International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AUTONOMOUS RESEARCH ASSISTANT

Mrs. Gunasundhari M, Dharanidhar A, Harish J, Krish Joyeaal A B, Kumaraguru G

Department of Artificial Intelligence and Data Science
Sri Shakthi Institute of Engineering and Technology

ABSTRACT:

This paper presents an Autonomous Research Assistant based on Retrieval-Augmented Generation (RAG) to provide intelligent, accurate, and real-time information retrieval for users. The system is designed to collect data from multiple sources, including web content and structured knowledge bases, enabling precise and context-aware responses. Users can interact with the system through a simple web interface by providing textual queries. When a query is submitted, the system converts it into vector embeddings and retrieves the most relevant information using a vector database powered by FAISS. The retrieved content is then processed using a large language model such as Gemini to generate a coherent and human-like response. The system also maintains a database to store user queries, responses, and logs for efficient data management and future analysis. This approach significantly improves the accuracy and reliability of generated responses by grounding them in real-time retrieved data, reducing hallucination commonly observed in standalone language models. The proposed system demonstrates effective performance and provides a scalable solution for intelligent research assistance across various domains.

Key Words: RAG, Autonomous Research Assistant, Semantic Search, Vector Database (FAISS), Large Language Models (Gemini), Natural Language Processing (NLP), Information Retrieval

1.INTRODUCTION

The exponential growth of digital information has made it increasingly challenging for users to efficiently search, analyze, and extract meaningful insights from vast amounts of data. Traditional search engines primarily provide a list of links, requiring users to manually navigate and interpret the information, which is often time-consuming and inefficient. Recent advancements in Artificial Intelligence, particularly in Natural Language Processing (NLP), have led to the development of large language models capable of generating human-like responses. However, these models often rely on pre-trained knowledge and may produce inaccurate or outdated information, a problem commonly referred to as hallucination. To overcome these limitations, this paper proposes an Autonomous Research Assistant based on Retrieval-Augmented Generation (RAG). The system combines real-time information retrieval with generative AI to deliver accurate and context-aware responses. Instead of relying solely on pre-trained knowledge, the system retrieves relevant data from external sources and incorporates it into the response generation process.

The proposed system utilizes a vector database powered by FAISS to perform efficient similarity search on embedded data. User queries are transformed into vector representations and matched with the most relevant information stored in the knowledge base. The retrieved content is then processed using a large language model such as Gemini to generate coherent and human-like responses.

2. BODY OF PAPER

The Autonomous Research Assistant is developed using a Retrieval-Augmented Generation (RAG) framework integrated with a Large Language Model (LLM) to provide accurate, context-aware, and real-time responses. The system begins by collecting data from multiple sources, including web content and structured knowledge repositories, which are processed to form a dynamic knowledge base. The collected data is segmented into smaller chunks and converted into vector embeddings, which are stored in a vector database powered by FAISS for efficient similarity search and retrieval.

When a user submits a query through the web interface, the system processes the input and converts it into embeddings. The retrieval module then performs a similarity search on the vector database to identify and extract the most relevant information corresponding to the query. This ensures that the generated response is grounded in actual retrieved data rather than relying solely on pre-trained knowledge, thereby improving accuracy and reducing hallucination.

The retrieved content is then passed as contextual input to the generative model, where a Large Language Model such as Gemini generates a coherent and human-like response. The output is presented to the user through a web interface in a structured and easily understandable format. Additionally, the system stores user queries, responses, and interaction logs in a database, enabling efficient data management and future analysis.

The integration of vector-based retrieval with generative models significantly enhances the performance of the system by ensuring relevant and context-aware responses. Compared to traditional search systems and standalone language models, the proposed approach provides improved precision, better contextual understanding, and a more efficient user experience. This architecture demonstrates the effectiveness of combining retrieval-based techniques with advanced language models for intelligent research and information assistance in real-world applications.

