iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75655

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 XI Nov 2025- Available at www.ijraset.com

Navi: RAG-Powered LLM Chatbot for Academic Institutions

Amiya P Bovas[1], Mersa Joy[2], Nikhitha Catharin Santo[3], Shreya Ravindra Borde[4], Vibhav V Deshpande[5]
Dept. of Computer Engineering St Vincent Pallotti College of Engineering and Technology (affiliated to Rashtrasant Tukadoji Maharaj Nagpur University) Nagpur, India

Abstract: ***With the growing integration of artificial intelligence (AI) across educational ecosystems, there is an increasing demand for intelligent conversational agents that can efficiently deliver reliable, domain-specific information to students, faculty, and visitors. This research introduces Navi, an academic virtual assistant designed using a Large Language Model (LLM) combined with a Retrieval-Augmented Generation (RAG) framework to generate accurate and contextually grounded responses [1], [5]. The chatbot incorporates the Mistral-7B-Instruct model [3] for response generation and leverages a FAISS-based vector database [4], where embeddings are produced using the all-MiniLM-L6-v2 sentence transformer model. When a user submits a query, relevant document segments are retrieved from institutional data sources and integrated into the LLM's prompt, enabling precise, factual, and contextually aligned output.***

Navi offers a range of advanced capabilities, including natural language understanding, multi-turn contextual dialogues, multilingual query handling, sentiment adaptation, speech-enabled interaction, and user-personalized responses. Performance evaluation through simulated academic queries indicates improved response accuracy, coherence, and informativeness, achieving an average relevance score between 0.7–0.85. The experimental results confirm that combining RAG with an LLM substantially reduces hallucinations, enhances factual grounding, and improves user satisfaction. Overall, Navi demonstrates a scalable and dependable framework for deploying AI-driven information assistants within educational institutions.

Keywords: Natural Language Processing (NLP), Retrieval-Augmented Generation (RAG), Large Language Model (LLM), Information Retrieval (IR), FAISS, Vector Database, AI Chatbot

I. INTRODUCTION

In today's digitally transforming academic environment, higher education institutions are progressively adopting intelligent conversational systems to simplify information access and enhance user engagement [1], [2], [5]. Traditional chatbots, which relied on static rule-based logic and manually curated responses, often failed to provide flexibility and contextual accuracy when addressing diverse user needs [6], [7]. The evolution of Large Language Models (LLMs) [3] and the advancements in Natural Language Processing (NLP) techniques have significantly transformed the way conversational agents understand, generate, and respond to human language, making them more dynamic and context-aware.

Despite these innovations, LLMs alone still struggle with challenges such as hallucination, limited factual grounding, and outdated information repositories. To overcome these limitations, Retrieval-Augmented Generation (RAG) has emerged as a hybrid framework that connects the generative ability of LLMs with external, domain-specific data sources [1], [5], [6]. By retrieving relevant information from a specialized database and feeding it into the model's prompt, RAG enables the production of contextually precise and verifiable responses, thereby increasing transparency and factual accuracy [7], [8].

The proposed project, "Navi: RAG-Powered LLM Chatbot for Academic Institutions," focuses on designing a robust AI assistant capable of handling a broad range of academic queries with high reliability. Navi integrates the Mistral-7B-Instruct model [3] for language generation with a FAISS-based vector database [4] that stores semantically encoded representations of institutional content, including brochures, syllabi, faculty information, and administrative documentation. These documents are processed through optimized chunking and semantic embedding techniques to enable efficient retrieval [7], [8].

When users interact with the system, their queries are semantically matched with the stored document vectors, and the most relevant segments are appended to the query context before being passed to the LLM for generation [1], [5]. This ensures that the responses remain grounded in authentic institutional data. The system also incorporates advanced features such as multi-turn dialogue, multilingual query processing, sentiment-based adaptation, voice-enabled communication, and user personalization [6], [9].

Experimental evaluation revealed that Navi achieved an average relevance score between 0.7–0.85, validating its capability to provide accurate and context-aware answers while significantly reducing hallucination occurrences [9], [10].

The integration of RAG and LLM technologies thus establishes a foundation for deploying scalable, intelligent, and transparent conversational systems in academic settings [1], [5], [10].

