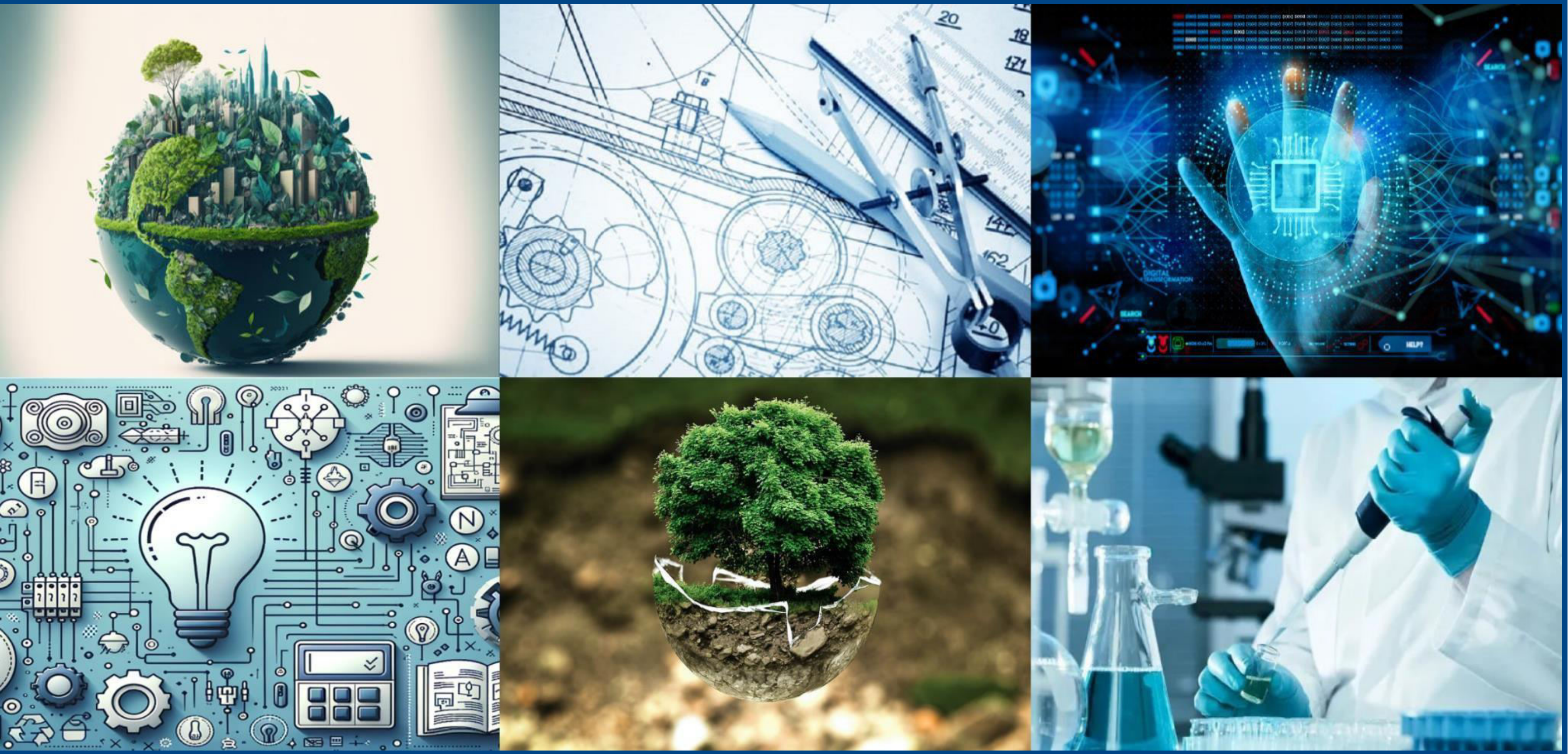
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Special Issue 2, November 2025

