iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73946

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Chatbot Integrated with Gen AI

Battu Nagasahithi[1], Dr. V Uma Rani[2], Dr. Sunitha Vanamala[3]
[1]*Post Graduate Student, MCA, Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India*
[2]*Professor and Head of Dept., Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India*
[3]*Lecturer, Department of Computer Science, TSWRDCW, Warangal, Telangana, India*

Abstract: ***Conversational AI systems powered by Large Language Models (LLMs) have improved natural human–computer interaction but remain limited by static training data and frequent inaccuracies. To address these challenges, this project implements a Retrieval-Augmented Generation (RAG) chatbot that grounds responses in user-uploaded documents. The system uses a modular full-stack architecture with a React frontend, Node.js/Express backend, and a Python microservice for document processing, embedding generation, and semantic retrieval through FAISS. A commercial LLM (Cohere) generates responses only after relevant context is retrieved, ensuring privacy since raw documents are never exposed for training. Testing confirmed that the chatbot delivers domain-specific, reliable answers while minimizing hallucinations and safeguarding sensitive data. The prototype establishes a scalable and secure framework for enterprise and educational use. Future work includes expanding to multimodal data, federated learning, and integration with knowledge graphs for greater adaptability and transparency***
Keywords: Chatbot, Generative AI, Retrieval-Augmented Generation (RAG), FAISS, Semantic Search, Large Language Models (LLMs), Privacy, Full-stack Architecture, Document-grounded QA.

I. INTRODUCTION

Chatbots have emerged as one of the most impactful applications of Artificial Intelligence (AI), supporting sectors such as customer service, healthcare, and education. With the development of Large Language Models (LLMs), modern chatbots can produce human-like responses and sustain natural conversations. Despite these advancements, conventional LLM-based systems have notable limitations.

LLMs are trained on massive but static datasets, making them prone to hallucinations—answers that sound correct but are factually inaccurate. They also lack the ability to adapt dynamically to domain-specific or real-time knowledge, which is critical in enterprise and academic use cases. Furthermore, sending user data to external servers for processing raises serious privacy concerns, preventing adoption in sensitive environments.

To address these challenges, this paper introduces a RAG-powered chatbot integrated with Generative AI. Unlike traditional systems, the chatbot retrieves relevant information from user-uploaded documents before generating a response. This ensures accuracy, domain adaptation, and privacy, while still leveraging the fluency of LLMs.

The proposed system is built on a modular full-stack architecture that separates the frontend, backend, AI microservice, and retrieval engine. The chatbot uses FAISS for fast similarity search, Cohere's LLM for response generation, and a privacy layer that prevents raw document exposure.

This paper contributes:

1) A document-grounded chatbot that reduces hallucinations and ensures verifiable responses.
2) A privacy-aware design that secures sensitive documents.
3) A modular full-stack architecture for scalability and adaptability.
4) A performance evaluation demonstrating accuracy, usability, and efficiency

A. Objective

The Primary objectives of the proposed system are:

- To design a chatbot capable of grounding responses in user-uploaded documents.
- To implement a modular system comprising frontend, backend, and AI microservices.
- To preserve privacy by restricting document exposure to the LLM.
- To test and validate the system through functional and integration evaluation.
- To provide a scalable framework adaptable to enterprise and educational use cases.

II. LITERATURE SURVEY

Research in conversational AI has evolved from rule-based chatbots such as ELIZA [1] and PARRY [2] to sophisticated LLM-based systems like GPT-3 [4]. While rule-based models lacked adaptability, transformer-based architectures [3] enabled contextual generation and large-scale pretraining.

Despite these breakthroughs, conventional LLMs face limitations including hallucinations [6], outdated knowledge, and security risks. To overcome these challenges, Retrieval-Augmented Generation (RAG) [7] integrates a retriever with a generator, grounding responses in external knowledge. Embedding techniques such as Sentence-BERT [10] and similarity search engines like FAISS [11] enable efficient retrieval. Recent work on Fusion-in-Decoder [14] and RETRO [15] has shown improvements in retrieval-generation pipelines. Privacy has also become a focus, with secure enterprise RAG architectures [18] demonstrating the feasibility of deploying such systems in sensitive domains.Recent studies also emphasize the importance of privacy-preserving architectures [17], especially in enterprise and academic contexts. These efforts collectively establish the foundation for the proposed RAG-based chatbot.

