise may still affect the accuracy of the final LLM responses. Retrieval enhancement based on literature data inevitably encounters issues such as format recognition errors and retrieval accuracy. Future research directions should focus on integrating PDF recognition with retrieval enhancement to better address these issues.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics14020386/s1, Figure S1: The number of publications in various subfield during 2000–2024; Figure S2: The number of question-answer pairs in various subfield during 2000–2024; Figure S3: The example on the generation of question-answer pairs; Figure S4: The process on prompt engineering optimization. (a) the case; (b) the template for question generation; (c) and the template for answer generation; Figure S5: The detailed components of advanced RAG Q&A system; Figure S6: The integrated question-answering system flowchart; Table S1: The keywords of the six research fields; Table S2: The standard examples on the question-answer pairs generated by proposed framework; Table S3: The performance of different question-answering system on the text relevance; Text S1: The details of six subfields.

Author Contributions: W.M.: methodology, formal analysis, validation, writing—original draft, writing—review and editing; Y.L.: conceptualization, methodology, formal analysis, writing—original draft, writing—review and editing, supervision; L.C.: writing—review and editing and validation; Z.D.: conceptualization, methodology, formal analysis, writing—original draft, writing—review and editing, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities (Beihang [501LKQB2022133003] and Southeast University [4025002413]). The APC was funded by Fundamental Research Funds for the Central Universities [4025002413].

Data Availability Statement: All data are available in the main text and supplementary information. The code that supports the findings of this study is openly available at the following URL/DOI: https://github.com/donkeyEEE/POPs_LLM (Accessed on 2 November 2024).

Conflicts of Interest: Author Wenjun Meng was employed by the company Beijing Huadian E-commerce Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Naidu, R.; Biswas, B.; Willett, I.R.; Cribb, J.; Singh, B.K.; Nathanail, C.P.; Coulon, F.; Semple, K.T.; Jones, K.C.; Barclay, A. Chemical pollution: A growing peril and potential catastrophic risk to humanity. *Environ. Int.* **2021**, *156*, 106616. [CrossRef] [PubMed]
2. Fuller, R.; Landrigan, P.J.; Balakrishnan, K.; Bathan, G.; Bose-O'Reilly, S.; Brauer, M.; Caravanos, J.; Chiles, T.; Cohen, A.; Corra, L. Pollution and health: A progress update. *Lancet Planet. Health* **2022**, *6*, e535–e547. [CrossRef] [PubMed]
3. Wang, Z.; Walker, G.W.; Muir, D.C.; Nagatani-Yoshida, K. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories. *Environ. Sci. Technol.* **2020**, *54*, 2575–2584. [CrossRef]
4. Dong, Z.; Liu, Y.; Duan, L.; Bekele, D.; Naidu, R. Uncertainties in human health risk assessment of environmental contaminants: A review and perspective. *Environ. Int.* **2015**, *85*, 120–132. [CrossRef]
5. Zeise, L.; Bois, F.Y.; Chiu, W.A.; Hattis, D.; Rusyn, I.; Guyton, K.Z. Addressing human variability in next-generation human health risk assessments of environmental chemicals. *Environ. Health Perspect.* **2013**, *121*, 23–31. [CrossRef]

6. Kavlock, R.J.; Austin, C.P.; Tice, R. Toxicity testing in the 21st century: Implications for human health risk assessment. *Risk Anal. Off. Publ. Soc. Risk Anal.* **2009**, *29*, 485. [CrossRef]
7. Lioy, P.J.; Smith, K.R. A discussion of exposure science in the 21st century: A vision and a strategy. *Environ. Health Perspect.* **2013**, *121*, 405. [CrossRef]
8. Malhotra, M.; Nair, T.G. Evolution of knowledge representation and retrieval techniques. *Int. J. Intell. Syst. Appl.* **2015**, *7*, 18. [CrossRef]
9. Jiang, Z.; Chi, C.; Zhan, Y. Research on medical question answering system based on knowledge graph. *IEEE Access* **2021**, *9*, 21094–21101. [CrossRef]
10. Mollá, D.; Vicedo, J.L. Question answering in restricted domains: An overview. *Comput. Linguist.* **2007**, *33*, 41–61. [CrossRef]
11. Liu, Z.; Yang, Q.; Zou, J. Lowering Costs and Increasing Benefits Through the Ensemble of LLMs and Machine Learning Models. In Proceedings of the International Conference on Intelligent Computing, Tianjin, China, 5–8 August 2024; pp. 368–379.
12. Majeed, A.; Hwang, S.O. Reliability Issues of LLMs: ChatGPT a Case Study. *IEEE Reliab. Mag.* **2024**, *1*, 36–46. [CrossRef]
13. Dierickx, L.; Van Dalen, A.; Opdahl, A.L.; Lindén, C.-G. Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review. In Proceedings of the Multidisciplinary International Symposium on Disinformation in Open Online Media, Münster, Germany, 2–4 September 2024; pp. 1–15.
14. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
15. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking large language models in retrieval-augmented generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 17754–17762.
16. Ayala, O.; Bechard, P. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), Mexico City, Mexico, 16–21 June 2024; pp. 228–238.
17. Ye, X.; Yavuz, S.; Hashimoto, K.; Zhou, Y.; Xiong, C. Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering. *arXiv* **2021**, arXiv:2109.08678.
18. Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J.T.; Yaghi, O.M. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *JACS* **2023**, *145*, 18048–18062. [CrossRef]
19. Li, M.; Kilicoglu, H.; Xu, H.; Zhang, R. Biomedrag: A retrieval augmented large language model for biomedicine. *arXiv* **2024**, arXiv:2405.00465. [CrossRef]
20. Louis, A.; van Dijck, G.; Spanakis, G. Interpretable long-form legal question answering with retrieval-augmented large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 22266–22275.
21. Lála, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodriques, S.G.; White, A.D. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv* **2023**, arXiv:2312.07559.
22. Zhu, J.-J.; Jiang, J.; Yang, M.; Ren, Z.J. ChatGPT and environmental research. *Environ. Sci. Technol.* **2023**, *57*, 17667–17670. [CrossRef]
23. Shi, H.; Zhao, Y. Integration of Advanced Large Language Models into the Construction of Adverse Outcome Pathways: Opportunities and Challenges. *Environ. Sci. Technol.* **2024**, *58*, 15355–15358. [CrossRef]
24. Wu, Y.; Xu, M.; Liu, S. Generative Artificial Intelligence: A New Engine for Advancing Environmental Science and Engineering. *Environ. Sci. Technol.* **2024**, *58*, 17524–17528. [CrossRef]
25. Ray, P.P. Leveraging deep learning and language models in revolutionizing water resource management, research, and policy making: A case for ChatGPT. *ACS EST Water* **2023**, *3*, 1984–1986. [CrossRef]
26. Zhu, J.-J.; Yang, M.; Jiang, J.; Bai, Y.; Chen, D.; Ren, Z.J. Enabling GPTs for Expert-Level Environmental Engineering Question Answering. *Environ. Sci. Technol. Lett.* **2024**, *11*, 1327–1333. [CrossRef]
27. Vaghefi, S.A.; Stammbach, D.; Muccione, V.; Bingler, J.; Ni, J.; Kraus, M.; Allen, S.; Colesanti-Senni, C.; Wekhof, T.; Schimanski, T. ChatClimate: Grounding conversational AI in climate science. *Commun. Earth Environ.* **2023**, *4*, 480. [CrossRef]
28. Ren, Y.; Zhang, T.; Dong, X.; Li, W.; Wang, Z.; He, J.; Zhang, H.; Jiao, L. WaterGPT: Training a large language model to become a hydrology expert. *Water* **2024**, *16*, 3075. [CrossRef]
29. Liang, W.; Su, W.; Zhong, L.; Yang, Z.; Li, T.; Liang, Y.; Ruan, T.; Jiang, G. Comprehensive Characterization of Oxidative Stress-Modulating Chemicals Using GPT-Based Text Mining. *Environ. Sci. Technol.* **2024**, *58*, 20540–20552. [CrossRef] [PubMed]
30. Barupal, D.K.; Fiehn, O. Generating the blood exposome database using a comprehensive text mining and database fusion approach. *Environ. Health Perspect.* **2019**, *127*, 097008. [CrossRef]
31. Lamurias, A.; Sousa, D.; Couto, F.M. Generating biomedical question answering corpora from Q&A forums. *IEEE Access* **2020**, *8*, 161042–161051.
32. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv* **2019**, arXiv:1909.06146.

33. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-augmented generation for ai-generated content: A survey. *arXiv* **2024**, arXiv:2402.19473.
34. Jacob, T.P.; Bizotto, B.L.S.; Sathiyanarayanan, M. Constructing the ChatGPT for PDF Files with Langchain–AI. In Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 24–26 April 2024; pp. 835–839.
35. Cormack, G.V.; Clarke, C.L.; Buettcher, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, 19–23 July 2009; pp. 758–759.
36. Salemi, A.; Zamani, H. Evaluating retrieval quality in retrieval-augmented generation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024; pp. 2395–2400.
37. Kim, J.H.; Kroh, G.; Chou, H.-A.; Yang, S.-H.; Frese, A.; Lynn, M.; Chu, K.-H.; Shan, L. Perfluorooctanesulfonic Acid Alters the Plant's Phosphate Transport Gene Network and Exhibits Antagonistic Effects on the Phosphate Uptake. *Environ. Sci. Technol.* **2024**, *58*, 5405–5418. [CrossRef]
38. Sheikholeslami, S.; Meister, M.; Wang, T.; Payberah, A.H.; Vlassov, V.; Dowling, J. Autoablation: Automated parallel ablation studies for deep learning. In Proceedings of the 1st Workshop on Machine Learning and Systems, Scotland, UK, 26 April 2021; pp. 55–61.