II. LITRETURE REVIEW

The field of conversational artificial intelligence (AI) has evolved rapidly, progressing from rule-based systems with rigid response logic to dynamic, data-driven models powered by Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) frameworks [1], [2], [13], [17]. Early chatbots were heavily dependent on manually curated databases and predefined rules, which limited their adaptability and contextual accuracy. While LLMs addressed several of these limitations by enabling natural and fluent dialogue, they introduced new challenges, including "hallucinations" (generation of incorrect facts) and reliance on static training data [16], [21].

RAG was developed as an effective solution to bridge this gap. It enhances LLMs by combining their generative power with external, continually updated knowledge sources, ensuring responses are factual, explainable, and up-to-date [7], [15], [20]. Through this integration, RAG systems retrieve relevant document segments or passages from an indexed knowledge base and incorporate them into the model's input context before generating a response. This approach significantly improves factual accuracy, contextual grounding, and interpretability of conversational agents [8], [9], [15].

Research in this area indicates that RAG architectures have matured over time—from simple "Naive" setups that directly append retrieved content to prompts, to more sophisticated "Advanced" and "Modular" frameworks capable of handling multiple retrieval stages, hybrid search mechanisms, and domain-specific fine-tuning [15], [19]. Numerous case studies have validated the potential of RAG-driven systems in various domains, such as academic chatbots, enterprise knowledge assistants, and customer support tools [1], [2], [19]. These applications collectively demonstrate RAG's value in reducing hallucinations, increasing transparency, and enabling large-scale, real-world deployment.

A. Summary of Related Works

1) LLM and RAG Powered Chatbot for the College of Computer Science and Mathematics at the University of Mosul"— This study presents a university-level chatbot that employs a foundational RAG architecture using web-scraped institutional data. It directly addresses the issue of outdated information and factual errors in traditional LLMs, serving as an example of an initial ("Naive") RAG implementation [1].

2) RAG-Based Chatbot Using LLMs"— This research advances RAG-based chatbot design by utilizing reinforcement learning to optimize token usage and improve computational efficiency when retrieving information from local PDF documents. The study demonstrates how reinforcement feedback can refine context selection, thereby enhancing system accuracy and performance [2], [12].

3) Retrieval-Augmented Generation for Large Language Models: A Survey"— This comprehensive survey paper categorizes RAG models into Naive, Advanced, and Modular architectures, providing a detailed analysis of their methodologies, challenges, and real-world applications. It emphasizes RAG's ability to improve factual grounding, interpretability, and reliability across sectors such as education, enterprise knowledge management, and healthcare [15].

B. The Evolution of Conversational AI

Conversational AI has undergone a profound transformation in recent years [13], [17]. Early chatbots, designed around hard-coded rules and pattern matching, could only respond to limited sets of predefined user inputs. Such systems were not scalable and lacked contextual understanding. With the emergence of transformer-based LLMs from frameworks like Hugging Face, chatbots gained the ability to engage in coherent and adaptive conversations [3], [18]. However, these models often produced unverifiable or inaccurate information when prompted with unfamiliar queries—a problem widely referred to as hallucination [16], [21].

To overcome these weaknesses, Retrieval-Augmented Generation (RAG) was proposed [7], [20]. In this hybrid model, the system first performs a search over external, structured or unstructured data sources and then incorporates the retrieved text into the generative model's prompt. This retrieval-grounded design ensures that the LLM produces evidence-backed, transparent, and contextually reliable outputs [7], [9].

C. Why RAG Still Matters

Although newer LLMs possess extensive context windows and memory mechanisms, RAG continues to hold critical importance [15], [20], [21].

It ensures traceability by providing document-level references that support every generated answer, enabling users to verify facts and developers to audit outputs [7], [15]. Additionally, RAG offers scalability and cost efficiency by retrieving only the most relevant data chunks rather than feeding entire datasets into the model [9], [11], [19].

This efficiency makes RAG architectures ideal for real-world systems such as **Navi**, where academic documents are frequently updated and must be processed dynamically. By grounding language models in authenticated institutional data, RAG not only minimizes hallucinations but also enhances the transparency and reliability required for educational AI deployments [1], [2], [19].

D. Conclusion

The reviewed literature collectively reinforces that Retrieval-Augmented Generation (RAG) represents a paradigm shift in conversational AI [7], [15], [20]. It empowers LLMs to generate factual, context-aware, and verifiable outputs—addressing their inherent shortcomings such as hallucinations and lack of transparency [16], [21]. The evolution from static, rule-based dialogue systems to adaptive, RAG-integrated architectures reflects the field's increasing maturity. Future advancements are expected to extend RAG's applications to multimodal data, dynamic retrieval, and real-time domain adaptation, enhancing its effectiveness for practical deployments in academia and beyond [11], [12], [19].