III. METHODOLOGY OF PROPOSED SYSTEM

A. *Proposed System*

The proposed chatbot integrates Retrieval-Augmented Generation (RAG**)** with Generative AI to ensure reliable, domain-specific, and privacy-preserving conversational capabilities. Unlike conventional LLM chatbots, this system is capable of dynamically adapting to user-provided knowledge bases through document uploads.

The uniqueness of this system lies in:

- Hybrid architecture: Combines retrieval-based methods with generative AI.
- Plug-and-play adaptability: Any domain knowledge can be added via document upload.
- Privacy by design: Only selected document snippets are shared with the LLM.

This makes the chatbot suitable for enterprise knowledge assistants, academic helpdesks, and research-oriented environments where accuracy and confidentiality are critical.

B. *Dataset Description*

The chatbot does not depend on a fixed dataset. Instead, its knowledge base is dynamically created from documents uploaded by users. Each uploaded document undergoes a preprocessing stage where it is segmented into smaller chunks. These chunks are converted into vector embeddings using a transformer-based encoder such as Sentence-BERT. The embeddings are stored in the FAISS vector store, which supports scalable similarity search.

Whenever a user uploads new material, the knowledge base is updated automatically without retraining. This design ensures that the chatbot is domain-adaptive and always up-to-date. Unlike static datasets used in conventional chatbots, this dynamic dataset generation makes the system more flexible and context-sensitive