Table -1: System Comparison

Feature	Traditional Assistant	Smart RAG Assistant
Data Source	Static/Dynamic Data	Knowledge Base
Interaction Mode	Text Only	Web interface
Response Type	Predefined	Context-Aware
Accuracy	Moderate	High
Information Source	Manual / Limited	Sematic Vector Search

2.1 SYSTEM ARCHITECTURE

The system architecture of the Autonomous Research Assistant is designed based on a Retrieval-Augmented Generation (RAG) framework to ensure accurate and context-aware information retrieval. The process begins with collecting data from multiple sources, including web content and knowledge repositories, which are processed and divided into smaller segments. These segments are converted into vector embeddings and stored in a vector database powered by FAISS for efficient storage and similarity-based retrieval. When a user submits a query through the web interface, the system converts the input into embeddings and forwards it to the retrieval module. The retrieval module performs a similarity search in the vector database to identify the most relevant information corresponding to the query. This retrieved content serves as contextual input for the generative component, ensuring that the response is grounded in actual data.The contextual information is then passed to a Large Language Model, such as Gemini, which generates a coherent, human-like, and context-aware response. The final output is presented to the user through the interface in a structured and easily understandable format.

The architecture ensures efficient handling of large-scale data through vector-based storage, enabling fast and scalable similarity search. The integration of retrieval and generative components enhances response accuracy and reduces hallucination. This design makes the system adaptable to various domains and data sources without requiring significant modifications, thereby improving scalability and flexibility for real-world applications.

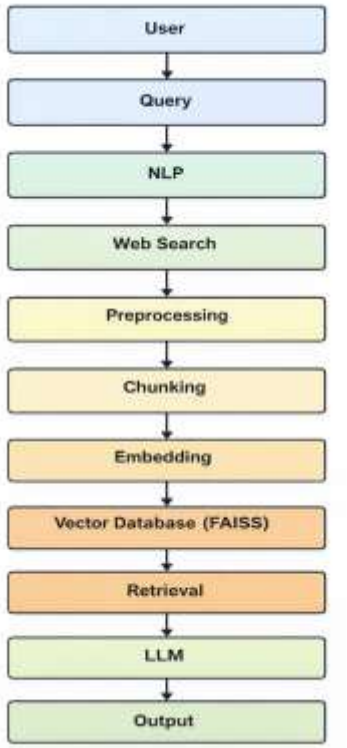

Fig - 1: System Architecture

2.2 WORKING OF THE SYSTEM

The system operates by first collecting data from various sources, including web content and knowledge repositories, and converting it into a searchable format. The collected data is preprocessed, segmented into smaller chunks, and transformed into vector embeddings. These embeddings are stored in a vector database powered by FAISS, enabling efficient and fast similarity-based retrieval. When a user submits a query through the web interface, the system processes the input and converts it into embeddings. The retrieval module then searches the vector database to extract the most relevant information corresponding to the user's query. This ensures that the response is based on actual retrieved data rather than relying solely on pre-trained knowledge. The retrieved content is then provided as context to a Large Language Model, such as Gemini, which generates a meaningful, coherent, and human-like response. The generated output is displayed to the user through the interface in a clear and structured format.

The working process is designed to be efficient and user-friendly, enabling quick response generation with minimal delay. The retrieval mechanism ensures that only relevant information is utilized, thereby improving accuracy and reducing noise. The integration of generative models enhances the quality and readability of responses, making them more conversational and easy to understand. The system is optimized for performance by combining efficient retrieval with intelligent generation, ensuring fast and reliable outputs. By leveraging real-time data and semantic search, the system provides accurate and context-aware responses, improving overall user experience and making it suitable for practical research and information retrieval applications.

2.3 RETRIEVAL MODULE

The Retrieval Module plays a crucial role in the Autonomous Research Assistant by identifying and extracting the most relevant information corresponding to user queries. It operates on a vector-based similarity search mechanism, where both user queries and stored data are represented as embeddings. These embeddings are stored in a vector database powered by FAISS, enabling efficient and scalable search operations.