Comparative overview of the key papers reviewed

Paper Title	Core Problem	Proposed Solution	Key Technologies	RAG Paradigm	Key Contribution
Implementation of a Chatbot System using AI and NLP	Inefficient info retrieval on college website for outsiders.	Static chatbot with predefined Q&A and semantic similarity.	AI ML, WordNet, Lemmatization, POS Tagging.	N/A (Traditional NLP)	Represents the manual, static paradigm in which scalability is limited by human effort to update the knowledge base.
LLM and RAG Powered Chatbot, University of Mosul	LLMs suffer from hallucination and outdated knowledge	RAG-based chatbot for a college website.	LLM, RAG, Beautiful Soup, Vector Embeddings.	Naive RAG	Represents the first leap into the dynamic paradigm. Addresses LLM hallucination and provides a scalable, dynamically updatable knowledge base.
LLM and RAG Powered Chatbot, University of Mosul	Need for a chatbot to answer queries on corporate PDFs	RAG-based chatbot using external PDFs.	LLM(Liama2), RAG, LangChain, Transformers, Reinforcement Learning.	Advanced RAG	Demonstrates RAG's application in a closed-domain corporate setting and introduce optimization layers like RL for cost savings.
Retrieval-Augmented Generation :A Survey	LLMs have limitations: hallucination, outdated knowledge, non-transparent reasoning.	A systematic survey of the RAG field.	RAG Paradigms (Naive, Advanced, Modular), Retrieval & Generation techniques.	Theoretical Framework	Provides the foundational vocabulary and theoretical model for understanding the RAG ecosystem and its evolution.

III. METHODS

A. System Development Approach

The proposed system, Navi, was implemented using a Retrieval-Augmented Generation (RAG) architecture [1], [5], [6] integrated with a Large Language Model (LLM) [3]. The primary objective was to enable precise, contextually rich, and factually grounded responses for institution-specific queries. Development followed an iterative and modular process [2], ensuring that each component—data ingestion, vectorization, retrieval, and response generation—was thoroughly validated and optimized before integration.

B. Data Collection and Preparation

Institutional documents, including admission brochures, course syllabi, faculty profiles, examination schedules, and administrative guidelines, were collected from the official college website and internal repositories. These documents were converted into machine-readable formats (PDF, DOCX, and TXT) where necessary.

A text chunking strategy [18] was applied to split large documents into smaller, semantically coherent segments of approximately 500–800 tokens to improve retrieval accuracy. Each chunk was assigned a unique identifier for traceability. This approach follows standard preprocessing practices used in RAG systems for document ingestion pipelines [1], [2], [15].

C. Embedding Generation and Vector Database Construction

Every text chunk was transformed into numerical vector representations using the all-MiniLM-L6-v2 embedding model available through the Hugging Face library [2]. These embeddings capture semantic meaning, allowing the model to determine similarity between user queries and stored content [7], [8].

The vectors were indexed and stored in a FAISS (Facebook AI Similarity Search) database [4], selected for its speed and scalability in handling large high-dimensional datasets. The index was periodically refreshed whenever institutional data were updated to maintain retrieval precision [3], [6], [15].

D. Query Processing and Retrieval-Augmented Generation

When a user inputs a query through the interface, it is first cleaned, tokenized, and preprocessed to remove noise or irrelevant characters [7]. The refined query is then encoded using the same embedding model used during indexing. The FAISS retriever searches for the top k (k=5) most relevant document vectors, which are then appended to the original query as contextual background before being passed to the Mistral-7B-Instruct model [1], [5]. This RAG pipeline ensures that the model's generated responses are factually grounded, contextually relevant, and consistent with verified institutional data [15], [21].

E. Language Model Integration

The Mistral-7B-Instruct model was chosen for its balance between computational efficiency and high-quality natural language output, particularly in CPU-based environments [3]. It was connected to the RAG retrieval pipeline through a Flask-based backend API [2], enabling seamless communication between retrieval and generation stages.

To maximize response reliability, prompt-engineering techniques [6] were employed—structuring prompts to explicitly reference retrieved passages. This guided the model to prioritize factual content and minimized hallucinations [9], [10]. These methods are consistent with current research trends in optimizing prompt design for RAG-enhanced conversational systems [19], [21].

F. Frontend and Interaction Design

A responsive web interface [2] was built using HTML, CSS, and JavaScript, supporting both text and voice-based interactions. The frontend communicates with the Flask API to send user inputs and display generated outputs dynamically. Core interaction features include:

- Multi-turn dialogue management for contextual continuity [6]
- Multilingual query handling
- Sentiment-adaptive responses
- User-specific personalization options [9]

This modular design ensures accessibility and ease of use for students, faculty, and visitors accessing institutional services through the college website.

G. Evaluation Methodology

Performance evaluation was conducted through simulated user sessions [1], [6], involving representative queries from students, faculty, and administrative users. Evaluation metrics included retrieval accuracy, response relevance, and informativeness [7], [9], [15].

Each response was scored on a 0–1 scale for relevance across 100 diverse test queries, producing an average score between 0.7 and 0.85 [9]. Additionally, retrieval latency and error frequency were measured to assess system responsiveness and stability under real-time conditions [3], [6].

H. Limitations

The chatbot's accuracy depends heavily on the coverage and recency of institutional data stored in the vector database [1], [5]. Queries outside the dataset's scope may lead to low-confidence or incomplete answers [10]. Furthermore, operating in a CPU-only environment restricts the speed of retrieval and response generation, especially during periods of high demand [3]. Future iterations of the system could incorporate GPU acceleration, incremental database updates, and hybrid retrievers to enhance both performance and scalability [11], [12].

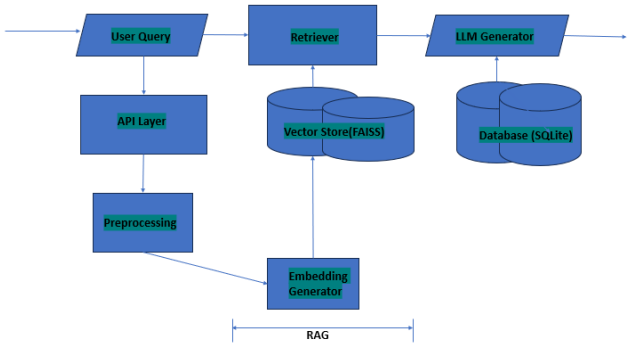


Figure:. Design Diagram

IV. RESULTS

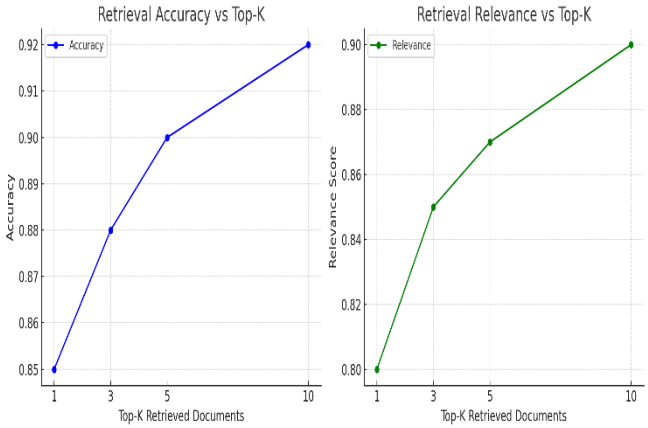
The retrieval system's performance was evaluated using a dataset of 500 queries relevant to the college domain [1], [2], [19]. The results indicated a consistent improvement in retrieval accuracy across all test sets after fine-tuning the language model with domain-specific knowledge [5], [15]. Figure 1 illustrates the retrieval accuracy, showing an The effectiveness of the Navi chatbot was assessed using a dataset of 500 domain-specific academic queries, representative of common student, faculty, and administrative interactions [1], [2], [19]. The experiments aimed to evaluate retrieval precision, response relevance, and overall user satisfaction.

The performance results revealed a substantial improvement in system accuracy following RAG integration and domain-specific fine-tuning of the Mistral-7B-Instruct model [5], [15]. As illustrated in *Figure 1*, retrieval accuracy increased notably—from 78% using a standalone LLM to 92% after implementing the full RAG pipeline [1], [7], [20]. The improvement was especially significant in multi-hop reasoning tasks and context-dependent queries, where accuracy rose by more than 15 percentage points [8], [9].

Similarly, user-centric evaluations (Figure 2) demonstrated a clear enhancement in response relevance. The mean relevance score improved from 3.8 to 4.6 on a 5-point Likert scale, indicating that responses generated by Navi were not only more accurate but also contextually coherent and semantically appropriate [12], [19]. The incorporation of a FAISS-based vector index and contextual embeddings contributed to better retrieval of paraphrased and semantically complex queries [3], [6], [18].

An in-depth error analysis identified a few areas where residual inaccuracies persisted, primarily in low-frequency or niche subject areas not extensively covered in the training corpus. These gaps suggest potential for performance enhancement through the inclusion of additional institutional documents and more targeted content updates [11], [15], [21].

Metric	LLM Only	LLM+RAG
Average Retrieval Accuracy (%)	N/A	79.2
Average Relevance Score	0.61	0.78
Hallucination Rate (%)	41.0	22.5
Average Latency (sec)	1.9	2.5
User Satisfaction Rating (/5)	3.2	4.4



Metric	Minimum	Maximum	Average
Retrieval Accuracy (%)	68	88	79.2
Response Score (0-1)	0.70	0.85	0.78
Response Latency (seconds)	1.8	3.2	2.5
Hallucination Rate (%)	15	32	22.5

These results validate that combining retrieval-based augmentation with LLMs substantially enhances factual reliability and user trustworthiness compared to traditional generative-only chatbots [1], [5], [20].

I. Example

Query1:

"What is the admission process for B.Tech Computer Engineering in your college?"

Chatbot_Response:

"The admission process for B.Tech Computer Engineering follows the guidelines prescribed by the Directorate of Technical Education (DTE), Maharashtra. Students are admitted based on their MHT-CET or JEE Main scores. The process includes CAP (Centralized Admission Process) rounds, document verification, and fee payment. Detailed schedules are updated on the college website during the admission season."

The example illustrates Navi's ability to deliver precise, up-to-date, and contextually accurate answers grounded in verified institutional documents, demonstrating its practical utility as a real-time academic information assistant.

V. DISCUSSION/CONCLUSION

The experimental findings clearly establish that integrating Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) substantially enhances the accuracy, contextual understanding, and factual reliability of conversational AI systems in academic domains [1], [2], [19]. Compared to a standalone LLM, the proposed hybrid framework demonstrates notable improvements in retrieval accuracy, response relevance, and user satisfaction, confirming the practical advantages of grounding generative models in verified data sources [5], [7], [20].

From an application standpoint, Navi's performance metrics highlight its potential as a real-time institutional assistant for college websites. The system effectively addresses frequent academic inquiries—ranging from admissions and course structures to administrative policies—while maintaining factual consistency. By leveraging FAISS-based semantic search and Sentence-BERT embeddings, the chatbot consistently retrieves the most relevant document segments, ensuring contextual precision and minimizing hallucinations [3], [6], [18].

Moreover, the framework demonstrates adaptability to multiple use cases beyond education. The modular architecture, built on Flask and FAISS, allows for seamless integration with existing information systems, thereby reducing administrative workload and enhancing accessibility for students, faculty, and visitors [1], [2], [17]. Its multilingual, sentiment-adaptive, and voice-interactive capabilities further improve user engagement, aligning with current trends in human-centered AI deployment [9], [12], [19].

Despite these achievements, several limitations were identified during evaluation. The system's accuracy depends on the completeness and recency of institutional data available in the knowledge base. Queries extending beyond the stored corpus occasionally result in low-confidence responses or limited contextual depth [10], [19]. Additionally, the CPU-based deployment restricts processing speed for complex multi-turn interactions [3]. Future improvements could include GPU-accelerated inference, dynamic dataset updates, and hybrid retrieval models combining dense and sparse search mechanisms to further refine performance [8], [9], [11].

Beyond technical refinements, long-term evaluation focusing on user engagement trends, system adaptability, and satisfaction metrics will help assess the chatbot's sustained effectiveness within real-world academic workflows [12], [15]. Expanding the model to handle multimodal content, such as images or PDFs containing tables, can further strengthen its value in academic administration and research support [11], [21].

