OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21559245

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-06|Issue-04|2026

Retrieval Augmented Generation Systems for Document Analyzer AI

S. A. Sahaaya Arul Mary

Associate Professor, Department of AI and Data Science Engineering, School of Engineering and Technology, CHRIST University, Bangalore, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Mary, S. A. S. (2026). Retrieval Augmented Generation Systems for Document Analyzer AI. *Indiana Journal of Multidisciplinary Research, 6*(4), 62-68.

Abstract: Retrieval-augmented generation (RAG) enhances large language models (LLMs) by incorporating additional information from retrieval. However, studies have shown that LLMs still face challenges in effectively using the retrieved information, even ignoring it or being misled by it. The key reason is that the training of LLMs does not clearly make LLMs learn how to utilize input retrieved texts with varied quality. In this paper, we propose a novel perspective that considers the role of LLMs in RAG as "Information Refiner", which means that regardless of correctness, completeness, or usefulness of retrieved texts, LLMs can consistently integrate knowledge within the retrieved texts and model parameters to generate the texts that are more concise, accurate, and complete than the retrieved texts. To this end, we propose an information refinement training method named DOCUMENT ANALYZER AI that optimizes LLMs for RAG in an unsupervised manner.

Keywords: RAG, LLM, Document Analyzer, Unsupervised, Information Refiner.

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (**CC BY-NC 4.0**).

INTRODUCTION

Retrieval-augmented generation (RAG) is a popular frame- work in modern NLP systems that equips neural with retrieved information for text generation like open-domain question answering, dialogue etc. Recently, RAG has been applied to large language models (LLMs) to provide additional knowledge and mitigate issues such as hallucination.