When a user submits a query, it is first converted into a vector representation using an embedding model. The Retrieval Module then compares this query embedding with the stored embeddings in the vector database using similarity measures such as cosine similarity. Based on this comparison, the system retrieves the top relevant data chunks that closely match the semantic meaning of the query.

This approach ensures that the retrieved information is contextually relevant rather than relying on simple keyword matching. By focusing on semantic similarity, the system is able to understand the intent behind the query and provide more accurate results. The retrieved data is then forwarded to the generative model as contextual input, which enhances the quality and reliability of the final response.

The use of vector-based retrieval significantly improves efficiency, allowing the system to handle large volumes of data with minimal latency. Additionally, it reduces irrelevant information retrieval, thereby improving overall system accuracy. The Retrieval Module is a key component of the

RAG architecture, ensuring that the responses generated by the system are grounded in relevant and meaningful data.

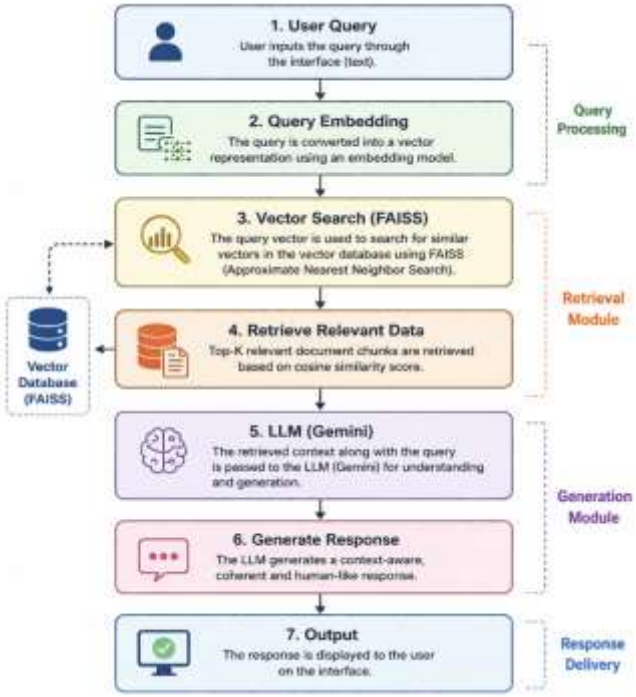

Fig -2: Retrieval Module Workflow

3. CONCLUSION

This paper presented an Autonomous Research Assistant based on Retrieval-Augmented Generation (RAG), designed to provide accurate, context-aware, and real-time information retrieval. By integrating vector-based similarity search using FAISS with advanced language models such as Gemini, the system effectively overcomes the limitations of traditional information retrieval methods and standalone language models. The proposed system enhances response accuracy by grounding generated outputs in relevant retrieved data, thereby reducing hallucination and improving reliability. The use of semantic search through embeddings allows the system to understand user intent more effectively compared to keyword-based approaches. Additionally, the scalable architecture enables efficient handling of large volumes of data while maintaining low latency in response generation.

Overall, the Autonomous Research Assistant demonstrates the effectiveness of combining retrieval mechanisms with generative models to deliver intelligent and meaningful responses. The system provides a practical and scalable solution for modern research and information retrieval tasks, with potential applications across various domains such as education, knowledge management, and decision support systems.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the institution and faculty members for their continuous guidance and support throughout the development of this project. The authors also thank their peers for their valuable suggestions and encouragement, which helped in successfully completing this work.

REFERENCES

1.Gao, Y., et al., “Retrieval-Augmented Genration for Large Language Models: A survey,” arXiv preprint arXiv:2312.10997, 2023.
2.Fan, X., et al., “A Survey on Retrieval-Augmented Generation for Large Language Models,” arXiv preprint arXiv:2405.06211, 2024.
3.Zhang, K., et al., “A Comprehensive Survey of Retrieval-Augmented Generation: Evolution and Future Directions,” arXiv preprint, 2024.
4.Li, H., et al., “Retrieval-Augmented Generation for Educational Applications,” Procedia Computer Science, 2025.