International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Enhancing large language models with a hybrid retrieval augmented generation system: A comparative analysis

S. A. Belhe, Parth Barse *, Dheeraj Chingunde, Rutuja Katkar and Vansh Koul

Department of Artificial Intelligence and Data Science, AISSMS IOIT Pune, Maharashtra, India.

International Journal of Science and Research Archive, 2025, 15(01), 1607-1612

Publication history: Received on 10 March 2025; revised on 26 April 2025; accepted on 28 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1170

Abstract

With the increasing reliance on cloud-based AI services for NLP tasks, organizations are facing significant challenges in ensuring the privacy and security of their internal data. Stringent data privacy regulations like GDPR and HIPAA require organizations to safeguard sensitive information and prevent it from leaving their local infrastructure. This project proposes a solution to address these concerns by leveraging Retrieval-Augmented Generation (RAG) techniques, which combine transformer-based language models with document retrieval systems to generate accurate, contextually relevant responses while ensuring data remains within the organization's local environment.

By integrating an on-premise system for document retrieval and response generation, we ensure that sensitive information is never exposed to external cloud servers, helping organizations comply with privacy regulations. The entire system is implemented in Python, designed to be scalable, flexible, and seamlessly integrated into existing infrastructure, making it a practical solution for organizations seeking to utilize advanced AI capabilities without compromising data security. This approach not only enhances privacy but also enables organizations to harness the power of AI-driven NLP tasks safely and efficiently.

Keywords: Retrieval-Augmented Generation (RAG); Large Language Models (Llms); Natural Language Processing (NLP); Transformer Models; Data Privacy; On-Premise Systems; Document Retrieval; Vector-Based Search; Contextual Understanding; GDPR Compliance; HIPAA Compliance; FAISS; Hugging Face Transformers; Secure Data Querying

1. Introduction

The unprecedented rise of remote work and digitalization has resulted in the exponential rise in organizational data. With the rise, data privacy concerns have also risen exponentially, particularly if cloud-based AI tools with NLP [5] support are utilized in the organization. In the majority of systems, data will need to pass through third-party servers, which raises the likelihood of a data breach and non-compliance with the law.

To address the aforementioned challenges, we introduce a prototype of on-premises Retrieval-Augmented Generation with the capacity to safely query organizational information. Our method involves the element of having all processing of data conducted on local servers, and hence organizations completely own data and privacy as they make use of AI capability. The architecture we introduce has the capacity to strictly adhere to both legal requirements and stringent confidentiality requirements that apply to various industry segments, thus making it highly robust to the prevailing data-intensive activities.

* Corresponding author: Parth Barse

Figure 1 YARVIS Logo

2. Literature review

The study, released in 2024, was conducted by researchers Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia.[1] The researchers developed a Multi-Model Retrieval Augmented Generation (MuRAG) pipeline that retrieves textual information, renders document pages into images to maintain text-image relationships, and integrates these with summaries to create a single knowledge base. The pipeline was evaluated on different question-answering datasets and compared to other multimodal large language models (LLMs). The findings indicated that it effectively managed intricate text-image relationships, resulting in better retrieval and generation performance across a range of datasets.

The paper published in 2024 [2], this paper was authored by researchers from Tongji University and Fudan University in China. The paper is an in-depth examination of different paradigms of Retrieval-Augmented Generation (RAG) of large language models, i.e., Naive, Advanced, and Modular RAG models. The study delves into the elementary techniques of retrieval, generation, and augmentation in these paradigms and tests the efficiency of these paradigms in reducing the limitations of large language models. One of the key conclusions of the study identifies that RAG significantly enhances the performance of large language models through higher accuracy, reliability, and traceability through the incorporation of external knowledge sources, thus significantly reducing hallucinations and stale responses.

The study was published in 2022 by Ahn, Lee, Shim, and Park. [3] The authors proposed a retrieval-augmented model with a topic-aware dual-matching reranker and a data-weighting approach specially designed for knowledge-grounded dialogues. The model was designed to enhance the relevance and diversity of the generated responses. The findings show that this approach achieved state-of-the-art performance in conversational AI, particularly in response generation that is both knowledge-grounded and contextually appropriate.

Authored by Asmitha M, Aashritha Danda, Hemanth Bysani, Rimjhim Padam Singh, and Sneha Kanchan, the paper was released in 2024.[4] Comparative analysis of various models of summarization, such as BERT, GPT, Pegasus, and Hugging Face models, was done in the study. The models were analyzed using ROUGE metrics to assess their summarization accuracy. The research identified that the Hugging Face model mbart-large-cc25, trained on the CNN/DailyMail dataset, had the best accuracy, particularly in the domain of abstractive summarization tasks.

This research was released in 2022 by Paras Nath Singh and Sagarika Behera. It discusses the use of transformer models such as BERT, GPT, and BART in various natural language processing tasks. [5] These tasks were performed using Python libraries such as PyTorch, TensorFlow, and the Hugging Face API and included POS tagging, bigram analysis, neural machine translation, summarization, paraphrasing, and sentiment analysis. The results showed the supremacy of transformer models over the conventional LSTM and RNN models, performing with incredible accuracy rates of 94% to 98% in multilingual sentiment analysis and performing well in paraphrasing and summarization tasks. As these tasks were performed using Python libraries such as PyTorch, TensorFlow.

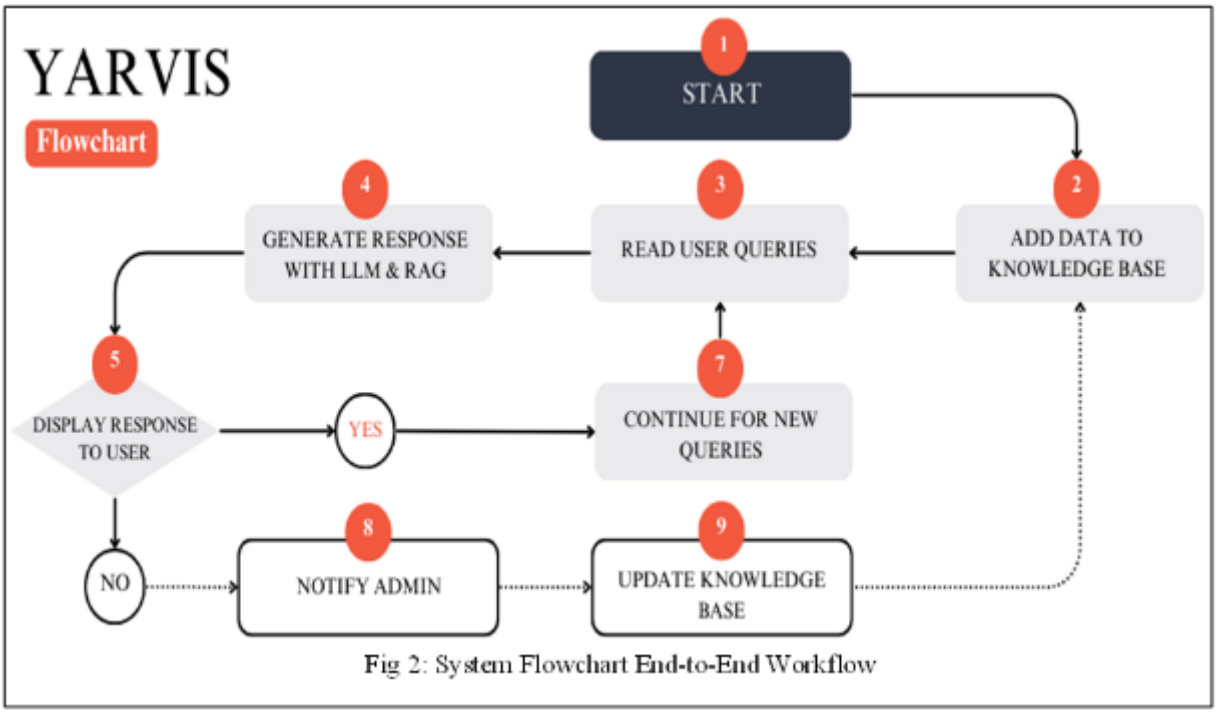

Figure 2 System Flowchart End–To– End Workflow

3. Methodology

This work focuses on designing a secure, in-house Retrieval-Augmented Generation (RAG) system for enabling organizations to query internal information using natural language questions with data privacy and regulatory compliance. The research process is structured as a series of significant steps, including data preparation, document retrieval, response generation, and deployment, each of which contributes to the overall robustness of the proposed system.

The first step is data preparation and collection. The data repositories are heterogeneous internal stores, including relational databases like PostgreSQL and MySQL, flat file formats like CSV, Excel, and JSON, and document stores with PDFs and Word documents. Data preprocessing is necessary to make the data consistent and usable, and this is achieved by filling gaps in missing values, correcting inconsistencies, and eliminating duplicates using pandas. Various text normalization methods are employed to attain uniformity in the text, including lower case conversion, removal of stopwords, and stemming. Tokenization is performed using libraries like spaCy or NLTK, and embeddings are created using pre-trained models like Sentence-BERT, enabling conversion of text data to numerical vectors for efficient search operations. The second step involves document retrieval, which is necessary for retrieving relevant documents that hold the contextual information required to produce correct responses.

Documents are indexed using Elasticsearch to enable fast and scalable full-text search functionality. In addition, additional metadata like tags and timestamps is added to enhance the relevance of the search results. For semantic search support, data is stored in the vector form by using dense embeddings. High-speed similarity searches for large collections of data are enabled using FAISS, or Facebook AI Similarity Search. Queries from the user are converted to embeddings, and based on the cosine similarity measure, the documents that best match are ranked to enable precise information retrieval. Upon identification of the relevant documents, the system shifts its focus towards response generation by employing advanced generative models like GPT-2, BART, or T5. The models are fine-tuned on domain data to enhance relevance and contextual correctness, and they are imported from Hugging Face's transformers library[4][5].

The system has two phases in the process: first, the retriever uses either Elasticsearch or FAISS to retrieve the top-k documents; the documents are then passed through the transformer model, which generates rich answers from the provided context. Beam search, top-k sampling, and temperature scaling are some of the methods used to optimize the accuracy and quality of the generated responses. Multi-turn dialogue is also enabled by the system through the preservation of interaction history, thus enabling contextually aware follow-up question answers. Strict security

practices are followed because of the sensitive nature of organizational data, providing confidentiality and integrity. Data is protected by encryption at rest and in transit using AES-256 encryption, and network traffic is protected by SSL/TLS protocols. Role-based access control provides assurance that sensitive datasets are accessed only by authorized users, and extensive logging captures user activity, query history, and system performance metrics.

These logs are regularly checked for security audits and regulatory compliance. The system is built for easy deployment and scalability. It comes packaged with Docker containers for uniformity in a different environment, and Docker Compose manages multi-container configurations with the RAG engine, database, and search engine. For organizations that scale slightly, the system can deploy on Kubernetes clusters; this supports horizontal scaling in instances of higher user demand. Performance is enhanced through mechanisms that cache results of commonly accessed queries and acceleration through the utilization of a GPU for quicker inference of models when dealing with large datasets.

This systematic approach makes sure the RAG system is robust, effective, and secure in the sense that it provides organizations with a guaranteed query technology to investigate their internal data without violating data confidentiality and being fully in sync with regulatory requirements. This technical report provides an overview of the whole workflow-from preprocessing data to producing the query response - through the system flowchart.

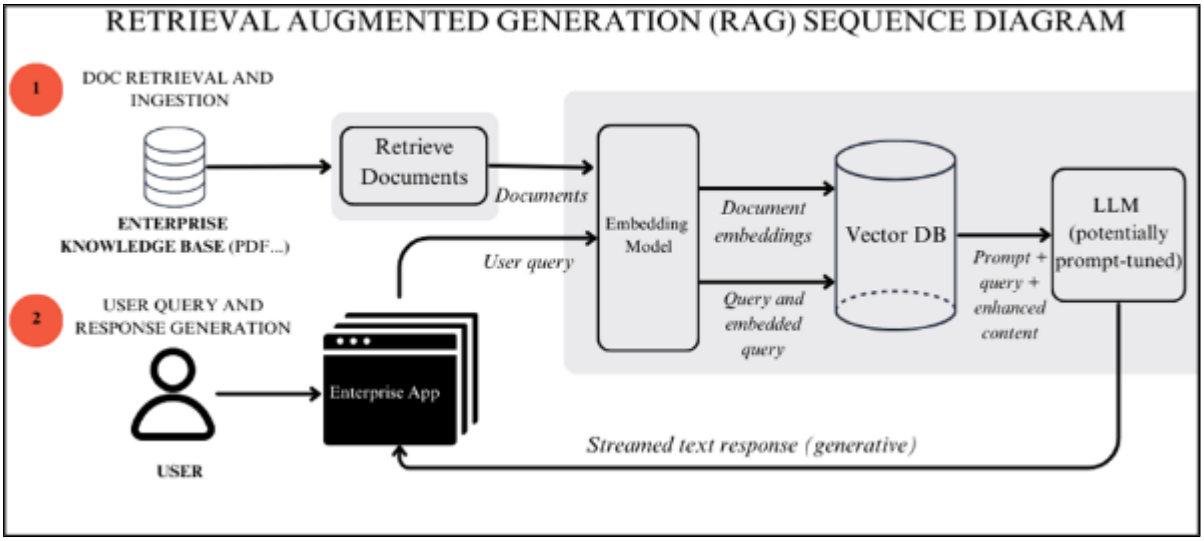

Figure 3 System Architecture

4. Project architecture

The architectural design of the system has been made such that the sensitive company information is kept confidential but can be accessed through advanced querying of data by means of Retrieval Augmented Generation (RAG) [3]. In essence, the system is composed of a standalone binary that is easy to deploy and install on local servers. This is so that all processing of data and artificial intelligence requests happens locally, hence eliminating the possibility of transferring sensitive data across the internet to third-party AI vendors.

The design is also modular, composed of several essential components: an intake layer to take in and process data, a retrieval module to structure the company's data in a query-efficient manner, and a generation module powered by cutting-edge language models. Each of these works together as one to allow companies to produce useful answers to queries based on their own data.

The architecture also has a secure interface for communication with the retrieval engine, via which users with permission can submit queries while maintaining control of their data. Since the entire solution is autonomous, organizations are not dependent on third-party APIs, which not only helps to protect data privacy but also provides full control of the infrastructure. Overall, the architecture is scalable, secure, and simple to deploy, and hence is perfect for enterprises that require confidentiality as well as advanced AI-driven insights.

5. Result

This section explains the findings of the study, such as an explanation of how the proposed Retrieval-Augmented Generation (RAG) system addresses the limitations of existing systems. The evaluation includes accuracy, performance, usability, scalability, privacy, and other system characteristics.

5.1. System accuracy and effectiveness

The RAG system successfully recovered relevant documents and produced context-aware responses. In contrast to baseline systems prone to producing irrelevant or incorrect responses, the application of fine-tuned transformer models and hybrid re-ranking enhanced coherence, particularly with intricate queries. It also supported multiple content types such as text, tables, and mixed documents.

5.2. Query performance and usability

By applying Elasticsearch and FAISS, the system provided quick and scalable response on datasets of different sizes. The user-friendly interface provides multi-turn dialogue support, providing enhanced interaction as compared to systems having no context continuity or user-friendly interface.

5.3. Data privacy and security

By being entirely on-premises, the system provides full control of the data, satisfying regulation compliance such as GDPR and HIPAA. Security is also enhanced by AES-256 encryption, SSL/TLS protocols, and role-based access control, securing sensitive organizational data.

5.4. Scalability, deployment, and administrative control

Its modular design makes it easy to deploy with Docker and Kubernetes, offering effortless scalability under heavy loads. Administrative controls enable efficient data management, making it enterprise-ready.

5.5. Future prospects

While the system is optimal when handling semi-structured and structured data, future releases will focus on handling unstructured data. Few-shot learning, real-time updates, and audio I/O will all increase flexibility and expand the applications.

5.6. Notification system

There is an in-built notification system to alert the admin about missing information when information is requested, facilitating real-time insertion of data and minimizing user frustration. This cutting-edge feature is absent in most existing systems.

5.7. Administrative interface

The secure admin interface allows only authorized personnel to add, update, or delete information without needing technical expertise. Role-based access control guarantees control and prevents illegal alteration, facilitating real-world practicality.

Table 1 Features

Feature	Existing Systems	Proposed System
Data Modalities	Limited to text-image; some support for plain text (e.g., MuRAG)	Supports text and enhances document processing
Re-Ranking Mechanism	Basic dual-matching in conversational RAG	Hybrid re-ranking with dual-matching, topic relevance, and context continuity
Retrieval Efficiency	Standard retrieval with limited speed for large data	Vector-based database integration, improving response time and scalability
Multi-Turn Conversation Handling	Limited memory, affects context in lengthy exchanges	Stores user interaction history, enabling better continuity and relevance

Summarization and Abstractive Generation	Primarily extractive models (e.g., Hugging Face)	Combination of extractive and abstractive techniques for comprehensive summaries
Response Accuracy and Relevance	Moderate accuracy with tendency to hallucinate in complex queries	Improved accuracy due to refined re-ranking and contextual understanding
User Interface	Basic or absent	Integrated UI with multi-turn history tracking and enhanced usability
Audio Output	Not available	Can be integrated in future for accessibility and voice-based interaction
Document Retrieval	Basic or limited	Advanced document retrieval using vector search (FAISS) and semantic similarity
Notification System	Not available	If data not available, system alerts admin to update the database
Data Security	Often cloud-based, less control over privacy	Data is stored and processed locally, ensuring full control and compliance
Admin Interface	Often missing or limited	Allows admin to add or update data directly through a secure local interface

6. Conclusion

This project provides a secure, on-premise solution that leverages Retrieval-Augmented Generation, enabling organizations to ask questions about their internal data using natural language. By processing all data within the organization's own infrastructure, we address the critical concern of data privacy while still offering powerful AI capabilities. Future initiatives will aim to incorporate real-time data sources and utilize advanced methods like few-shot learning to enhance the system's flexibility and accuracy.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] P. Joshi, A. Gupta, P. Kumar and M. Sisodia, "Robust Multi Model RAG Pipeline For Documents Containing Text, Table & Images," 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2024, pp. 993-999.

[2] S. Vakayil, D. S. Juliet, A. J and S. Vakayil, "RAG Based LLM Chatbot Using Llama-2," 2024 7th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India, 2024, pp. 1-5.

[3] Y. Ahn, S. -G. Lee, J. Shim and J. Park, "Retrieval Augmented Response Generation for Knowledge-Grounded Conversation in the Wild," in IEEE Access, vol. 10, pp. 131374- 131385, 2022.

[4] P. N. Singh and S. Behera, "The Transformers' Ability to Implement for Solving Intricacies of Language Processing," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-7.

[5] M, A. Danda, H. Bysani, R. P. Singh and S. Kanchan, "Automation of Text Summarization Using Hugging Face NLP," 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India, 2024, pp. 1-7.

[6] M. Maryamah, M. M. Irfani, E. B. Tri Raharjo, N. A. Rahmi, M. Ghani and I. K. Raharjana, "Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access," 2024 16th International Conference on Knowledge and Smart Technology (KST), Krabi, Thailand, 2024, pp. 259-264.

[7] H. K. Chaubey, G. Tripathi, R. Ranjan and S. k. Gopalaiyengar, "Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development," 2024 International Conference on Future Technologies for Smart Society (ICFTSS), Kuala Lumpur, Malaysia, 2024, pp. 169-172

[8] A. Sar et al., "PDF-Based Chatbot Development Using LLAMA2 and LangChain: Training and Deployment for Document Interaction," 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, 2024, pp. 1-6