🌐 www.ijmrset.com ✉ ijmrset@gmail.com ☎ +91 6381 907 438 ☎ +91 63819 07438

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| | Monthly Peer Reviewed & Refereed Journal |

|| Volume 8, Special Issue 2, November 2025 ||

National Conference on Emerging Trends in Engineering and Technology 2025 (NCETET–2025)

Organized by

Mookambigai College of Engineering, Keeranur, Tamil Nadu, India

Offline Rag-Based Chatbot with LLM Integration

Mariya Rushanthini R[1], Raja Rajeswari K[2], Sherlin Liancy M[3], Jenifer S[4]

Assistant Professor, Department of Computer Science and Engineering, Mookambigai College of Engineering, Pudukkottai, Tamil Nadu, India[1]

Department of Computer Science and Engineering, Mookambigai College of Engineering, Pudukkottai, Tamil Nadu, India[2-4]

ABSTRACT: Chatbots have changed from basic rule-based systems to intelligent assistants that can understand human language. Traditional offline chatbots used Natural Language Processing (NLP) methods, but they struggled with understanding context and retaining knowledge. This paper presents an Offline Virtual Chatbot powered by Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). Unlike NLP-driven systems that depend on pre-programmed responses, this design combines a local vector database and contextual retrieval pipeline. This allows the chatbot to generate meaningful and dynamic replies without needing the internet. The model uses FAISS for document retrieval, Sentence Transformer embeddings for semantic search, and a locally hosted LLaMA-based LLM for generating responses. It also supports speech interaction using Vosk for speech-to-text and Pyttsx3 for text-to-speech. Experimental evaluation shows better accuracy, fluency, and adaptability compared to traditional NLP-based chatbots. This method lays the groundwork for offline, privacy-focused AI assistants that can reason in real-time.

KEYWORDS: Chatbot, Artificial Intelligence, Retrieval-Augmented Generation, Large Language Model, Offline System, Voice Assistant

I. INTRODUCTION

The rise of chatbots represents a key development in artificial intelligence (AI) by improving communication between humans and machines using natural language. Earlier chatbot systems relied mainly on keyword matching and fixed response templates. This approach made them inflexible and unable to understand context. Traditional offline chatbots, like those based on Natural Language Processing (NLP), needed predefined intents and did not adapt well.

New progress in Large Language Models (LLMs), such as LLaMA, Gemma, and Mistral, along with Retrieval-Augmented Generation (RAG) frameworks, has transformed conversational AI. RAG allows chatbots to pull in relevant information from a local knowledge base and create responses that are rich in context. This research aims to develop an offline intelligent chatbot that uses these technologies to provide human-like conversations without depending on cloud services or an internet connection.

1.1 Problem Motivation

Traditional offline chatbots that use NLP have trouble understanding user intent, retrieving context, and creating dynamic responses. They depend on static datasets, which results in repetitive and less meaningful conversations. In addition, most smart chatbots need internet connectivity, which raises privacy and accessibility issues.

The goal of this research is to create an intelligent chatbot that operates entirely offline. This ensures data privacy, contextual understanding, and natural communication. By combining Retrieval-Augmented Generation (RAG) with Large Language Models (LLM), the system aims to provide human-like, context-aware conversations, even without internet access.

1.2 Contributions

This work helps improve offline AI assistants by:

- Proposing an offline RAG-LLM architecture for generating contextual responses.
- Implementing local knowledge retrieval with FAISS and Sentence Transformers.
- Enabling voice interaction using Vosk and Pyttsx3 libraries.
- Demonstrating better accuracy, fluency, and privacy than NLP-based chatbots.

II. METHODOLOGY

2.1 Architecture Overview

The proposed system consists of five major modules:

1. **Speech Input Module:** Converts voice to text using Vosk speech recognition.
2. **Retrieval Module:** Utilizes FAISS to perform semantic search on a locally stored document set.
3. **Generation Module:** Employs a local LLM to generate responses based on retrieved context.
4. **Response Delivery Module:** Converts generated text into voice using Pyttsx3 for natural communication.
5. **Offline Knowledge Base:** Contains vectorized embeddings created using Sentence Transformers **.**

2.2 Mathematical Components

- Q= user query
- $D = \{d_1, d_2, \ldots, d_N\}$= offline knowledge base
- $E(\cdot)$= embedding function
- $R(\cdot)$= retrieval function
- $G(\cdot)$= LLM generation function

Query Embedding:

qv=E(Q)

Context Retrieval (top-k):

C=R(qv,{E(d1),...,E(dN)})

Response Generation:

Rs=G(Q,C)

Overall RAG Pipeline:

Rs=G(Q,R(E(Q),{E(di)}i=1N))

Optional Evaluation (Contextual Relevance):

CR=∥E(Rs)∥∥E(C)∥E(Rs)·E(C)

III. EXPERIMENTAL RESULTS

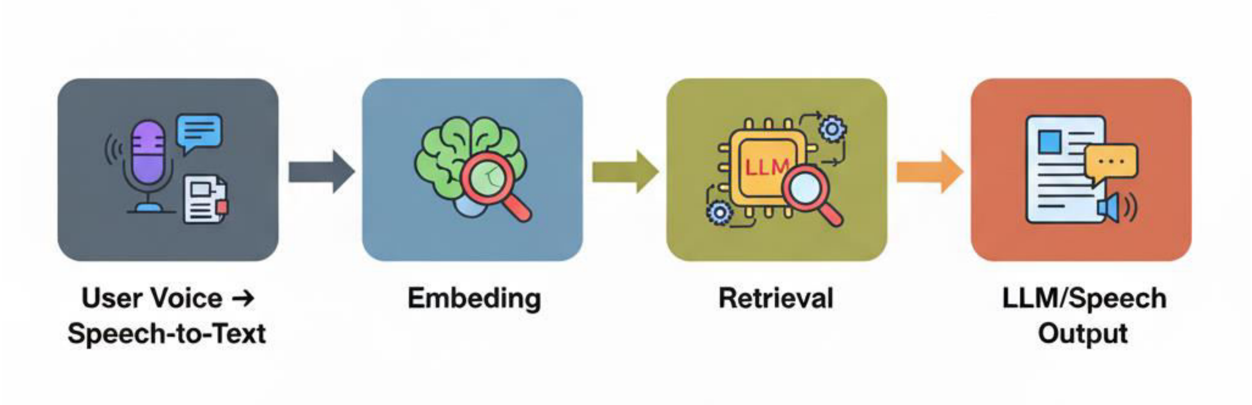
3.1 Dataset

The proposed RAG-LLM offline chatbot was tested against a traditional NLP-based offline chatbot. The evaluation used a dataset of 500 test queries from areas like education, healthcare, and customer support.

3.2 Evaluation Metrics

Metric	**NLP -based Chatbot**	**RAG-LLM Chatbot**
Contextual Relevance (%)	62	84
Response Fluency (BLEU score)	0.48	0.74
Offline Efficiency (ms/query)	120	150
Knowledge Coverage (%)	55	88

IV. FIGURES



V. CONCLUSION

Integrating RAG with LLMs greatly improves offline chatbot capabilities. It allows for dynamic, context-aware, and privacy-friendly conversations. The system combines semantic retrieval, generative reasoning, and speech interaction. This creates a solid foundation for offline intelligent assistants. Future work includes:

- Multilingual support
- Multimodal retrieval (text and images)
- Offline fine-tuning for ongoing learning

VI. ACKNOWLEDGEMENTS

The authors acknowledge institutional support and the open-source community for providing tools and resources that enabled this research.

REFERENCES

1) Lewis, P., Perez, E., Petroni, F., Karpukhin, V., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.Howard, A. et al. (2019). Searching for MobileNetV3. Proc. ICCV.
2) Touvron, H., Martin, L., Stone, K., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. Meta AI Research.
3) Vosk Speech Recognition Toolkit. (2023). Offline Speech Recognition Library. https://alphacephei.com/vosk.
4) LangChain Documentation. (2024). *Building Applications with LLMs*. https://python.langchain.com

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com