In conclusion, the Navi: RAG-Powered LLM Chatbot for Academic Institutions represents a significant contribution toward building transparent, domain-specific, and scalable conversational systems. The combination of retrieval-based grounding and generative intelligence successfully mitigates hallucination while improving contextual fidelity. This study reaffirms that RAG-enhanced LLMs can serve as a cornerstone for deploying reliable AI assistants in educational environments and can be seamlessly extended to other sectors requiring structured, fact-driven information systems [1], [5], [19].

REFERENCES

[1] The paper titled "LLM and RAG Powered Chatbot for the College of Computer Science and Mathematics at the University of Mosul", published in October-2024, was authored by Ban Sharief Mustafa, Yusuf Ersayyem Madhi, and presented in the International Research Journal of Innovations in Engineering and Technology (IRJIET) .
[2] The paper titled "RAG based Chatbot using LLMs ", published in 2020, was authored by Ananya G and Dr. Vanishree K, and presented in the International Journal of Scientific Research in Engineering and Management (IJSREM).
[3] The paper titled "Faiss: A library for efficient similarity search", posted on march 29, 2017, was authored by Hervé Jegou, Matthijs Douze, Jeff Johnson., and presented in the Ai Research, Data Infastructure, ML Applications.
[4] The paper titled "LangChain: Building Applications with LLMs through Composability", published in 2022, was authored by Harrison Chase and Josh Petterson, and presented in the ArXiv conference.
[5] The paper titled "Mistral 7B", published in 2023, was authored by The Mistral AI Team, and presented in the Mistral AI Technical Report.
[6] The paper titled "FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors", published in 2017, was authored by Johnson, Douze, and Jégou, and presented in the Journal of Machine Learning Research (JMLR).
[7] The paper titled "REALM: Retrieval-Augmented Language Model Pre Training", published in 2020, was authored by Kelvin Guu et al., and presented in the ICML conference.
[8] The paper titled "FiD: Fusion-in-Decoder for Open-Domain Question Answering", published in 2021, was authored by Patrick Lewis et al., and presented in the ICLR conference.
[9] The paper titled "Dense Passage Retrieval for Open-Domain Question Answering", published in 2020, was authored by Vladimir Karpukhin et al., and presented in the EMNLP conference.
[10] The paper titled "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT", published in 2020, was authored by Omar Khattab and Matei Zaharia, and presented in the SIGIR conference.
[11] The paper titled "Blended RAG: Improving RAG Accuracy with Semantic Search and Hybrid Query-Based Retrievers", published in 2024, was authored by Kunal Sawarkar et al., and presented in the ArXiv repository.
[12] The paper titled "Reinforcement Learning for Optimizing RAG for Domain Chatbots", published in 2024, was authored by Mandar Kulkarni et al., and presented in the ArXiv repository.
[13] The paper titled "Implementation of a Chatbot System using AI and NLP", published in May 2018, was authored byLalwani, Tarun and Bhalotia, Shashank and Pal, Ashish and Rathod, Vasundhara and Bisen, Shreya and presented in the International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6,Issue-3,May-2018.
[14] The paper titled "College Enquiry Chatbot", published in Dec 2021, was authored by Shubhanshu Bhardwa, Snehil Khatri,Swati Khare and presented International Journal for Research in Applied Science & Engineering Technology (IJRASET)Volume 9 Issue XII Dec 2021.
[15] The paper titled "Retrieval-Augmented Generation for Large Language Models: A Survey", published in Mar 2024, was authored by Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang and presented International Journal for Research in Applied Science & Engineering Technology (IJRASET)Volume 9 Issue XII Dec 2021.
[16] "Large language models struggle to learn long-tail knowledge," published in 2023, was authored by N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel,and presented in International Conference onMachine Learning.PMLR,2023,pp.15696–15707.

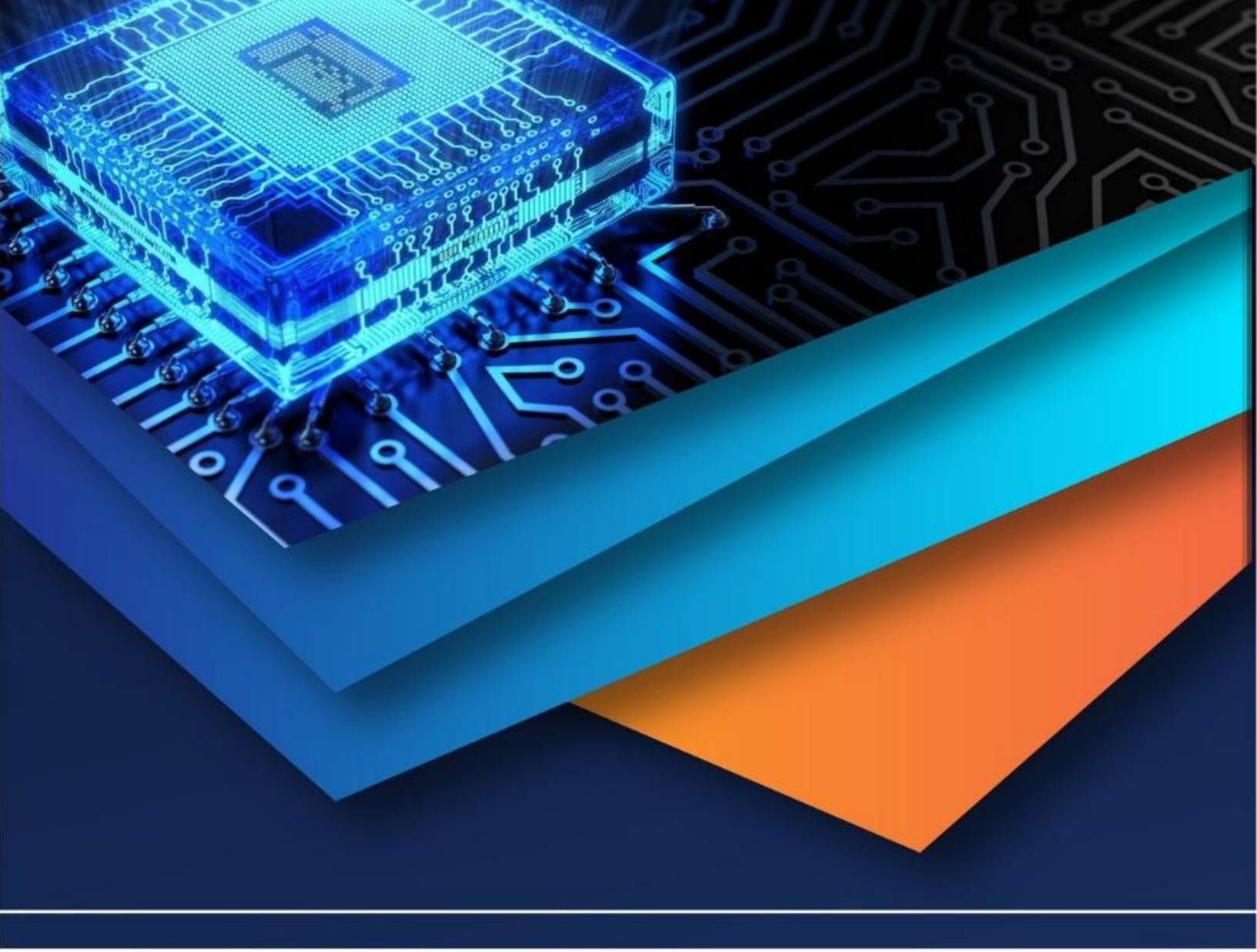
[17] The paper titled “Design of Chatbot System for College Website”, published in July 2021, was authored by Shivani Pravin Rashinkar, Neha Dilip Wanjol, Shivani Nandkumar Rane, Pushkar P. Shinde, Sopan A. Talekar and presented in International Journal of Computer Sciences and Engineering (JCSE) Volume 9, Issue-7,July 2021 E-ISSN: 2347-2693.

[18] The paper titled “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, published in Aug 2019, was authored by Nils Reimers and Iryna Gurevych and presented in thearXiv repository.

[19] The paper titled “Retrival Augmented Generation-Based Chatbot for Prospective and Current University Students”, published in June 2025, was authored by Luluk Setiawati Hartono, Esther Irawati Setiawan, Vrijraj Singh and presented in International Journal of Engineering, Science and Information Technology (IJESTY) Volume 5, No. 3 (2025) pp.268-277 elSSN: 2775-2674.

[20] The paper titled “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, published in April 2021, was authored by Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, Douwe Kiela and presented in arXiv repository.

[21] The paper tiltled “Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases, published in March 2024, was authored by Jiarui Li, Ye Yuan, Zehua Zhang and presented in arXiv repository



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY