

Supplementary Materials

For

**Using the retrieval-augmented generation to improve the question answering  
system in human health risk assessment: the development and application**

Wenjun Meng<sup>1,2</sup>, Yuzhe Li<sup>3\*</sup>, Lili Chen<sup>4</sup>, and Zhaomin Dong<sup>3,4\*</sup>

<sup>1</sup>, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100181,  
China;

<sup>2</sup>, Beijing Huadian E-commerce Technology Co., Ltd., Beijing 100164, China

<sup>3</sup>, School of Materials Science and Engineering, Beihang University, Beijing 100191, China

<sup>4</sup>, School of Public Health, Southeast University, Nanjing 210096, China

\*Corresponding authors:

Correspondence to Yuzhe Li, School of Materials Science and Engineering, Beihang University,  
Beijing, China; 20377027@buaa.edu.cn

Correspondence to Zhaomin Dong, School of Materials Science and Engineering, Beihang  
University, Beijing, China; dongzm@buaa.edu.cn

**Text S1.** The details of six subfields

In order to build a comprehensive knowledge base in the field of environmental pollutants and considering the limitations of retrieval algorithm performance due to the scale of the knowledge base, this study divides the field of human health risk assessment into six subfields based on the following criteria:

**Analytical methods:** As a fundamental and critical area, this focuses on the chemical and physical analysis techniques of environmental samples, including the use of instruments and experimental methods. Research in this field is vital as it provides precise data that underpins all environmental science studies, serving as the basis for decision-making and research.

**Transport and fate:** This field explores the behavior patterns of chemicals in the environment, including their biodegradation, photochemical transformations, and other processes. Research here aids in predicting the fate and long-term effects of pollutants in the environment, guiding environmental management and pollution control.

**Exposure:** This area studies the distribution and migration pathways of pollutants in the environment and the exposure of humans and ecosystems. Research in this field helps us understand how pollutants spread from sources to environments accessible to humans, providing essential information for environmental risk assessment and pollution prevention.

**Toxicokinetics:** This area examines how organisms respond to environmental stressors, such as their ability to accumulate, metabolize, and excrete xenobiotics. Research in this field reveals how pollutants can affect different organisms through the food chain, potentially leading to biomagnification effects that have far-reaching impacts on ecosystem health and human well-being.

**Toxicity:** This field focuses on the harmful effects of pollutants on human health and ecosystems. Through toxicological assessments and risk evaluations, research in this area provides a scientific basis for developing public health policies and environmental protection measures.

**Risk assessment:** This field primarily assesses the impact of environmental factors on public health, studying the health risks associated with specific pollutants. Research in this area provides the necessary scientific support for developing effective public health interventions and environmental policies.

The selection of keywords for the six subfields should clearly reflect the core research

content and the use of specialized terminology in each area. Below is a detailed explanation of the basis for keyword selection in each subfield:

**Analytical methods:** Keywords include “analytical techniques,” “chemical analysis,” and “instrumental analysis.” These terms collectively cover the fundamental techniques for analyzing environmental samples. “Analytical techniques” refers broadly to various analysis methods applicable to different environmental samples. “Chemical analysis” focuses on detecting chemical components, while “instrumental analysis” emphasizes precise measurements conducted using instruments. This set of keywords ensures comprehensive coverage from basic to advanced analytical techniques.

**Transport and fate:** Keywords include “transport mechanisms,” “pollutant fate,” and “biodegradation.” These terms explore the behavior of pollutants in the environment and their ultimate destinies. “Transport mechanisms” study the processes of pollutant transfer, “pollutant fate” focuses on their stability and transformation in the environment, and “biodegradation” refers to the natural degradation process under microbial action.

**Exposure:** Keywords include “exposure pathways,” “contaminant monitoring,” and “personal exposure.” “Exposure pathways” describe the routes through which pollutants spread from sources to humans or ecosystems. “Contaminant monitoring” focuses on the monitoring of pollutants in the environment, which is a key step in assessing environmental quality. “Personal exposure” evaluates the direct contact of humans with pollutants, which is crucial for public health research.

**Toxicokinetics:** Keywords include “bioaccumulation,” “biomagnification,” and “ecotoxicity.” “Bioaccumulation” describes the accumulation of pollutants within organisms, “biomagnification” studies the amplification effects of pollutants in the food chain, and “ecotoxicity” assesses the toxic impacts of chemicals on components of ecosystems.

**Toxicity:** Keywords include “acute toxicity,” “chronic toxicity,” and “genotoxicity.” This set of terms describes short-term, long-term, and genetic toxicity, respectively. “Acute toxicity” measures the effects of high-dose exposure over a short period, “chronic toxicity” concerns the effects of long-term low-dose exposure, while “genotoxicity” studies the potential damage of chemicals to genetic material.

**Human health risk:** Keywords include “epidemiological studies,” “risk assessment,” and

“public health impact.” These terms link environmental factors to public health. “Epidemiological studies” are used to observe and analyze the impacts of environmental factors on population health, “risk assessment” evaluates potential health risks, and “public health impact” studies the overall effects of environmental issues on public health.

**Table S1.** The keywords of the six research fields

Groups	Keywords
analytical method	(analytical techniques [Title/Abstract]) OR (chemical analysis [Title/Abstract]) OR (instrumental analysis [Title/Abstract])
transport and fate	(exposure pathways [Title/Abstract]) OR (contaminant monitoring [Title/Abstract]) OR (personal exposure [Title/Abstract])
exposure	(transport mechanisms [Title/Abstract]) OR (pollutant fate [Title/Abstract]) OR (biodegradation [Title/Abstract])
toxicokinetics	(bioaccumulation [Title/Abstract]) OR (biomagnification [Title/Abstract]) OR (ecotoxicity [Title/Abstract])
toxicity	(acute toxicity [Title/Abstract]) OR (chronic toxicity [Title/Abstract]) OR (genotoxicity [Title/Abstract])
human health risk	(epidemiological studies [Title/Abstract]) OR (risk assessment [Title/Abstract]) OR (public health impact [Title/Abstract])

**Table S2.** The standard examples on the question-answer pairs generated by proposed framework

Publication time	DOI	Question	Answer
2023	10.1002/mrc.5350	What are the benefits and challenges of using synergistic sample preparation and multiple analytical techniques in untargeted metabolomics?	A multiplatform (multiple analytical techniques) approach can improve upon the number of metabolites reliably detected and correctly assigned. This can be further improved by applying synergistic sample preparation along with the use of combinatorial or sequential non-destructive and destructive techniques. Similarly, peak detection and metabolite identification strategies that employ multiple probabilistic approaches have led to better annotation decisions. Applying these techniques also addresses the issues of reproducibility found in single platform methods. Nevertheless, the analysis of large data sets from disparate analytical techniques presents unique challenges. While the general data processing workflow is similar across multiple platforms, many software packages are only fully capable of processing data types from a single analytical instrument. Traditional statistical methods such as principal component analysis were not designed to handle multiple, distinct data sets. Instead, multivariate analysis requires multiblock or other model types for understanding the contribution from multiple instruments.
2007	10.1007/s11306-007-0082-2	What are the proposed minimum reporting standards for chemical analysis in metabolomics experiments?	The proposed minimum reporting standards for chemical analysis in metabolomics experiments include sample preparation, experimental analysis, quality control, metabolite identification, and data pre-processing. These standards primarily focus on mass spectrometry and nuclear magnetic resonance spectroscopy techniques. Adherence to these standards is crucial for providing a biological and empirical context for metabolomics data sets, facilitating experimental replication, data

			<p>comparison, and overall research quality and reproducibility. The Chemical Analysis Working Group (CAWG) encourages additional input on other techniques related to chemical analysis in metabolomics experiments through the online discussion forum provided by the Metabolomics Standards Initiative.</p>
			<p>Recent analytical methodologies in lipid analysis primarily involve the use of mass spectrometry (MS) techniques, such as direct infusion MS or in combination with separation techniques like liquid chromatography or gas chromatography. Mass spectrometry is crucial for lipid identification and quantification. Additionally, emerging strategies like Ion Mobility-Mass Spectrometry and Ion Fragmentation via Ozonolysis are being utilized for lipidomic analysis to structurally characterize lipids with specificity, addressing the challenges posed by the high isomeric complexity of the lipidome.</p>
2024	10.3390/ijms25042249	What are the recent analytical methodologies used in lipid analysis?	<p>Furthermore, supercritical fluid chromatography (SFC) is a novel separation method used in metabolomics that can increase metabolome coverage, decrease cost, and analysis time. Despite its potential advantages, the application of SFC-MS in metabolomics is currently limited.</p> <p>In summary, recent advancements in lipid analysis involve a combination of mass spectrometry techniques, including emerging strategies like Ion Mobility-Mass Spectrometry and Ion Fragmentation via Ozonolysis, as well as innovations in separation techniques like supercritical fluid chromatography. These methodologies are essential.</p>

---

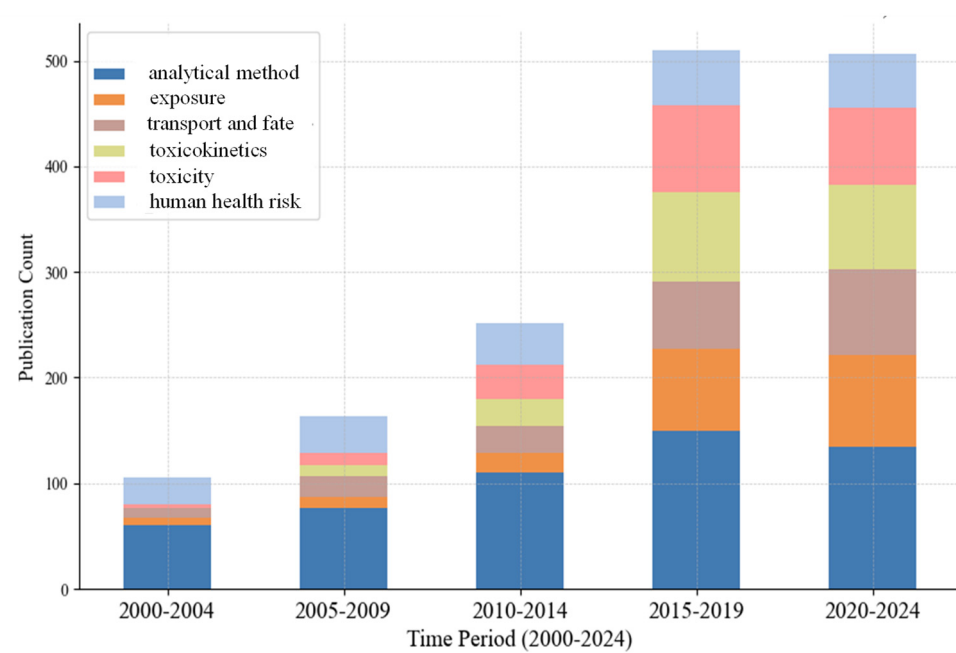
for accurate lipid identification, quantification, and structural  
characterization.

---

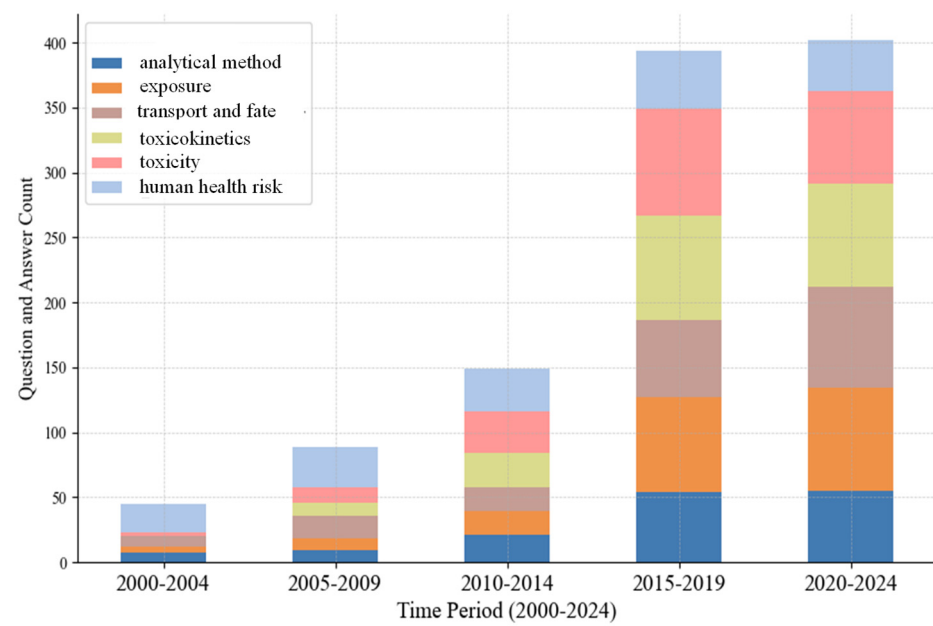


**Table S3.** The performance of different question-answering system on the text relevance.

Subfield	gpt-3.5-turbo	gpt-4	glm-3-turbo	glm-4	Advanced RAG
analytical method	0.959	0.965	0.928	0.953	0.957
transport and fate	0.941	0.962	0.933	0.865	0.949
exposure	0.966	0.958	0.873	0.940	0.962
toxicokinetics	0.931	0.930	0.851	0.949	0.952
toxicity	0.867	0.847	0.694	0.704	0.963
human health risk	0.933	0.940	0.887	0.907	0.952



**Figure S1.** The number of publications in various subfield during 2000-2024



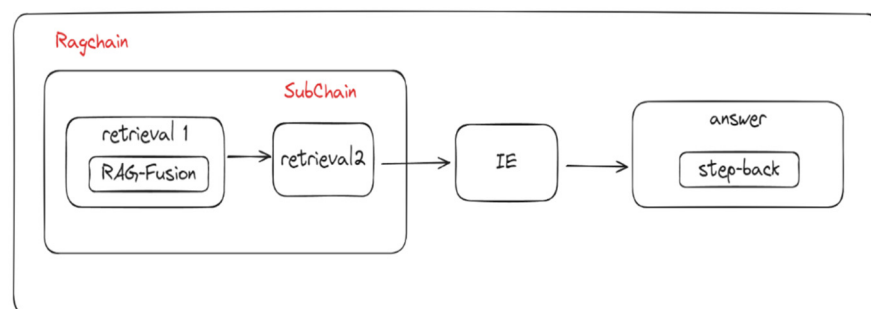
**Figure S2.** The number of question-answer pairs in various subfield during 2000-2024

Question: What are the current strategies for eliminating phthalic acid esters (PAEs) from environments using bacteria-driven biodegradation?

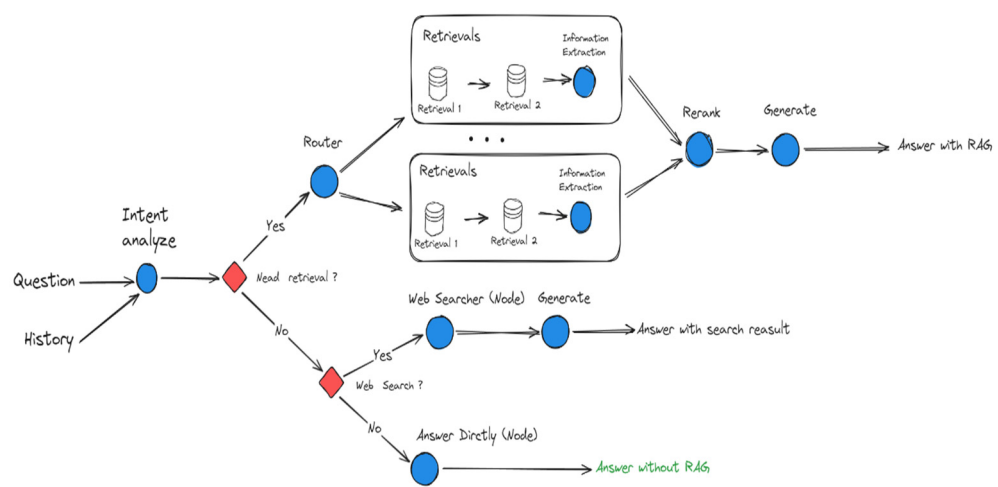
Answer: The strategy for eliminating phthalic acid esters (PAEs) from the environment using bacteria-driven biodegradation primarily relies on the selection and utilization of bacteria with the capacity to degrade PAEs, which can convert these compounds into harmless metabolic byproducts. Additionally, synthetic microbial ecology is employed to construct artificial bacterial communities that optimize the degradation process under controlled conditions, enhancing the efficiency of bioremediation and offering a cost-effective and environmentally friendly solution for environments contaminated with PAEs. This is considered an inexpensive yet effective approach to restoring polluted environments.

**Figure S3.** The example on the generation of question-answer pairs





**Figure S5.** The detailed components of advanced RAG Q&A system



**Figure S6.** The integrated question-answering system flowchart