C. *System Architecture*

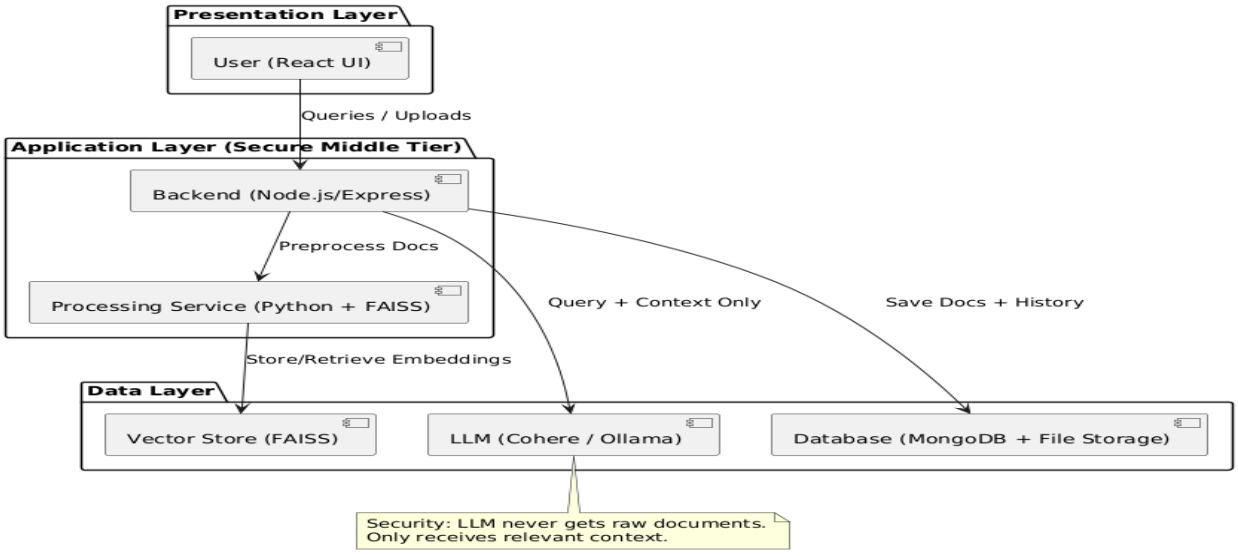


Figure-1: System Architecture of the proposed chatbot

The system follows a three-layered architecture enforcing modularity and privacy.
At the top, the Presentation Layer provides a user interface built with ReactJS, enabling document uploads and conversational interaction.
The Application Layer, built on Node.js and Python microservices, manages request handling, preprocessing, chunking, and embedding generation. The Python service ensures that computationally intensive tasks such as embedding creation are separated from the main backend, enhancing scalability.
At the core lies the Data and Intelligence Layer, which integrates the FAISS vector database, the Cohere LLM, and the privacy enforcement module. FAISS enables fast semantic retrieval, while Cohere generates context-aware answers based on retrieved snippets. The privacy module guarantees that raw documents remain within the system, preventing sensitive data exposure.
This layered design ensures the system is modular, secure, and adaptable across diverse application domains.

D. Methodology

The workflow of the chatbot proceeds as follows.

1) Document Upload: Users upload domain-specific files through the frontend. These are forwarded to the Python microservice.
2) Preprocessing and Embedding Generation: The documents are cleaned, divided into smaller text chunks, and converted into embeddings using Sentence-BERT. These embeddings are stored in the FAISS database.
3) Query Submission: When a user submits a query, it is converted into an embedding by the microservice.
4) Semantic Retrieval: FAISS compares the query embedding with stored embeddings and retrieves the most relevant document chunks.
5) Privacy Enforcement: Only the retrieved snippets are forwarded; the raw documents remain secure within the system.
6) Response Generation: The Cohere LLM combines the query with retrieved context to generate a natural and accurate answer.
7) Answer Delivery: The backend delivers the response to the frontend, where it is displayed in the chat interface.

Through this structured workflow, the system ensures that answers are factually correct, privacy-preserving, and dynamically adaptable to new knowledge.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

A. Key Features

The developed chatbot demonstrates the following core features:

1) Document Upload & Processing – Allows PDF/TXT uploads, automatically extracts text, generates embeddings, and stores them securely.
2) Semantic Retrieval – Uses FAISS to retrieve top-k relevant chunks for query answering, ensuring domain relevance.
3) Context-Aware Generation – Integrates Cohere LLM to generate grounded responses that minimize hallucinations.
4) Data Privacy – User documents remain private; only retrieved snippets are passed to the LLM, ensuring confidentiality.
5) Error Handling & Robustness – Invalid file formats are safely rejected, and failed API calls are managed with clear error messages.

B. Results

The system was tested across multiple functional and performance scenarios.

1) Functional Correctness – Document upload, retrieval, and response generation modules worked as intended. Queries referencing uploaded documents consistently produced accurate answers.
2) Performance – Average query response time was 3–5 seconds, with retrieval latency below 150 ms, meeting the defined performance requirements.
3) Privacy Validation – API request inspection confirmed that only retrieved text snippets were shared with the LLM, not raw documents.
4) Error Handling – Unsupported file formats (e.g., .exe) were safely rejected, validating the robustness of the system.

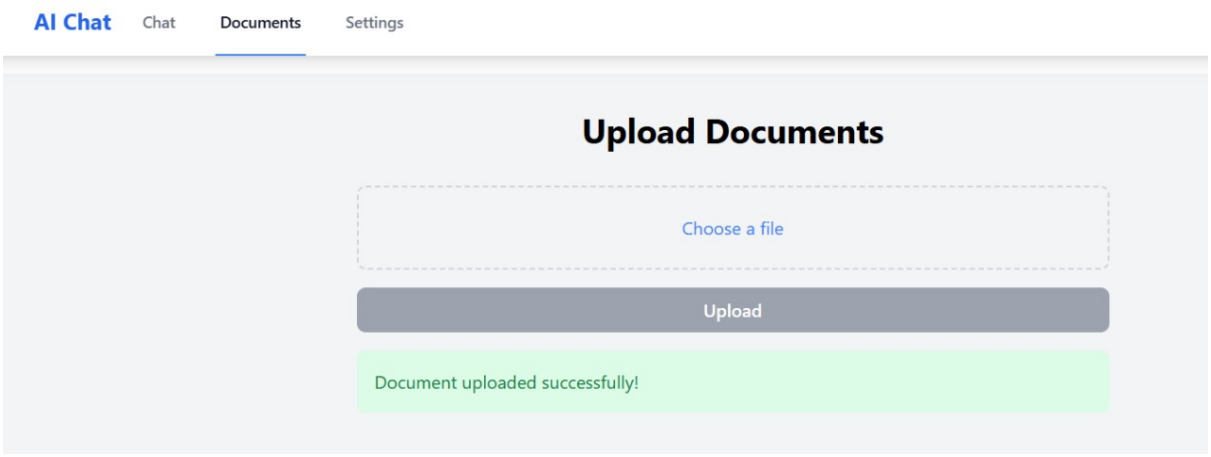


Figure-1:Document Upload Page – Successful Upload

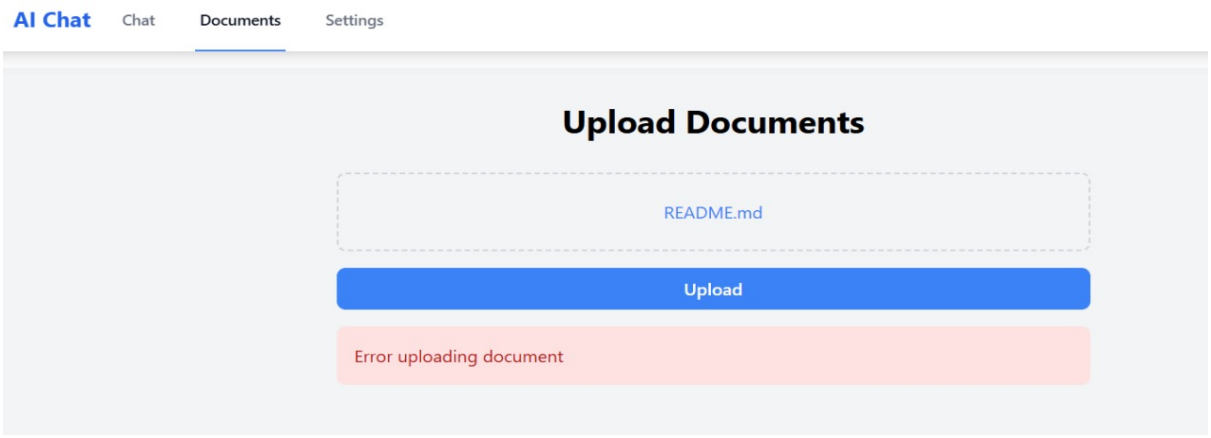


Figure-2: Document Upload Page – Error Handling

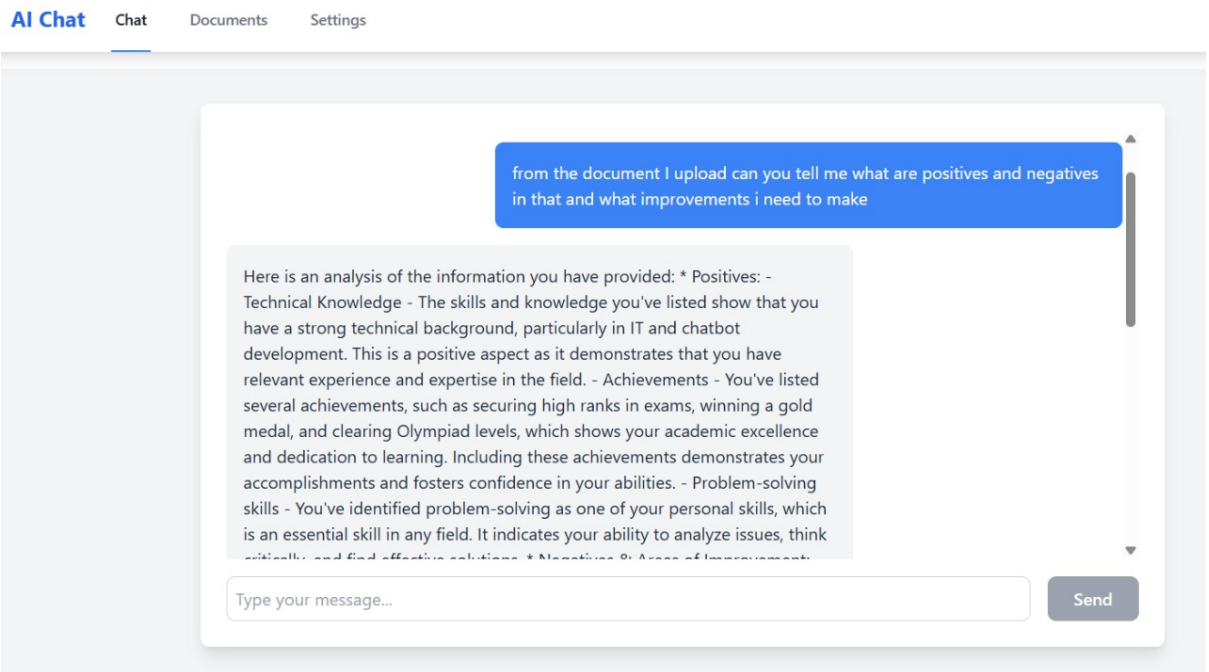


Figure-3: Document-Grounded Response

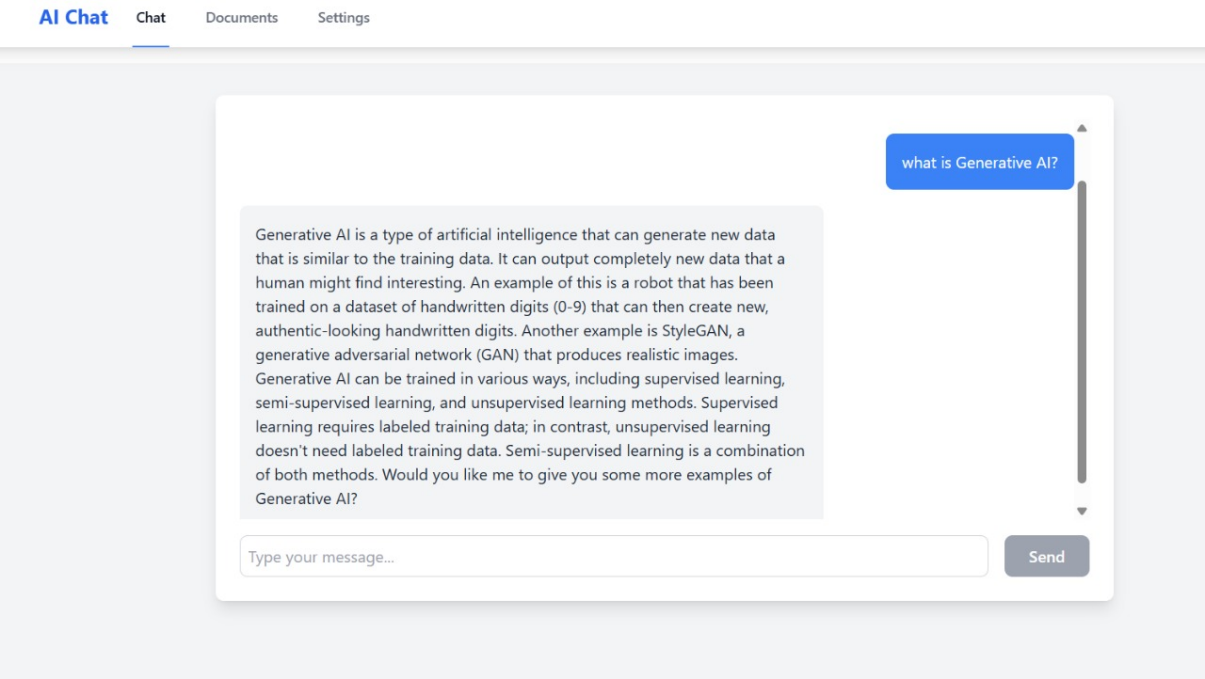


Figure-4: Query without Document

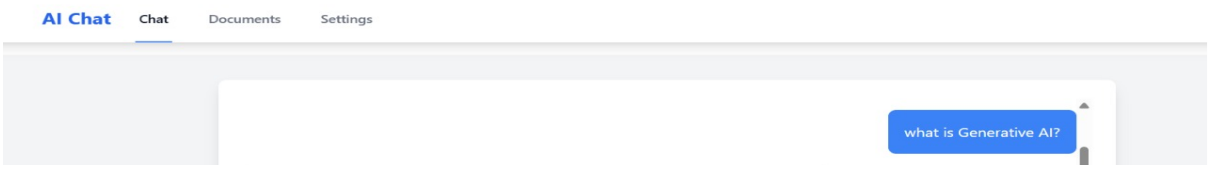


Figure-5: Chat Interface

The experimental outcomes confirm that the proposed chatbot:

- Reduces hallucinations by grounding responses in retrieved evidence.
- Preserves data privacy by restricting document exposure.
- Achieves efficient performance suitable for enterprise and educational use.

V. LIMITATIONS AND FUTURE SCOPE

Although the proposed chatbot demonstrates strong performance in document-grounded conversational tasks, it is subject to certain limitations. The current implementation depends on external APIs such as Cohere for response generation, which introduces reliance on internet connectivity. Additionally, retrieval accuracy may decrease when handling extremely large or noisy datasets, since the quality of embeddings directly impacts the relevance of retrieved snippets. The present version supports only text-based documents, limiting its applicability for multimedia knowledge sources.

In the future, the system can be extended in several directions. Support for multimodal data such as images, audio, and video could enhance versatility. Incorporating knowledge graphs and domain ontologies may further improve retrieval accuracy and contextual understanding. Integration of reinforcement learning techniques could enable the chatbot to learn from user feedback and continually refine its responses. Finally, enterprise deployment at scale may be facilitated by developing on-premise, fully offline models to remove dependency on third-party APIs.

.

VI. CONCLUSION

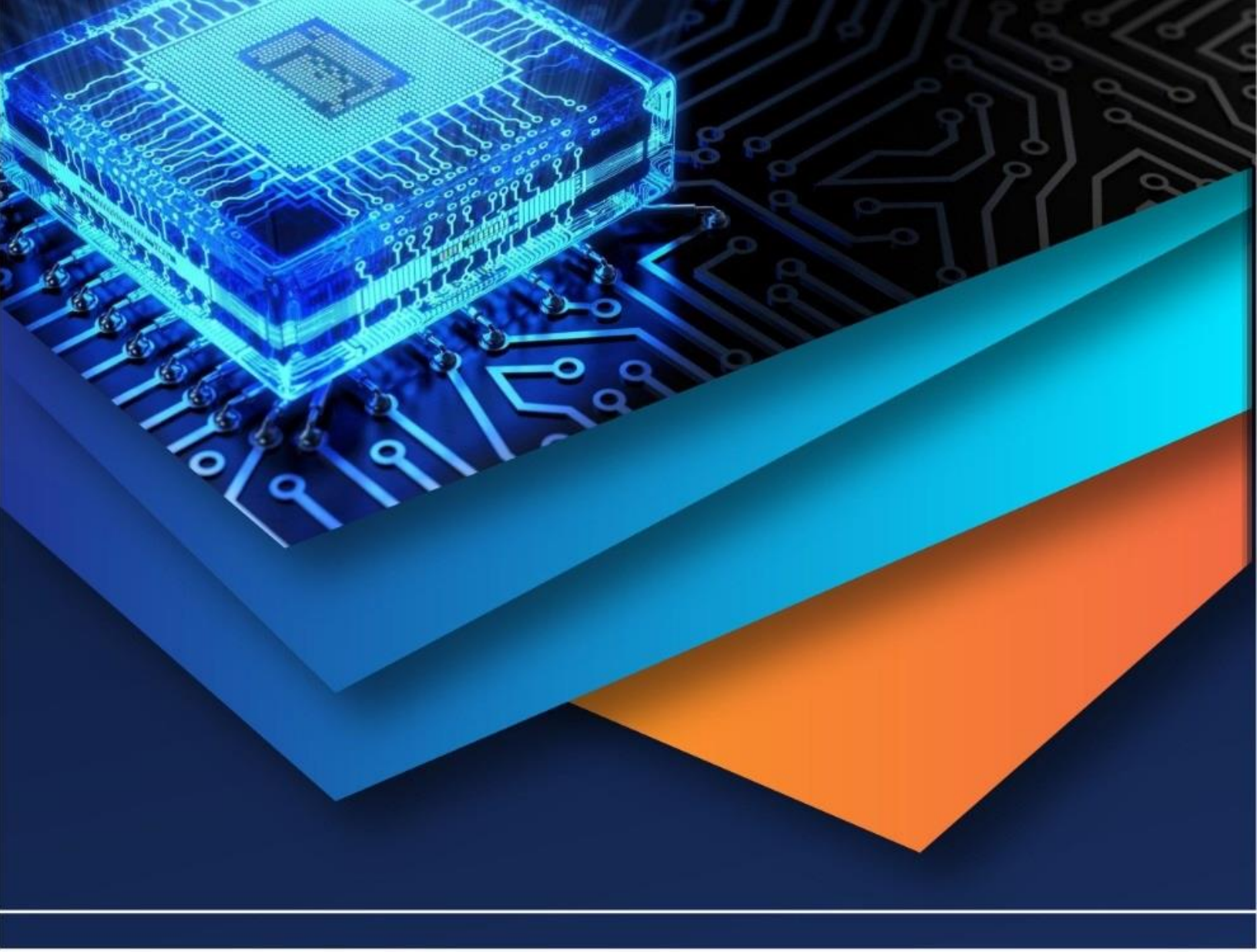
This paper presented a Retrieval-Augmented Generation (RAG)-based chatbot integrated with Generative AI, designed to address the limitations of conventional LLM-driven conversational systems.

By grounding responses in uploaded documents, the system ensures accuracy, adaptability, and privacy preservation. The modular three-layered architecture enables scalability, while the privacy enforcement mechanism guarantees the security of sensitive documents.

Experimental results validated the system's ability to provide factually correct, domain-specific, and efficient responses with reduced hallucinations compared to baseline LLMs. With its adaptable design, the chatbot has strong potential for deployment in academic, enterprise, and research domains, offering a reliable and privacy-conscious conversational interface.

REFERENCES

[1] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.
[2] K. Colby, Artificial Paranoia: A computer simulation of paranoid processes. Pergamon Press, 1975.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, … and D. Amodei, "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.
[5] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[6] S. Ji, T. Xu, Y. Yang, and C. Yu, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.
[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, … and S. Riedel, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
[8] O. Vinyals and Q. Le, "A neural conversational model," arXiv preprint arXiv:1506.05869, 2015.
[9] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. de Oliveira Pinto, J. Kaplan, … and S. Borgeaud, "Evaluating large language models trained on code," arXiv preprint arXiv:2107.03374, 2021.
[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
[11] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[12] Y. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781.
[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[14] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2021.
[15] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, … and K. Kavukcuoglu, "Improving language models by retrieving from trillions of tokens," arXiv preprint arXiv:2112.04426, 2022.
[16] R. Lewis et al., "Question answering with retrieval-augmented generation models," Transactions of the Association for Computational Linguistics (TACL), vol. 9, pp. 1–15, 2021.
[17] A. Özgür, S. Singh, and M. Ahmad, "Privacy-preserving architectures for enterprise conversational AI," in Proc. International Conference on Data Engineering (ICDE), 2024.
[18] J. Gao, X. He, and J. Li, "Neural approaches to conversational AI," Foundations and Trends in Information Retrieval, vol. 13, no. 2–3, pp. 127–298, 2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY