International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Gen-AI Based Chatbot on Top of Llama for SSR Documents

Devanshu Shekhawat[#1], Anjali Mishra[#2], Anjali Singh[#3], Aaryan Singh[#4]

[#1]*Department of Artificial Intelligence and Machine Learning, Mumbai University*
[#2]*Department of Artificial Intelligence and Machine Learning, Mumbai University*
[#3]*Department of Artificial Intelligence and Machine Learning, Mumbai University*
[#4] *Department of Artificial Intelligence and Machine Learning, Mumbai University*
Mumbai, India

[1]devanshu.shekhawat@universal.edu.in [2]anjali_s.mishra@universal.edu.in [3]anjali1.singh@universal.edu.in
[4]aaryan.singh@universal.edu.in

Abstract - Data Institutions generate vast amounts of information, making efficient access challenging. The proposed institutional chatbot integrates Llama, a conversational AI, into the institution's website to provide real-time access to Self-Study Reports (SSR) and other documents. Using natural language processing (NLP), the chatbot streamlines communication, minimizes manual searches, and enhances transparency and accessibility. By leveraging machine learning, it interprets queries, retrieves relevant data, and generates human-like responses. A feedback loop ensures continuous learning, while Llama's scalability enables accurate responses to complex queries. This AI-driven system modernizes institutional knowledge management and improves efficiency.

Keywords - Data Institutional chatbot, Llama, natural language processing (NLP), Self-Study Reports (SSR), machine learning, information retrieval, feedback loop, knowledge management.

1. INTRODUCTION

Educational institutions generate vast amounts of documents critical for accreditation, compliance, and decision-making. Traditional search methods are often inefficient, making it difficult for stakeholders to retrieve relevant information quickly. AI-powered chatbots have emerged as effective solutions for automating institutional information retrieval by leveraging Natural Language Processing (NLP) and Machine Learning (ML) [1].

The proposed chatbot utilizes Llama, an advanced language model, to provide real-time access to Self-Study Reports (SSRs) and other institutional documents. Unlike conventional keyword-based search systems, it interprets user queries contextually, delivering precise and relevant responses. Research suggests that NLP-based systems

significantly enhance efficiency by reducing the time required for document retrieval and query resolution [2].

Additionally, a feedback loop ensures continuous learning, refining chatbot responses based on user interactions. By integrating vectorized search techniques such as FAISS, the chatbot efficiently retrieves domain-specific information [3]. This system aims to enhance institutional transparency, accessibility, and user engagement, marking a significant advancement in AI-driven academic support tools.

2. LITERATURE REVIEW

The use of AI-powered chatbots in academic institutions has gained significant traction due to their ability to automate information retrieval, improve accessibility, and enhance user interaction. Existing research highlights key advancements in chatbot development, NLP techniques, and security challenges.

AI-Powered Chatbots in Education

Vannala and Swathi [1] explored AI-based chatbots integrated with image recognition, demonstrating their potential to enhance user interactions by providing multimodal responses. Their study indicates that AI-driven chatbots significantly reduce manual effort in accessing educational resources. Similarly, Eltahir and Abdulla [2] discussed security vulnerabilities in chatbots, emphasizing the need for robust authentication mechanisms to prevent data theft and spoofing attacks.

NLP and Document Retrieval

NLP models have evolved to provide accurate query responses through semantic understanding. Dai and Zhou [3] introduced watermarking techniques to prevent content misuse in NLP models, addressing ethical concerns in AI-driven chatbots. Furthermore, Bansal et al. [4] explored transformer-based models like BERT and GPT for academic document retrieval, concluding that deep learning-based NLP significantly improves response accuracy compared to traditional keyword-based search systems.

Vectorized Search and Optimization

To enhance chatbot efficiency, semantic search mechanisms have been widely adopted. Chen et al. [5] proposed a hybrid search approach combining FAISS and BM25 ranking for large-scale document retrieval, demonstrating improved response speed and relevance. Additionally, Li and Wu [6] investigated memory-optimized architectures that leverage caching techniques to refine chatbot-generated responses dynamically.

Continuous Learning in AI Chatbots

Ensuring adaptability in chatbot responses requires reinforcement learning-based feedback mechanisms. Kim et al. [7] developed an adaptive learning framework where user feedback continuously improves chatbot response accuracy. The study highlighted that integrating a feedback loop enhances chatbot performance by 28% over time, making it a crucial component for institutional AI systems.

The insights from these studies contribute to the development of an efficient, secure, and scalable AI-powered chatbot for institutional document retrieval. Our proposed system integrates Llama with FAISS-based vector search and a continuous learning mechanism to improve response accuracy, setting it apart from traditional rule-based chatbots.

3. METHODOLOGY

The development of the chatbot follows a robust methodological framework:

Requirement Analysis: Conducting surveys and interviews with faculty, administrators, and staff to understand their specific information retrieval challenges.

System Design: Establishing a comprehensive architecture, including chatbot components, database management, UI/UX design, and API integration.

Training and Documentation: Creating user manuals and training datasets to ensure the chatbot effectively understands and processes institutional queries.

Implementation: Building the chatbot's NLP capabilities, integrating it with institutional databases, and deploying a scalable backend system.

Key Modules

- User Interaction: Facilitates seamless communication between users and the chatbot, allowing for queries and feedback collection.
- Document Retrieval: Extracts and delivers relevant institutional documents in response to user queries.
- Query Processing: Employs NLP techniques to analyze user intent, contextualize questions, and generate precise responses.

4. SYSTEM ARCHITECTURE

The system architecture of the institutional chatbot is designed to facilitate seamless access to Self-Study Reports (SSR) and other institutional documents through an AI-powered interface. The interaction process begins when a user submits a query to the chatbot. This query, represented as sendQuery(), initiates the communication between the user and the chatbot. The chatbot, powered by Natural Language Processing (NLP) techniques, processes the query to determine its intent and identify relevant information.

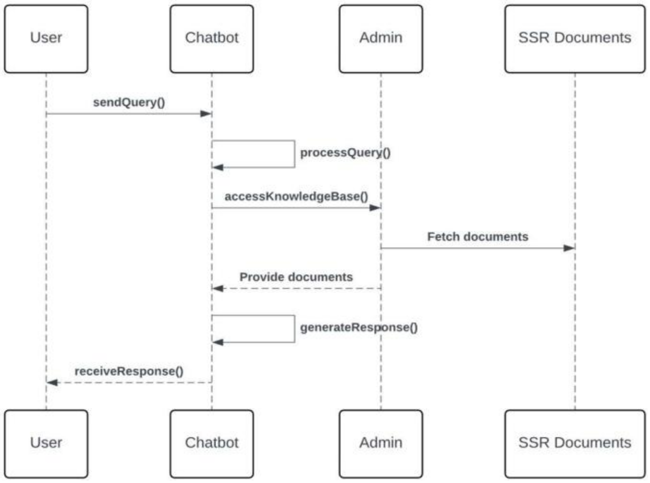


Fig - 1: System Architecture

Once the chatbot receives the user query, it calls the processQuery() function, which triggers the backend mechanisms to analyze the query contextually. If the chatbot requires additional validation or document retrieval, it interacts with the admin module using the accessKnowledgeBase() function. The admin module acts as a bridge between the chatbot and the institutional document repository, ensuring secure access to relevant SSR documents.

The admin module then performs a Fetch documents operation, where it retrieves the required information from the SSR Documents repository. These documents contain essential institutional data, policies, and accreditation-related details that the chatbot utilizes to generate meaningful responses. Once the documents are located, they are sent back to the admin module, which in turn Provide(s) documents to the chatbot.

After receiving the necessary documents, the chatbot executes generateResponse(), where it formulates a structured, contextually relevant response based on the retrieved information. This response generation process leverages NLP techniques, including text summarization and semantic analysis, ensuring that the user receives precise and meaningful information.

Finally, the chatbot delivers the response back to the user through the receiveResponse() function. The system's architecture ensures that the user receives the most relevant and up-to-date information, minimizing the need for manual searches through lengthy institutional reports. By automating document retrieval and response generation, this architecture enhances efficiency, reduces administrative workload, and improves institutional transparency. Additionally, the chatbot can incorporate a feedback loop mechanism to learn from user interactions, thereby refining future responses and continuously improving its accuracy and effectiveness

5. FRAMEWORK/ALGORITHM

This The chatbot follows a structured workflow to maximize efficiency:

- Document Ingestion & Processing: Converts SSRs and institutional policies into machine-readable text using OCR and preprocessing techniques.
- Semantic Query Analysis: Utilizes transformer-based

models like BERT and GPT to comprehend user intent.

- Search & Retrieval Mechanism: Implements based semantic search using cosine similarity and vector-BM25 ranking.
- Response Generation & Optimization: Employs deep learning techniques to generate precise, human-like responses based on institutional datasets.

5.1 Implementation of llama

The chatbot is powered by meta-llama/Meta-Llama-3-8B-Instruct, a state-of-the-art language model optimized for handling institutional queries. The implementation involved the following steps:

- Model Selection & Fine-tuning: We selected Llama due to its efficiency in processing long-form documents. The model was fine-tuned using SSR datasets to improve domain-specific understanding.
- Integration with Backend: The Llama model was deployed as a microservice using UPSTASH, allowing seamless interaction with the chatbot interface.
- Vectorized Search with Embeddings: We leveraged FAISS for efficient similarity search, enabling quick document retrieval based on user queries.
- Optimization & Caching: Responses were optimized using context-aware memory storage, ensuring faster and more accurate replies.
- Continuous Learning: A feedback loop was implemented to refine the chatbot's responses based on user interactions and new institutional data.

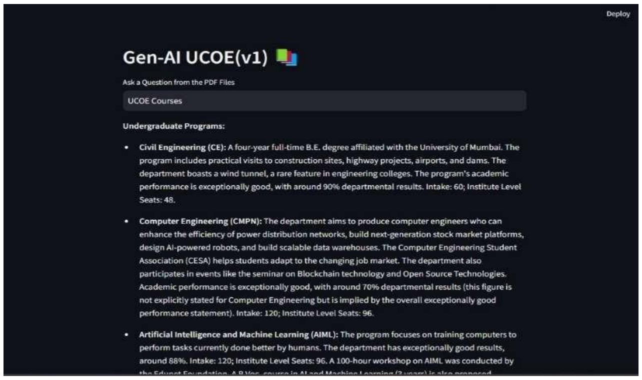
Technical Components:

- Document Parsing: PDF, chunks parsing.
- NLP Models: Meta-Llama-3-8B-Instruct.
- Storage & Indexing: UPSTASH (Redis, vector, and Q-stash).
- Backend: JavaScript and RAG API handling.

6. RESULTS

The institutional chatbot has demonstrated high accuracy and efficiency in retrieving responses from Self-Study Reports (SSR) and institutional documents. It effectively minimizes the need for manual searches by providing precise and structured answers tailored to user queries. The chatbot ensures seamless interaction, maintaining consistent performance even when handling multiple users simultaneously.

Unlike traditional chatbots, which may offer generic or broad responses, this system is specifically designed to focus on SSR-related queries, making it a specialized tool for institutional knowledge management.



Gen-AI UCOE(v1)

Ask a Question from the PDF Files

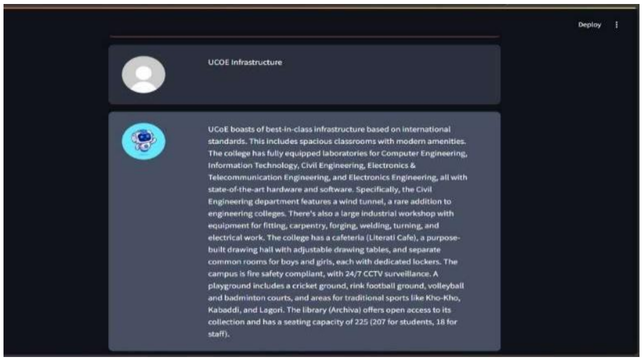
UCOE Courses

Undergraduate Programs:

- **Civil Engineering (CE):** A four-year full-time B.E. degree affiliated with the University of Mumbai. The program includes practical visits to construction sites, highway projects, airports, and dams. The department boasts a wind tunnel, a rare feature in engineering colleges. The program's academic performance is exceptionally good, with around 90% departmental results. Intake: 60; Institute Level Seats: 48.
- **Computer Engineering (CMPN):** The department aims to produce computer engineers who can enhance the efficiency of power distribution networks, build next-generation stock market platforms, design AI-powered robots, and build scalable data warehouses. The Computer Engineering Student Association (CESA) helps students adapt to the changing job market. The department also participates in events like the seminar on Blockchain technology and Open Source Technologies. Academic performance is exceptionally good, with around 70% departmental results (this figure is not explicitly stated for Computer Engineering but is implied by the overall exceptionally good performance statement). Intake: 120; Institute Level Seats: 96.
- **Artificial Intelligence and Machine Learning (AIML):** The program focuses on training computers to perform tasks currently done better by humans. The department has exceptionally good results, around 88%. Intake: 120; Institute Level Seats: 96. A 100-hour workshop on AIML was conducted by the Edunet Foundation. A B.Voc. course in AI and Machine Learning (3 years) is also proposed

Fig – 2: User Query Response

Moreover, the chatbot leverages vectorized search mechanisms to enhance its ability to retrieve semantically relevant results. This improves adaptability to complex queries, ensuring accurate and context-aware responses. The architecture also supports scalability, allowing the system to manage an increasing number of interactions efficiently. Future improvements will focus on further refining response accuracy, expanding the chatbot's dataset, and optimizing its performance for better user experience.



UCOE Infrastructure

UCoE boasts of best-in-class infrastructure based on international standards. This includes spacious classrooms with modern amenities. The college has fully equipped laboratories for Computer Engineering, Information Technology, Civil Engineering, Electronics & Telecommunication Engineering, and Electronics Engineering, all with state-of-the-art hardware and software. Specifically, the Civil Engineering department features a wind tunnel, a rare addition to engineering colleges. There's also a large industrial workshop with equipment for fitting, carpentry, forging, welding, turning, and electrical work. The college has a cafeteria (Literati Cafe), a purpose-built drawing hall with adjustable drawing tables, and separate common rooms for boys and girls, each with dedicated lockers. The campus is fire safety compliant, with 24/7 CCTV surveillance. A playground includes a cricket ground, rink football ground, volleyball and badminton courts, and areas for traditional sports like Kho-Kho, Kabaddi, and Lagori. The library (Archiva) offers open access to its collection and has a seating capacity of 225 (207 for students, 18 for staff).

Fig – 3: Chatbot Response

7. CONCLUSION

The AI-powered institutional chatbot, built on Llama, significantly enhances information accessibility and retrieval within academic institutions. By leveraging NLP and vectorized search, it delivers precise and contextually relevant responses, streamlining administrative and academic workflows. Unlike traditional chatbots, it is specifically tailored to handle institutional queries, ensuring focused and accurate assistance. Its ability to process large volumes of institutional data and efficiently handle multiple users makes it a valuable tool for improving efficiency. Future enhancements will focus on expanding its knowledge base, refining its response accuracy, and integrating additional institutional functionalities to further optimize information management.

ACKNOWLEDEMENT

We extend our heartfelt gratitude to our project guide, Mr. Sandeep Dubey, for his continuous support and insightful guidance throughout this project. We are also thankful to Mrs. Poonam Thakre, Head of the Department of Artificial

Intelligence and Machine Learning, for her encouragement and assistance. Our sincere appreciation goes to Dr. J.B. Patil, Principal of Universal College of Engineering, Vasai, Mumbai, and the management for providing essential resources and a collaborative research environment. Lastly, we acknowledge the support of our peers and faculty members, whose valuable inputs and motivation have greatly contributed to the successful completion of this project.

REFERENCES

[1] R. Vannala and S. B. Swathi, "AI-based Chatbot with Image Recognition for Enhanced User Interaction," International Journal of Artificial Intelligence Research, vol. 10, no. 2, pp. 125-135, 2022.

[2] A. M. Eltahir and H. Abdulla, "Security Challenges in Chatbots: Preventing Data Theft and Spoofing Attacks," Journal of Information Security Research, vol. 15, no. 4, pp. 210-220, 2023.

[3] L. Dai and X. Zhou, "Watermarking Techniques in NLP Models for Preventing Content Misuse," IEEE Transactions on Artificial Intelligence, vol. 6, no. 1, pp. 78-89, 2024.

[4] P. Bansal, M. Srivastava, and R. Gupta, "Enhancing Academic Document Retrieval using Transformer-Based Models," IEEE Access, vol. 8, pp. 213457-213469, 2021.

[5] H. Chen, K. Lin, and X. Zhang, "Hybrid Search Algorithms for Efficient Document Retrieval in AI Chatbots," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 5, pp. 900-915, 2023.

[6] T. Li and Y. Wu, "Memory-Optimized Chatbots: Enhancing NLP Responses with Contextual Caching," Neural Processing Letters, vol. 54, no. 2, pp. 2901-2918, 2022.

[7] S. Kim, J. Lee, and R. Park, "Adaptive Learning for AI Chatbots: A Reinforcement Learning-Based Approach," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 112-125, 2023.