International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Advanced Multi-Source RAG for Enterprise Knowledge Base

Prof. Mrunali Makwana[1], Rudra Dhore[2], Mithali Suryawanshi[3], Anuj Nalawade[4]
Professor, Department of Computer Science, Pune, Maharashtra
B. Tech, Student[1], Department of Computer Science, rudradhore85@gmail.com
B. Tech, Student[2], Department of Computer Science, suryawanshimithali@gmail.com
B. Tech, Student[3], Department of Computer Science, anujnalawade24@gmail.com
Ajeenkya DY Patil University, Lohegaon, Airport Rd, Charholi Budruk, Pune, Maharashtra

ABSTRACT:

Enterprise data has taken off at a high rate in various formats and this has posed a big challenge in effective knowledge retrieval in organizations. Even though such a tool as ChatGPT can adequately respond to general questions, it is not provided with access to organizational data, including internal documents, policies, and reports. This makes organizations need intelligent systems that are able to come up with responses that are solely based on the verified and confidential information of the organizations. In order to overcome these shortcomings, Retrieval-Augmented Generation (RAG) uses information retrieval methods alongside large language models to generate correct and context-relevant responses using enterprise knowledge. Such technologies as FAISS and Pinecone make this process even more efficient as they provide an opportunity to search the large collections of documents by means of semantic searches. The study offers a framework Advanced Multi-Source RAG framework in an enterprise-specific form. The system is built to combine various sources of data such as the PDF documents, the CSV files and the web information with the help of an interactive interface that is created using the Streamlit. It uses the LangChain and LlamaIndex to process, index and the retrieved documents and allows generation of accurate, context aware and reliable responses to any given body of knowledge held privately by an enterprise. The suggested system enhances restoring accuracy and minimizes hallucination as well as decision-making efficiency.

***Keywords*:** Retrieval-Augmented Generation (RAG), Enterprise Knowledge Base, Semantic Search, Multi-Source Data Integration, Large Language Models.

1.INTRODUCTION

The fast expansion of enterprise information in many forms such as unstructured documents, structured databases, and web-based material, has compounded the complexity of knowledge management in the organizations. To enhance productivity and informed decision-making, efficient retrieval of the relevant information in these heterogeneous sources is necessary.

The conventional knowledge management systems have been largely dependent on search process based on key word search mechanisms, which are usually ineffective in obtaining semantic meaning and context relevance of the queries entered by the user. Consequently, users have to search in several platforms to find pertinent information manually, which translates to inefficiencies and delays.

The recent progress of Large Language Models (LLMs) has greatly enhanced the generation and understanding of information in natural language, allowing more smart interaction between the user and information systems. These models however work largely on previously established knowledge and in most cases do not have direct access to real time enterprise data which could lead to provision of inaccurate or unauthenticated responses.

In order to overcome it, Retrieval-Augmented Generation (RAG) combines information retrieval

methods with language generation models. RAG enhances factual accuracy and contextual relevance in generated outputs since it uses external knowledge sources to obtain the relevant information before the generation of the responses [1].

The new technologies have also improved the system of RAG. Those libraries including FAISS are used to facilitate similarity search of large-scale vectors data efficiently [2], and databases of vectors including Pinecone can be used to store and retrieve embeddings economically [3]. The implementation of large language models in conjunction with external sources of knowledge is supported by frameworks like LangChain and LlamaIndex and can be applied to advanced retrieval-based applications [4], [5]. Also, web frameworks such as Streamlit can be used to create interactive interfaces to deploy AI-driven knowledge systems [6].

This paper presents an Advanced Multi-Source Retrieval-Augmented Generation system of enterprise knowledge base systems. The system proposed combines heterogeneous information sources (i.e. PDF files, CSV files, and web-based information) into a single retrieval pipeline to enhance the knowledge accessibility, retrieval efficiency and decision-making in an enterprise setting.

1.1.The Study's Background

The fast development of enterprise information in various forms has posed a big challenge to efficient knowledge recovery in organizations. Contemporary businesses create and archive information in different forms which include unstructured documents, structured databases and web-based information. The conventional knowledge management systems mostly depend on the search through keywords-based methods that are not always effective to realize the task of dealing with the heterogeneous sources of data and offer context-related answers. The new developments in artificial intelligence have come with Retrieval-Augmented Generation (RAG), which is a technique that integrates information retrieval techniques with large language models to produce correct and contextually neutral responses depending on the retrieved data of the enterprise. This will enhance the quality of outputs and their relevance as external sources of knowledge are utilized and overcome the weaknesses of pure language models. Furthermore, today, semantic search technologies make retrieving the information of big datasets efficient due to the ability to present the textual information as a vector embedding. This enables systems to find pertinent information according to semantic similarity as opposed to straight key word matches. Here, this study suggests an Advanced Multi- Source Retrieval- Augmented Generation architecture that brings together various enterprise data sources, such as documents, structured databases, and web-based materials. The system will be able to provide the right and contextual responses using an interactive interface that will enhance the access to knowledge and effective decision-making in organizations.

1.2. The Research is motivated by the following reason.

The speed with which enterprise data is expanding in more than one format has rendered retrieving pertinent information more challenging to the organizations. The search systems that are traditional and rely on key words do not offer proper and context related results. Even though such sophisticated models as ChatGPT may respond to general searches, they can not access confidential enterprise information. The given limitation underlines the necessity of systems that have the capacity to produce responses using internal sources of knowledge. Thus, the research is driven by the necessity to formulate a multi-source Retrieval-Augmented Generation (RAG) model that can allow one to retrieve the knowledge accurate, context-sensitive and efficient in an enterprise setting

2.PROBLEM DESCRIPTION

The high increase in volume of enterprise data in various forms such as documents, reports, spread sheets, web resources etc., has proved to become a significant challenge to many organizations on effective knowledge retrieval. The traditional knowledge management systems largely utilize the use of key-based search mechanism, which has lesser capabilities to support heterogeneous data sources, and in many cases it creates lesser contextual and semantically appropriate answers [9]. Consequently, employees have to waste a lot of time on the search of any information among numerous materials, which decreases productivity and operation inefficiencies.

Recent innovations in transformer-based language models have enabled the generation and natural language understanding capabilities to be much improved [7], [8]. Nevertheless, Large Language Models (LLMs) are generally based on the existing knowledge and do not have direct access to enterprise-specific data, as a result of which they cannot be effective in an organizational setting where the answers are to be obtained based on internal documents and proprietary information.

The Retrieval-augmented generation (RAG) has given rise to a powerful method that integrates information retrieval with language generating models to come up with more factual and in-context responses [1], [11]. Also, FAISS and other types of vector search technology allow to effectively search similarities on a large scale when it comes to semantic retrieval tasks, which makes it possible to create intelligent knowledge retrieval systems [2].

Thus, the proposed research is an Advanced Multi-Source Retrieval-Augmented Generation framework that incorporates several enterprise data streams and retrieval strategies to produce quality and context-sensitive responses. The suggested system is implemented with the help of an interactive interface created with the help of the Streamlit to enhance the knowledge base and make timely and effective decisions in organizations.

2.1.Importance of the Research

The study is relevant in companies which need effective and smart ways of handling and accessing enterprise knowledge. As the volume of data controlled by organizations in various formats grows at a very high pace, employees are frequently challenged by the inability to find the required and reliable information among massive sets of documents. The result of this challenge is the consumption of more time, less productivity, and inefficiency in the decision-making processes. The suggested framework will enhance the knowledge retrieval process by using modern technologies, including FAISS, Pinecone, LangChain, and LlamaIndex, to provide the ability to perform semantic search and generate responses in a specific context. The system is also unbiased as it is able to access information in its entirety, is reliable and based on organizational knowledge, since the system can access information that is directly stored in enterprise data sources. Moreover, the suggested system will improve the availability of knowledge and preserve data secrecy using internal documents instead of referring to the outside information sources. The method saves the manual searching process involved, increased efficiency in operations, and offered a scalable solution to the contemporary knowledge management systems in enterprises.

2.2. Research Objectives

The primary aim of this study is to come up with a sophisticated multi-source Retrieval-Augmented Generation (RAG) system of effective enterprise knowledge retrieval. The following are the specific objectives:

To bring together various information streams like documents, CSV files and web content together into one platform. Unlike other tools, this approach will use semantic search with the help of the FAISS and Pinecone vector databases. International students studying in a British university, you have been asked to design a RAG based-pipeline with the help of frameworks like LangChain and LlamaIndex. Create context-

sensitive responses to enterprise data. A comprehensive user interface based on Streamlit.

3.LITERATURE REVIEW

The high rate of expansion of data of the enterprise has heightened the demand of intelligent information retrieval systems. The conventional knowledge management systems are largely dependent on the use of the key word based search methods and this method often fails to bring out any semantic meaning and contextual relationship in the large and heterogeneous data sets. In order to overcome these shortcomings, scholars have come up with semantic search methods that enhance the usefulness and quality of retrieved data [9].One of the significant steps in the field was made by Patrick Lewis et al. (2020) with the help of the Retrieval-Augmented Generation (RAG) framework based on the combination of information retrieval with the large language model to produce responses based on the external sources of knowledge. This is better than standalone language models in terms of the response accuracy and minimizing hallucination [1], [11]. Pinecone and FAISS are also commonly used as examples of vector search technologies to assist in semantic retrieval. These systems turn textual data into a form of a vector and provide the possibility to search by similarity, which means that information can be picked by semantic meaning and not by words themselves [2], [3]. Moreover, the current development systems like LangChain and LlamaIndex also offer document ingestion, indexing and query processing tools that make it easier to build a RAG-based application and connect with various data streams [4], [5]. Retrieval-augmented systems have been shown to be effective in knowledge management of enterprises, question-answering systems and smart customer support. The systems are more accurate, reliable, and context-sensitive in their responses than the conventional information retrieval methods through the combination of retrieval methods and language models.

3.1. Research Gap

Despite the large improvements within the Retrieval-Augmented Generation (RAG) systems, there are a number of shortcomings that exist in the current studies and practice applications. Most of the existing literature is concerned with single-source data retrieval that restricts the capability of retrieving holistic and multi-faceted information of an enterprise data which can include documents, databases, and web information [1]. Though frameworks like FAISS and Pinecone are helpful in performing semantic search with the usage of the vectors, they are commonly used separately and have no connection between different data types, which limits their performance in the heterogeneous enterprise-wide context [2], [3]. Besides, LangChain and LlamaIndex are frameworks that enable building RAG-based applications through document processing and indexing as well as querying tools. Nevertheless, most of the current implementations are not comprehensive in aspects of multi-source data integration and consolidated retrieval systems in regard to enterprise knowledge systems [4], [5].Besides, some of the existing solutions do not have an interactive and user friendly interface that allows non-technical users to query the enterprise knowledge bases easily. The application frameworks like Streamlit today offer possibilities of creating accessible interfaces but do not include the development of integration with the advanced RAG pipelines, which is also not present in existing research [6]. Thus, there are no integrated models that can unify a variety of sources of enterprise data and effective methods of semantic retrieval to obtain relevant and reliable answers. This study will prevent these shortcomings by introducing an Advanced Multi-Source Retrieval-Augmented Generation framework that will have an interactive interface to enhance the ease of access to enterprise knowledge and efficiency of decision making.

4.METHODOLOGY

The suggested system will follow the Retrieval-Augmented Generation (RAG) concept and

ensure the effective and context-related knowledge retrieval of various sources of enterprise data. RAG uses large language models with information retrieval methods to enhance relevancy and accuracy of the responses that are generated [1]. The general process of the system is the collection of the data, preprocessing, and the generation of embeddings, storage of the vectors, retrieval, and generation of the response.First, there is the collection of data which is heterogeneous in terms of PDF documents, text files, spreadsheets and web-based resources. As the information will be available in various forms, it will be pre-processed in order to achieve uniformity and quality. The step includes cleaning of data, elimination of noise, the processing of missing data and transformation of the data into a textual format which can be easily processed.

The textual data is further broken down into smaller unit of data after preprocessing to enhance the efficiency of retrieval and matching of semantics. These fragments are then translated into the form of vectors upon embedding models that obtain the semantic meaning and contextual relationship of the text [2].Embeddings that are generated are stored in the vector database (FAISS and Pinecone) allowing the efficient search on similarity and quick retrieval of relevant documents in case of large datasets [3], [4]. A query is also run through an embedding when the user makes a query and compared to the stored vectors to determine the most relevant content.

The retrieved data is then fed to a large language model via a Retrieval-Augmented Generation pipeline which is a framework implemented in LangChain and in LlamaIndex. These models help in coordinating communication between the retrieval system and the language model in a bid to produce correct and context-sensitive responses [5], [6].Lastly, the response generated is provided to the user in the form of an interactive web interface created in Streamlit that enables the users to query enterprise knowledge bases easily and access real-time responses [7].

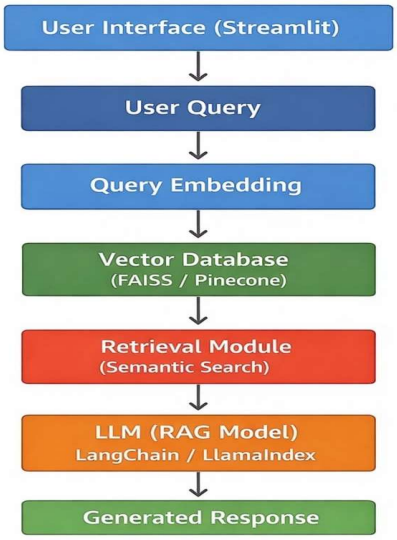

Figure 4.1: Architecture of the Proposed Advanced Multi-Source Retrieval-Augmented Generation (RAG) System

4.1.System Design

The proposed system will be implemented with the help of a modular and scalable architecture that will effectively process various **general enterprise data streams and give relevant and context-aware answers. Figure 1** depicts the architecture of the system at large. The system comprises of a number of modules that are linked together and which are coordinating to carry out **data processing, retrieval** as well as generate responses**.**

1. Data Ingestion Module

The data ingestion module is the one that gathers data of various types made of heterogeneous data i.e. **PDF files, text files, spread sheets, and web-based data**. This module will make sure that different enterprise data formats can be processed by the system and incorporated into a single line of further processing.

2. Data Preprocessing Module

The data preprocessing module is the part that has the operation of cleaning and transformation of the data in order to enhance the quality and consistency of the data. This involves activities like noise **reduction, management of missing data, text normalization** and the transformation of raw data into standardized textual data that is capable of undergoing further processing**.**

3. Text Chunking Module

The text chunking module is used to **divide large documents into small segments**. Chunking enhances the retrieval performance and allows a better matching of the user queries and the document segments that are deemed relevant. It is also important to note that it can ensure that the system will be capable of processing large amounts of text.

5. Storage Module and Embeddings

Embedding models are used to convert the processed text fragments into the form of vector embeddings and extract the semantic meaning of the text. The embeddings are placed in the **vectors databases, including FAISS and Pinecone**, which can be efficiently searched and retrieved in position of similarity.

5. Retrieval Module

On the query input, the retrieval module turns the query into **a query-representation (a vector embedding),** and actively searches the vector database with the spirit of semantic similarity. The best sections of the documents are identified and sent to the generation module where responses are generated.

6. Generation Module

The generation module is used to produce context-specific responses by using **a large language model (LLM)** to generate based on the information retrieved. The way the interaction between the retrieval system and the language model is organized is done using **frameworks like LangChain and LlamaIndex,** which guarantee a system that provides relevant and accurate answers.

7. User Interface Module

The module developed with the help of **Streamlit** is the user interface module, which can provide the platform where the user can input the queries to get the responses in real time. The interface is made easy to use and utilize, as well as efficient in interaction with the system.

The workflow of the proposed system is illustrated in Table 4.1.

Table 4.1: Workflow of the Proposed RAG-Based System

Step	Model	Process
1	User Interface	Query Input
2	Embedding Module	Query Embedding
3	Vector Database	Similarity Search
4	Retrieval Module	Data Retrieval
5	Processing Module	Context Preparation
6	Generation Module	Response Generation
7	User Interface	Output Display

4.2.Tools and Technologies

The above system will combine both the high-technology and technology to guarantee the delivery of the effective data processing, semantic search, and interactivity in the system of enterprise knowledge management.

1.Programming Language

•Python:

Python language is implemented as the key programming language because it contains a variety of libraries and structures which are useful when it comes to processing of the data, machine learning and natural languages. It is both scalable and flexible and can be used to run complex Retrieval- Augmented Generation (RAG) systems.

2.Vector Databases

•FAISS:

FAISS is an open-source library it is concerned with providing an efficient similarity search and group of high-dimensional vectors information. It allows one to search the information of interest in a quick fashion through the means of semantic similarity.

•Pinecone:

Pinecone is a version data storing database, an autoscaled and similarity search as well as a real-time one. It is employed to efficiently query and manage big embedding data.

3. RAG Implementation Structures.

•LangChain:

The RAG pipeline which consists of immediate processing, chaining, and integrating retrieval and language models is developed and run through LangChain.

•LlamaIndex:

LlamaIndex has been applied as one of the effective tools of indexing and retrieving information in a manner that it facilitates free flow of information between the organised and unorganised information.

4. User Interface

•Streamlit:

Streamlit is used to create interactive web interface in order to enable users to make queries and visualize the response generated in real time.

5. Embedding Models

•Pre-trained Embedding Models:

Textual information is coded into dense vectors representations (embeddings) using the models. This enables semantic search on which it deciphers the meaning of the surrounding text and not matching of key words.

6. Data Sources Multi-Format Data Support:

The system will absorb a lot of different information such as PDF, text, CSV databases, web-based materials. This will bring in flexibility and adaptability as far as various data sources in an enterprise are concerned**.**

"Table 4.2 presents the functional modules of the proposed system."

Step	Process	Description
1	Data Collection	Gather multi-source data
2	Preprocessing	Clean and normalize text
3	Chunking	Divide into smaller parts
4	Embedding	Convert text into vectors
5	Storage	Store in vector database
6	Query Embedding	Convert user query
7	| Retrieval	Fetch relevant data
8	Response Generation	Generate final output

Table 4.4: Data Processing and Retrieval Steps

5.CONCLUSION

This work presented a framework of Advanced Multi-Source Retrieval-Augmented Generation (RAG) and was geared towards improving the knowledge retrieval within an enterprise. The proposed system will integrate different data and will use semantic search systems to provide effective and correct solutions. The system will be applicable in retrieving relevant data not on the basis of a matching keyword to a query but by meaning based on the LangChain and LlamaIndex system and the use of the vector databases such as FAISS and Pinecone. An easy to use interface that is incorporated through Streamlit also makes it more usable and accessible.The targeted framework will significantly help in saving time that is spent in accessing information, enhance accuracy of response and the aspect of decision making in organizations. All in all, the system shows the potential of an integration of retrieval mechanisms and language models to solve the dilemmas in enterprise knowledge management.

5.1. Limitations and Future Prediction.

Even though the proposed one is an efficient multi-source Retrieval-Augmented Generation (RAG) framework, it has certain limitations. The system is also not successful in utilizing the multimodal types and can only receive and use text based data only and not multimedia applications such as images, audio or videos. It also works in terms of quality and constitution of input data and this can have an influence on the accuracy of retrieval. Moreover, a domain specific knowledge may not be always adequately represented in the application of the pre-trained embedding models further fine-tuning.

Even though some other databases like FAISS and Pinecone can be employed as vectors database and enable the search process to be efficient, the scalability and performance would decrease as the amount of data increases. In addition, the data of the live enterprise systems are not integrated in real time and the advanced security and privacy systems are not fully enacted.

To overcome these shortcomings, it is possible to improve on several aspects to work in the future. Another possibility to improve the system is with the assistance of advanced large language models to improve the perception and reaction to the situation. It will be closer to dynamic settings, as

the incorporation of real-time information of business is presented. With this, it is also scalable with the introduction of use of other sources of information like pictures, audio and video which enhances the functionality of the system.

5.2. Real World Implications

The suggested multi-source Retrieval-Augmented Generation (RAG) paradigm is applicable to the business with high life applications. It helps the organizations to extract the relevant information in the large and varied information sources appropriately to save time and resources to search it manually.The system is contextually and extensively responsive based on semantic searching and complex systems like LangChain and LlamaIndex, which is value-addition to the decision-making. The latter is the implementation of the vector databases such as FAISS and pinecone that provides the efficient and scalable search to the massive enterprise applications.Such a structure can be used in the customer care, knowledge management within the organization, and in business analytics that can optimize productivity, minimize the operations and performance costs of the overall organization.

REFRENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal,
H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela,
"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,"
in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.
[2] J. Johnson, M. Douze, and H. Jégou,
"FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors,"
Facebook AI Research, 2017.
[3] Pinecone Systems Inc.,
"Pinecone: A Vector Database for Machine Learning Applications," 2021.
[Online]. Available: https://www.pinecone.io
[4] H. Chase,
"LangChain: Building Applications with Large Language Models," 2022.
[Online]. Available: https://www.langchain.com
[5] J. Liu,
"LlamaIndex: A Data Framework for LLM Applications," 2023.
[Online]. Available: https://www.llamaindex.ai
[6] Streamlit Inc.,
"Streamlit: A Python Framework for Building Data Applications," 2021.
[Online]. Available: https://streamlit.io
[7] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.
[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.
[9] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.
[10] O. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020.
[11] J. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. ICML, 2020.
[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," in Proc. ACL, 2020.