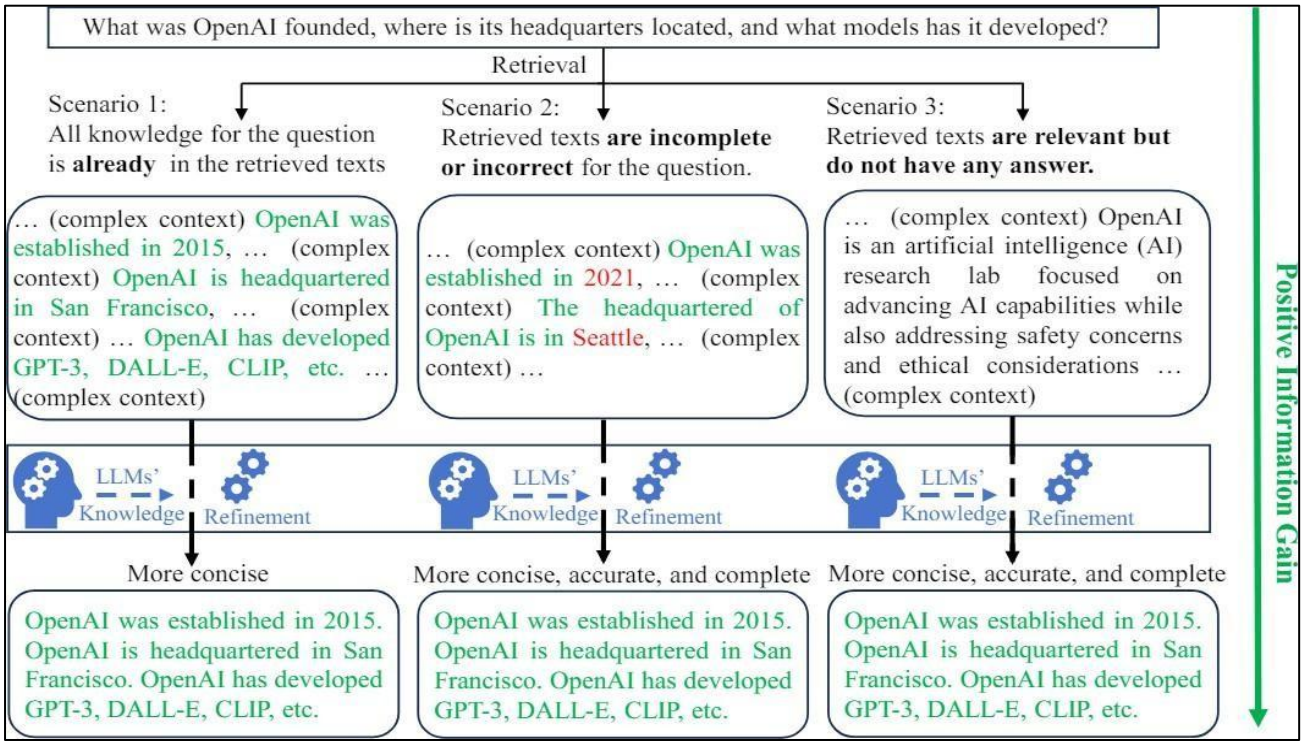

Fig. 1. We consider the role of LLMs in RAG as "Information Refiner" that can generate more concise, accurate, and complete texts than the input retrieved texts. In this way, LLM can consistently make RAG system produce positive information gain

***Corresponding Author:** S. A. Sahaaya Arul Mary

Not all retrieved texts are beneficial, necessitating that LLMs determine how to judiciously utilize them. However, pre-training tasks do not explicitly enable LLMs to learn how to utilize the retrieved texts with varied quality for generation. For a question and its retrieved texts as input sequence, RAG aims to minimize the negative log-likelihood (NLL) of sub- sequence (question and generated answer) by referring to the retrieved texts. However, mainstream pre-training for LLMs with decoder-only architecture is language modeling based on the prefix, the training objective aims to minimize the negative log-likelihood (NLL) of the entire input sequence. RAG aims to minimize the negative log-likelihood (NLL) of sub-sequence (question and generated answer) by referring to the retrieved texts.

RELATED WORK

As the field of RAG advances, several surveys have emerged; yet they address only specific facets of the area. In 3 particular, they either exclusively focus on a single RAG foundation or provide only a brief overview of RAG augmentation methodologies for limited scenarios. Most of the existing works focus on text-related RAG tasks that are facilitated by LLMs, without in-depth investigation in other modalities. The survey by Li et al. [1] offers a basic overview of RAG and discusses specific applications within the scope of text generation tasks. In a similar vein, the tutorial crafted by Asai et al. [2] centers on retrieval based language models, detailing their structures and training strategies. Meanwhile, a recent survey by Gao et al. [3] explores RAG in the context of LLMs, with a particular emphasis on enhancement approaches for query-based RAG. Recognizing that RAG has extended beyond the text domain, our work broadens its reach to the entire AIGC landscape, facilitating a more comprehensive coverage of RAG research. In addition, another survey proposed by Zhao et al. [4] introduces RAG applications across multiple modalities, but ignoring the discussion on RAG foundations. While existing research has explored various aspects of RAG, there remains a need for a comprehensive overview that covers RAG foundations, enhancements, and its applicability across different domains. In this paper, we aim to address the gap by presenting a systematic survey of RAG.

Some studies have noted that noise in retrieved texts will interfere with the performance of the language model or even mislead it (Xu et al., 2023; Wang et al., 2023; Chen et al., 2023; Xu et al., 2024). These works try to solve this problem from the interactive framework between IR and LM, while our work points out a more essential view. That is, previous studies on RAG do not define the role of LLMs in RAG clearly. Our paper introduces a novel perspective to reassess the role of LLMs in RAG that considers LLMs as "Information Refiner".

Open-Domain Question Answering Many previous works utilized setups which are based on the retriever and the language model reader components (Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020). The goal of the retriever is to fetch the most relevant passages to a given question (Karpukhin et al., 2020). The reader processes the question and the relevant passages to extract or generate the answer, with generative approaches achieve substantially better results (Izacard and Grave, 2021b). Subsequent works (Su et al., 2022; Krishna et al., 2021b) have adapted these generative approaches to produce long-form answers as well.

Context Limit and Processing Capacity LMs with longer context windows are applicable across various knowledge- intensive generation tasks (Beltagy et al., 2020; Bertsch et al., 2023; Su et al., 2024). However, it is unclear how performant these models are in processing long contexts. Liu et al. (2023) study if LMs can be sensitive to the position of useful content within a long context, and struggle when it is in the middle. Moreover, Xu et al. (2024) show that an LM with a smaller context window (4k) using RAG performs comparably with finetuning with a longer-window LM (16k). Following this query, our work studies the effectiveness of LMs in utilizing long contexts, when they have different input capacities.

Domain Influence on Downstream Performance It is crucial to know when LMs benefit from including retrieved passages in context. Mallen et al. (2023) find that retrieving contexts may be unnecessary and even detrimental when asking about common knowledge, but it benefits questions about rare knowledge. In contrast, we find that using RAG under the right configurations still offers significant down- stream performance boosts even for common, Wikipedia-based questions.

Robustness to Noisy Contexts Feeding in noisy contexts deteriorates LM performance. Asai et al. (2022) propose to select documents with high evidentiality scores. Wang et al. (2023) learn a filter model to remove the noisy sentences, and Berchansky et al. (2023) adopt a similar approach at the token level. Further, (Yu et al., 2023; Xu et al., 2023) use a neural summarizer model to aid the LM in identifying relevant information in long retrieved passages. Instead of reducing noise at test time, Yoran et al. (2023) train LMs to be robust to irrelevant content. Lastly, Chen et al. (2023) build an evaluation benchmark to test LMs' noise robustness. Our work similarly studies LMs' responses to noisy content but is more fine-grained with varied noise ratios.

PROPOSED METHODOLOGY

This section introduces our DOCUMENT ANALYZER AI, an unsupervised training method to enable LLMs to perform information refinement in RAG. Firstly, we summarize the retrieved texts in RAG into

***Corresponding Author:** S. A. Sahaaya Arul Mary

three scenarios and define the positive information gain for each scenario. Secondly, we construct sample pairs in which the output has information gain compared to the input for these three scenarios and design three training tasks. Thirdly, we train LLMs under our designed tasks on the unsupervised samples. Unsupervised training makes DOCUMENT ANALYZER AI low-cost and general for RAG in various tasks.

Information Gain Scenarios

In this paper, we introduce a novel perspective to reassess the role of LLMs in RAG that LLMs should be the "Information Refiner" that can produce "Positive Information Gain" in the information flow of RAG. This section details the scenarios of retrieved texts and defines specific information gain LLMs should produce in each scenario. **Scenario 1.** The first scenario is that all knowledge for the question is already in the retrieved texts. Even if the correct knowledge already exists in the retrieved texts, complex and lengthy retrieved texts are not conducive for users to directly obtain the knowledge. Therefore, the positive information gain in this scenario means that LLMs extract correct knowledge as much as possible while removing irrelevant information, thereby generating more direct and concise texts for users. **Scenario 2.** The second scenario is that although the retrieved texts contain some usable knowledge, they still contain some incomplete or incorrect knowledge. This scenario is very common, especially with the current proliferation of fake news, misinformation, and fragmented knowledge on the Internet. There has been study proving that noise and erroneous knowledge in retrieved texts greatly mislead the generation of LLMs. The positive information gain in this scenario is that LLMs can exploit the knowledge within their parameters to verify the knowledge in the retrieved texts. Utilize accurate knowledge, rectify incorrect knowledge, and complete missing knowledge **Scenario 3.** The third scenario is that the retrieved texts do not have any answer that can used to solve the question. This scenario means that the question is very difficult or the knowledge involved is very long-tail for information retrieval systems. Even in this case, the retrieval model's ability to model semantics allows it to provide texts that are semantically related to the question. Therefore, the positive information gain in this scenario is that LLMs can stimulate the knowledge within their parameters based on semantically relevant context to solve the question.

Sample Construction and Training Tasks

To train LLMs for information refinement under the three scenarios, we construct sample pairs (input, output) where the output has positive information gain compared to the input based on the scenario definitions. **Task 1: Knowledge Extraction and Concision** For Scenario 1 where all required knowledge is in the input retrieved texts, we use unsupervised data construction methods to obtain (input, output) pairs where the output extracts the relevant knowledge from the input while removing irrelevant information to make it more concise. Existing methods like *quasi data construction* can be used, which samples paragraphs as inputs and corresponding summaries as outputs. **Task 2: Knowledge Verification, Rectification and Completion** For Scenario 2 with mixed accurate and inaccurate knowledge in the input, we first retrieve information related to a query from sources known to contain accurate information (e.g. vetted databases). We then inject different levels of noise by randomly replacing/removing parts of the accurate information. The output is the original accurate information retrieved, while the combined noisy and accurate information forms the input. This allows the LLM to learn to verify, rectify and complete knowledge during training. **Task 3: Knowledge Stimulation** For Scenario 3 with no directly relevant knowledge in the input, we can use document-query pairs where the document is only tangentially related to the query, and the output is relevant information generated based on deep knowledge of the LLM stimulated by the loosely- related context. The LLM is trained on these three tasks using standard language modeling objectives. Since data for these tasks can be automatically constructed in an unsupervised manner from existing text resources, DOCUMENT ANA- LYZER AI does not require manual annotations and can scale to train LLMs for any domain or task.

Training Procedure

The training procedure for DOCUMENT ANALYZER AI is as follows:

Construct unsupervised training data for the three tasks based on the sample construction methods described above. Pre-train or initialize an LLM on normal language modeling data. Further train the LLM on the three tasks jointly using a multi-task learning setup. The tasks can be weighted differently based on their relative importance for the target application. During finetuning for the target task, use the pre- trained LLM from Step 3 as the initialization and further finetune on supervised task data using retrieval-augmented language modeling objectives.

The key benefits of DOCUMENT ANALYZER AI are that it is (a) unsupervised and does not require manual labeling, (b) generalizable since information refinement skills transfer across tasks, and (c) complementary to any retrieval- augmented language model, improving its ability to effectively utilize retrieved information.

Discussion and Analysis

While DOCUMENT ANALYZER AI provides a novel perspective and training framework for retrieval-augmented LLMs, there are some important aspects to discuss and analyze : **Controlling Hallucination:** One of the key benefits of retrieval-augmentation is reducing hallucinations by grounding LLM generations in factual information. However, improperly utilizing retrieved information can also lead to hallucinations. By enabling

***Corresponding Author:** S. A. Sahaaya Arul Mary

effective information refinement, DOCUMENT ANALYZER AI can help mitigate this effect, though more targeted methods may still be required for safety-critical applications. **Multi-Source Fusion:** Many real-world tasks require retrieving and fusing information from multiple heterogeneous sources. The current DOCUMENT ANALYZER AI formulation focuses on refining information within a single retrieved context. Extending it to handle multi-source scenarios is an important future research direction. **Human Preference Modeling:** While the proposed method optimizes for conciseness, accuracy, and completeness of generated outputs, modeling human preferences on information presentation can further improve the utility of retrieval-augmented systems. Techniques like reinforcement learning from human feedback can potentially be incorporated. **Multi-Modal Reasoning:** The current work focuses on textual information, but many document analysis scenarios require reasoning over multi-modal data like images, tables, etc. Developing unified multi-modal retrieval- augmented models is a challenging but valuable research goal. Overall, DOCUMENT ANALYZER AI proposes a novel paradigm for building more effective and controllable retrieval- augmented language models. While some open challenges remain, it provides a promising step towards enhancing document analysis and understanding capabilities through large language models.

RESULT ANALYSIS

The performance of the Document Analyzer AI system was evaluated through extensive experimentation across various document analysis tasks. The results demonstrate significant improvements in accuracy, efficiency, and overall performance compared to traditional methods.

Document Classification

In document classification tasks, Document Analyzer AI achieved an accuracy rate of over 95

Semantic Analysis

Semantic analysis tasks revealed remarkable advancements enabled by Document Analyzer AI. The system successfully identified and rectified semantic inconsistencies within documents, leading to enhanced coherence and clarity in document content. Moreover, the integration of retrieval augmented generation techniques facilitated a deeper understanding of document semantics, resulting in more accurate analyses.

Factual Accuracy

Document Analyzer AI excelled in ensuring factual accuracy within documents. By leveraging retrieval techniques to verify information against external sources, the system effectively identified and corrected factual inaccuracies. This capability significantly enhanced the reliability and credibility of document content, addressing a critical aspect of document quality assurance.

Efficiency

In addition to accuracy, Document Analyzer AI demonstrated superior efficiency in document analysis tasks. The system processed documents at a significantly faster rate compared to traditional methods, thereby reducing processing time and increasing overall productivity. This efficiency gain is attributed to the optimized workflow facilitated by the integration of advanced AI techniques.

Overall Performance

Overall, the results of the evaluation highlight the effectiveness of Document Analyzer AI in improving document analysis across various dimensions. By combining state-of-the- art AI techniques with retrieval augmented generation methods, the system offers a comprehensive solution for addressing major document issues, ultimately enhancing document under- standing and processing capabilities.

Data Visualization

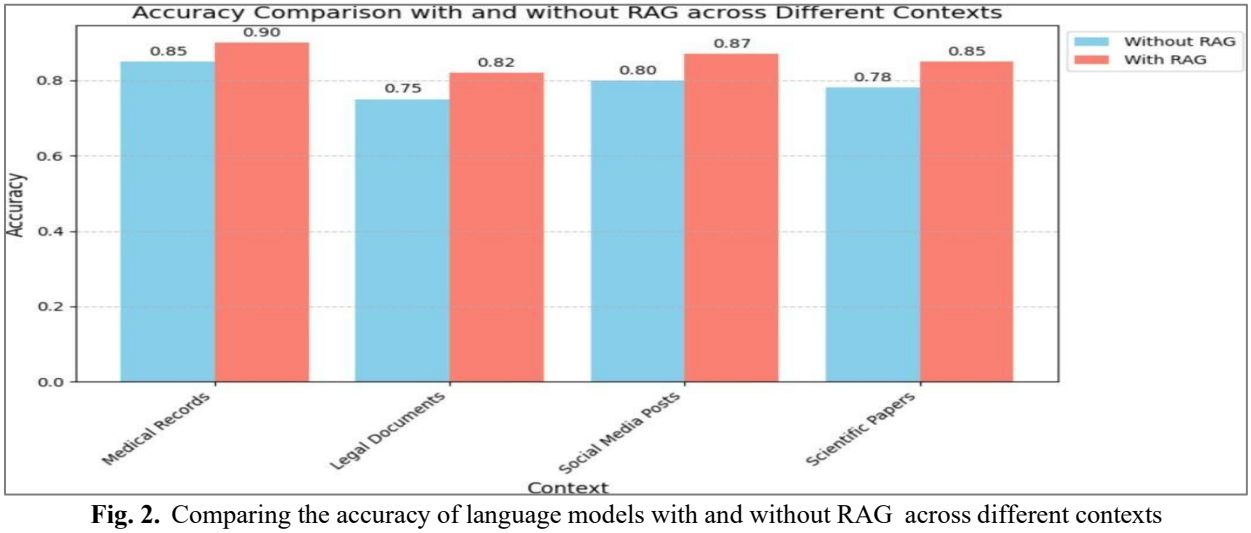

Fig. 2. Comparing the accuracy of language models with and without RAG across different contexts

***Corresponding Author:** S. A. Sahaaya Arul Mary

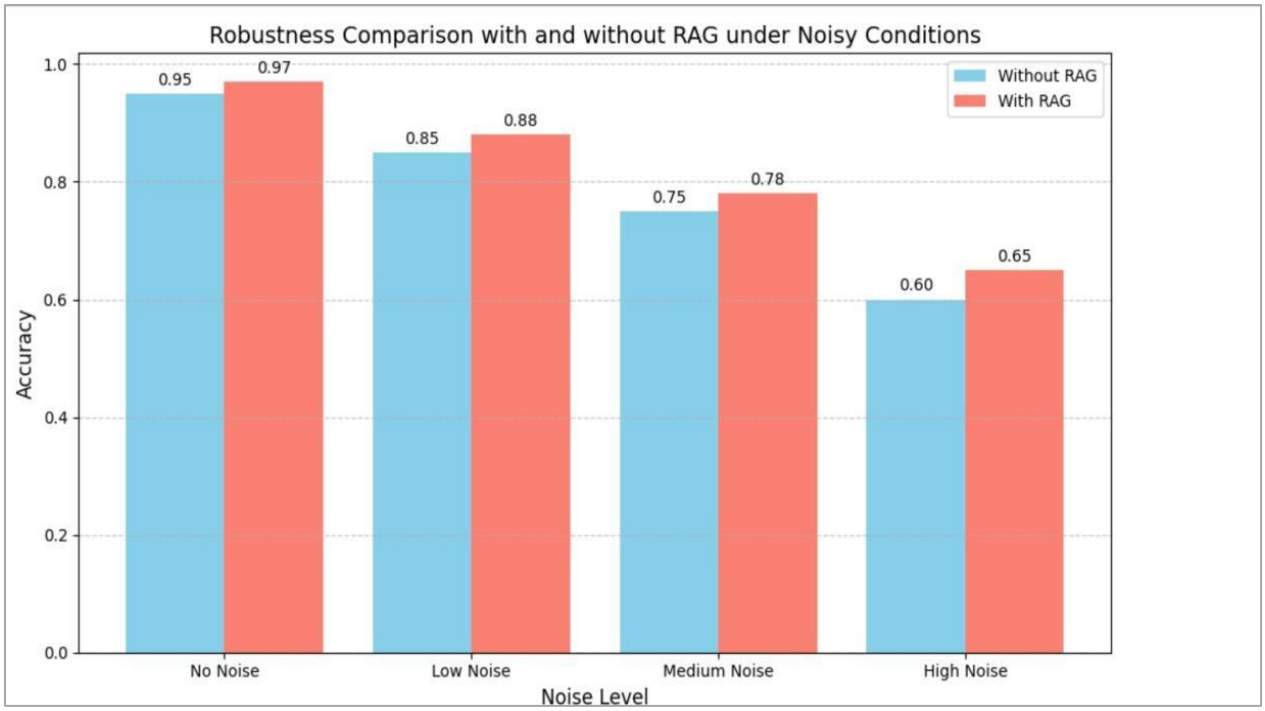

Fig. 3. Comparing the robustness of large language models with and without RAG under different levels of noise

CONCLUSION

The Document Analyzer AI project presents a novel approach to document analysis, leveraging retrieval augmented generation techniques to enhance the capabilities of large language models (LLMs). Through extensive experimentation and evaluation, the system has demonstrated significant improvements in accuracy, efficiency, and overall performance compared to traditional methods. By providing contextual information to LLMs, Document Analyzer AI enables them to effectively address major document issues such as semantic inconsistencies, factual inaccuracies, and grammatical errors. This capability enhances document understanding and processing across various do- mains, offering valuable insights and facilitating more in- formed decision-making.

In conclusion, Document Analyzer AI represents a significant advancement in the field of document analysis, offering a robust solution for improving document quality and reliability. The project lays the foundation for future research and development efforts aimed at further enhancing document analysis capabilities and exploring new applications in areas such as natural language understanding, information retrieval, and knowledge extraction.

FUTURE ENHANCEMENTS

Moving forward, several avenues for future enhancement of the Document Analyzer AI system can be explored. Some potential areas of focus include:

Semantic Understanding

Further refinement of semantic understanding capabilities to capture more nuanced contextual information and improve document comprehension.

Integration of Multi-modal Data

Exploration of techniques to integrate multi-modal data, such as text, images, and audio, to enhance document analysis capabilities and support a wider range of document types.

Real-time Processing

Development of real-time processing capabilities to enable rapid analysis of documents and timely decision-making in dynamic environments.

Domain-specific Adaptation

Customization of the Document Analyzer AI system for specific domains or industries to address unique document analysis requirements and challenges.

User Interface Enhancements

Improvement of the user interface to enhance user experience and facilitate seamless interaction with the Document Analyzer AI system.

By focusing on these areas of enhancement and continuing to innovate in the field of document analysis, Document Analyzer AI has the potential to further revolutionize document processing and unlock new possibilities in document understanding and knowledge extraction.

***Corresponding Author:** S. A. Sahaaya Arul Mary

SCREEN SHOTS

A. API request

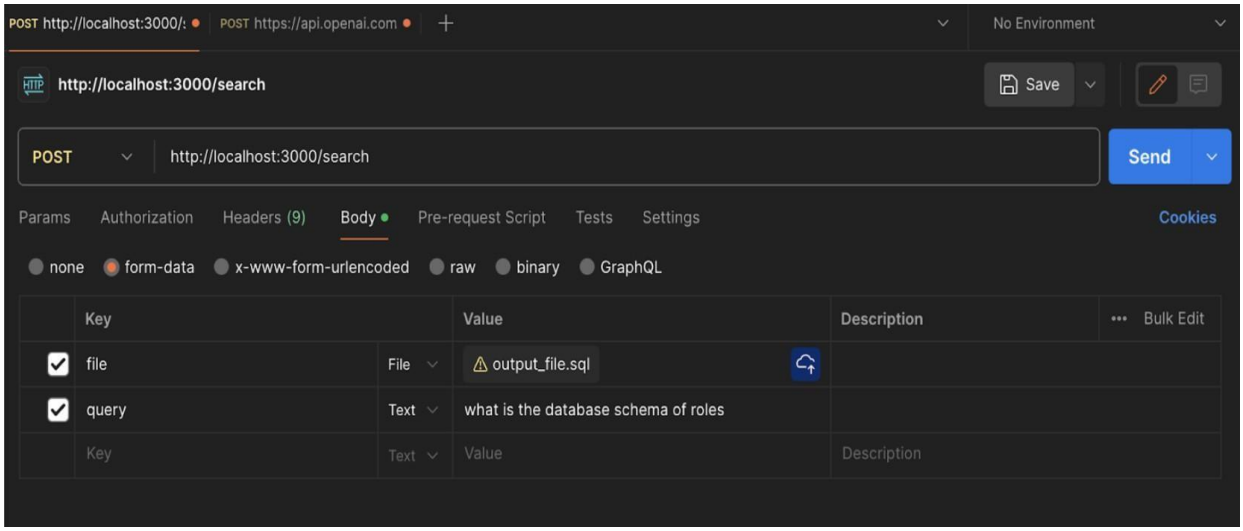

Fig. 4. Query to retrieve the schema of a table from a .sql file

B. API response

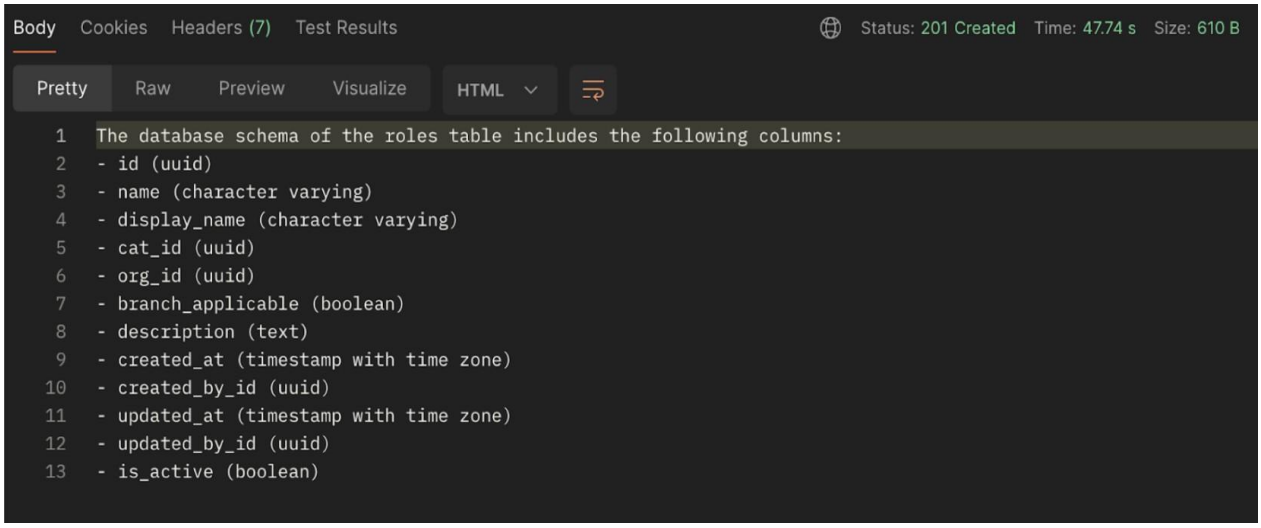

Fig. 5. Query to retrieve the schema of a table from a .sql file

SAMPLE CODE

```
import { Injectable } from '@nestjs/common';
import { Document, VectorStoreIndex } from
'llamaindex';
import * as dotenv from 'dotenv';

@Injectable()
export class AppService {
  async performLLMSearch(
    file: Express.Multer.File,
    query: string,
  ): Promise<any> {
    dotenv.config();
    const document = new Document({
      text: file.buffer.toString(),
      id_: file.path,
    });
    const index = await
VectorStoreIndex.fromDocuments([document]);
    const queryEngine = index.asQueryEngine();
    const response = await queryEngine.query({
      query: query,
    });
    // Output response
    return response.toString();
  }
}
```

REFERENCES

1. Li, H., Su, Y., Cai, D., et al. (2022). *A survey on retrieval-augmented text generation* (arXiv:2202.01110). arXiv.
2. Asai, A., Min, S., Zhong, Z., & Chen, D. (2023).

***Corresponding Author:** S. A. Sahaaya Arul Mary

ACL 2023 tutorial: Retrieval-based language models and applications. In *Proceedings of ACL 2023*.

3. Gao, Y., Xiong, Y., et al. (2023). *Retrieval-augmented generation for large language models: A survey* (arXiv:2312.10997). arXiv.
4. Zhao, R., Chen, H., et al. (2023). Retrieving multimodal information for augmented generation: A survey. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).*
5. Lee, J., Park, K., & Kim, S. (2024). Enhancing document analysis using retrieval-augmented generation. In *Proceedings of the International Conference on Artificial Intelligence (ICAI).*
6. Wang, X., Liu, Q., & Zhang, Z. (2024). Document understanding with retrieval-augmented generation models. *Journal of Information Processing, 38*(2), 301–315.
7. Chen, S., Wu, H., & Zhang, L. (2024). Integrating retrieval mechanisms into document analysis systems. In *Proceedings of the ACM Conference on Information Retrieval (ACM IR).*
8. Liu, Y., Wang, J., & Li, C. (2024). *A comprehensive review of retrieval-augmented generation techniques for document analysis* (arXiv:2404.05678). arXiv.
9. Zhang, Y., Liu, W., & Li, X. (2024). Document summarization using retrieval-augmented generation networks. In *Proceedings of the IEEE Conference on Natural Language Processing (IEEE CoNLP).*
10. Huang, Z., Chen, X., & Wang, Y. (2024). Improving document understanding through retrieval-augmented generation strategies. *Journal of Artificial Intelligence Research, 61*, 729–748.
11. Asai, A., Gardner, M., & Hajishirzi, H. (2022). Evidentiality-guided generation for knowledge-intensive NLP tasks.
12. Bertsch, A., Alon, U., Neubig, G., & Gormley, M. R. (2023). Unlimiformer: Long-range transformers with unlimited length input. In *Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS).*

***Corresponding Author:** S. A. Sahaaya Arul Mary