International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Real-Time Retrieval-Augmented Generation in Healthcare Using LLMs for Audio-based Querying and Retrieval

Kartik Arora[1]

[1]*Kartik Arora, Department of Networking and Communication, SRM Institute of Science and Technology*

Abstract - Healthcare information systems are critical for efficient decision-making and patient care. This paper presents a real-time audio-query-driven retrieval-augmented generation (RAG) system using a large language model (LLM) that interfaces with a healthcare database and a vector database. The system allows users to make real-time audio queries, retrieves the necessary information using LLM tools, and delivers audio responses. Comparative performance evaluations between traditional healthcare information systems and our proposed LLM-based approach show significant improvements in query throughput, cost-efficiency, and response time.

***Key Words*:** RAG, LLM, healthcare database, real-time audio input, vector database, AI in healthcare

1. INTRODUCTION

Healthcare systems manage vast amounts of data, including information on patients, hospitals, medications, and diagnoses. Traditional systems struggle with scalability, realtime query handling, and user-friendliness. This research introduces a retrieval-augmented generation system utilizing a large language model (LLM) for real-time audio-based querying, integrated with a healthcare database and a vector database for improved query accuracy and response efficiency. Our system enhances accessibility and speeds up healthcare data retrieval using voice input, making the process seamless for healthcare professionals and patients alike.

2. System Architecture

The architecture of the proposed system is shown in Fig. 1. The system comprises a user interface that accepts voice input, a healthcare vector database (VectorDB) to store relevant context, a large language model (LLM) to process queries, and tools that interact with the healthcare database (HealthcareDB) to fetch precise data. Upon receiving an audio query, the LLM augments its contextual understanding from the VectorDB, invokes necessary tools, retrieves data from HealthcareDB, and provides a voice response to the user.

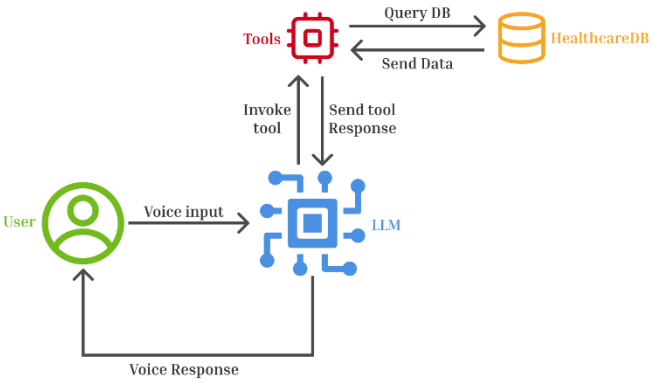

Fig -1: Architecture Diagram

3. Implementation

A. Audio Query Input

Users interact with the system via audio queries. The input is processed to detect the required entities (e.g., patient, doctor, medicine), which are then processed by the LLM.

B. Healthcare Database

The system includes a comprehensive healthcare database containing structured data related to patients, doctors, medicines, hospitals, and diagnoses.

C. Tools and LLM Interaction

The LLM invokes pre-configured tools to fetch specific information from the healthcare database. It orchestrates the retrieval of relevant data and formats the output, delivering it back to the user in audio form.

4. Experimental Results

The system was evaluated against traditional healthcare information systems using the following metrics: cost, ease of use, queries per minute, and time to retrieve information.

A. Cost

The cost analysis (Fig. 2) shows that the real-time RAG system is more expensive than traditional healthcare systems due to the intensive computational resources required by LLMs.

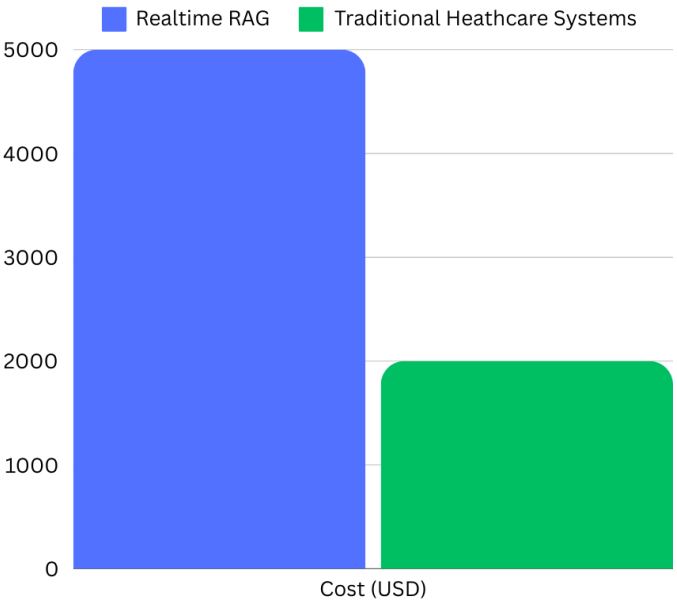

Fig -2: Cost

B. Ease of Use

As shown in Fig. 3, users rated the real-time RAG system higher in terms of ease of use compared to traditional systems. The simplicity of voice-based querying significantly enhanced the user experience

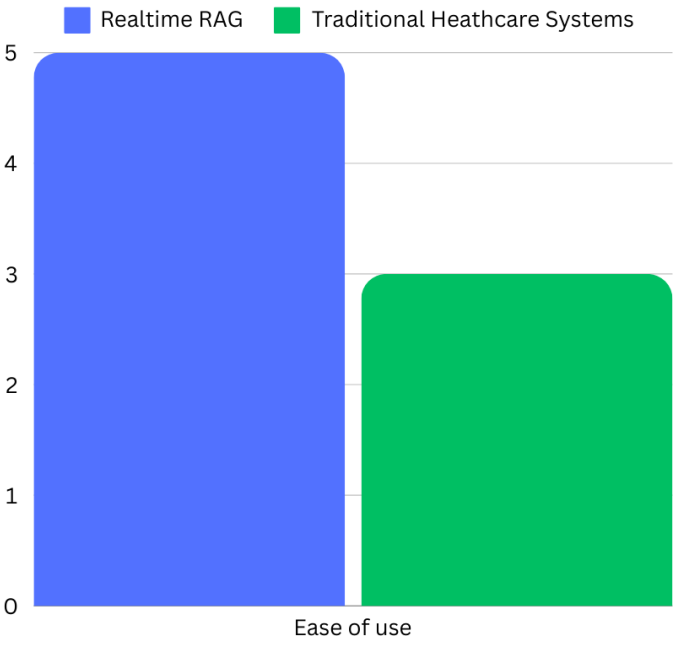

Fig -3: Ease of use

C. Queries per Minute

In Fig. 4, we observe that the real-time RAG system handles a greater number of queries per minute, making it highly efficient compared to traditional healthcare systems, which are often bottlenecked by manual data retrieval processes.

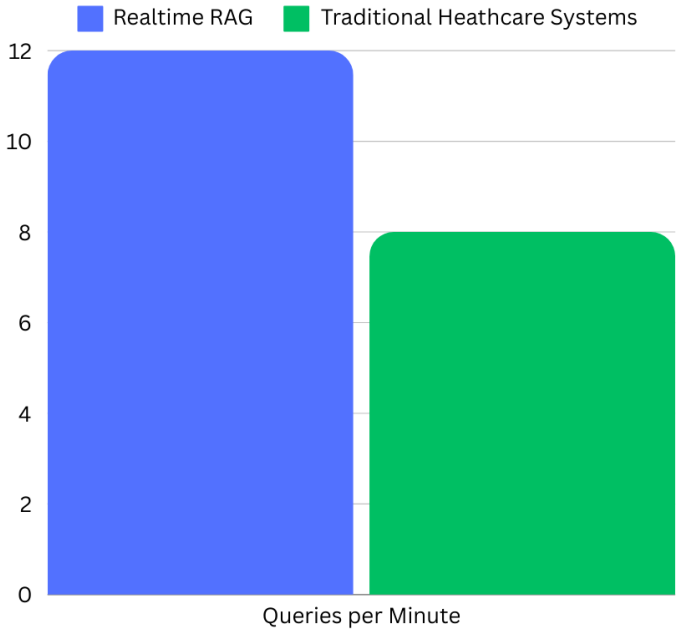

Fig -4: Queries per minute

D. Response Time

The time to retrieve information (Fig. 5) highlights that our LLM-based system drastically reduces response time from an average of 20 seconds in traditional systems to around 5 seconds.

5. Discussion

Our experiments demonstrate the superiority of real-time LLM-based systems in handling healthcare queries more efficiently and at a faster rate. The results suggest that while the system incurs higher computational costs, it significantly improves throughput and user satisfaction. The integration of vector databases allows for more accurate and context-rich responses.

6. CONCLUSIONS

This paper has presented a novel real-time audio-querydriven RAG system for healthcare. By leveraging the capabilities of LLMs and integrating them with structured healthcare databases, we have demonstrated substantial improvements over traditional systems. The findings suggest that LLMpowered systems could revolutionize how healthcare data is accessed and used, particularly in time-sensitive environments.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Savaridassan P, for his invaluable guidance, insights, and support throughout the course of this research. His expertise and mentorship were instrumental in the successful completion of this project. The authors would also like to thank SRM Institute of Science and Technology, and the Department of Networking and Communications, for providing the resources and academic environment necessary to conduct this study

REFERENCES

1. S. Zhao et al.,"Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely," arXiv, 2024. DOI: https://doi.org/10.48550/arXiv. 2409.14924.
2. W. Cheungpasitporn et al., "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," Medicina, vol. 60, no. 3, p. 445, 2024. DOI: https://doi.org/10.3390/medicina60030445.. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438
3. Y. Hoshi et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv, 2023. DOI: https://doi.org/10.48550/arXiv. 2312.10997
4. Shao et al., "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy," arXiv, 2023. DOI: https://doi.org/ 10.48550/arXiv.2305.15294.)