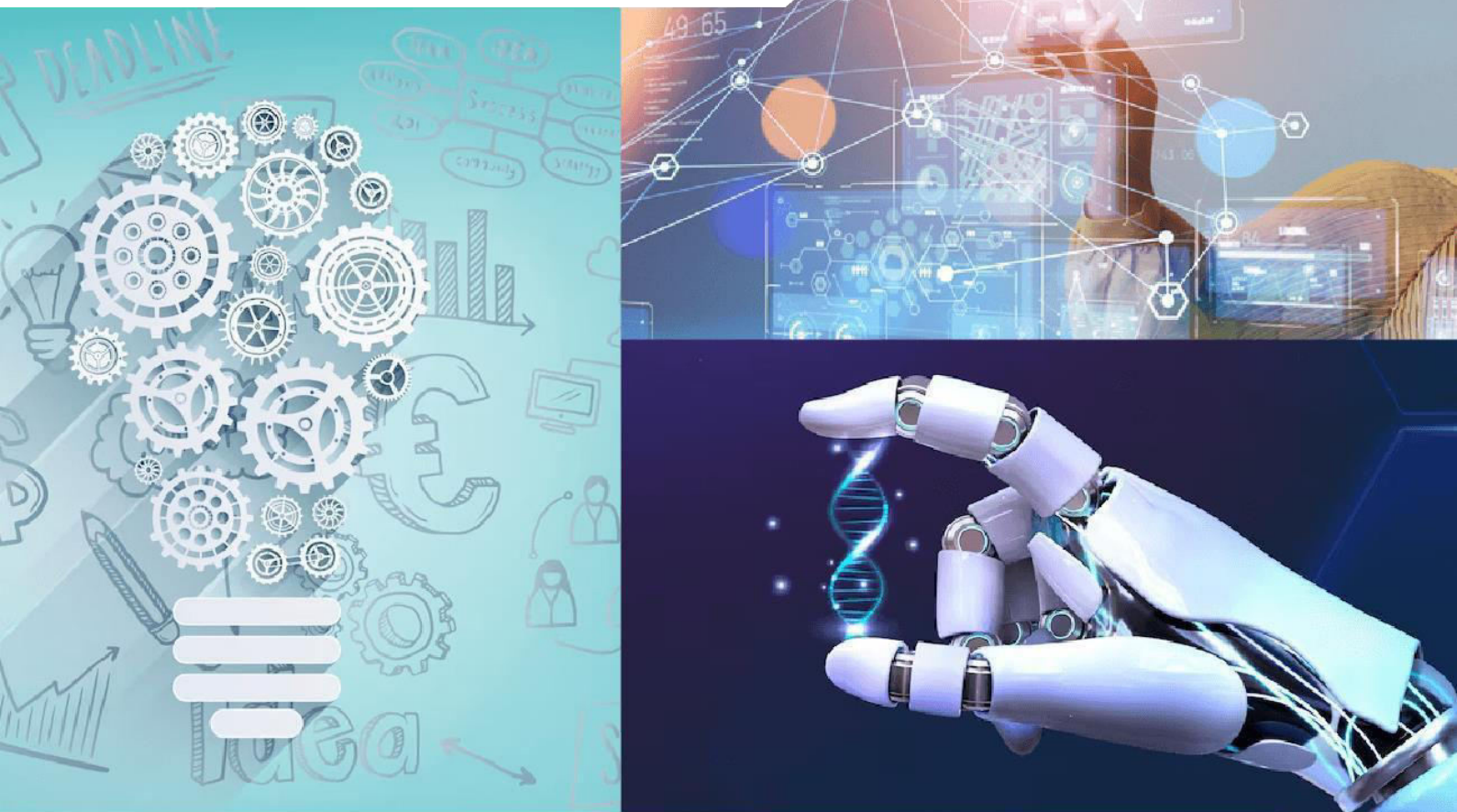




# **International Journal of Advanced Research in Education and Technology (IJARETY)**

**Volume 13, Issue 3, May - June 2026**

**Impact Factor: 9.971**



# Language Agnostic Chatbot: A Retrieval-Augmented Voice-Enabled Assistant for Academic Information Management

Manasa L

Dept. of Master of Computer Applications, CMR Institute of Technology, Bengaluru, India

**ABSTRACT**-Educational institutions handle a large amount of academic and administrative information every day, including examination schedules, fee notifications, scholarship announcements, placement updates, hostel notices, and academic calendars. Even though colleges and universities have adopted digital communication systems, students still face difficulty finding accurate information quickly because updates are spread across multiple platforms such as websites, emails, student portals, notice boards, and messaging applications. This fragmented communication structure often creates confusion, delays, and repeated dependency on administrative departments.

This paper presents Language Agnostic Chatbot, a language-agnostic academic chatbot developed using Retrieval-Augmented Generation (RAG). The proposed system retrieves information directly from institutional documents before generating responses, improving reliability and reducing unsupported outputs. Language Agnostic Chatbot supports both text and voice interaction and can process scanned documents using Optical Character Recognition (OCR).

The framework combines OCR, semantic vector search, transformer-based embeddings, and Large Language Models to generate context-aware responses [2], [3], [8]. ChromaDB is used for vector storage, while FastAPI is used for backend services [5], [6]. The system also supports multilingual communication, improving accessibility for users communicating in regional languages [9]. Experimental evaluation demonstrates stable performance, low response latency, and accurate retrieval of institutional information. The proposed framework reduces repetitive administrative workload and improves accessibility to academic information in educational environments.

**KEYWORDS:** Educational Chatbot, Retrieval-Augmented Generation, OCR, Semantic Search, Multilingual AI, Smart Campus.

## I. INTRODUCTION

Educational institutions continuously generate and manage a large amount of information related to examinations, internal assessments, fee payments, placements, hostel notices, scholarships, and academic schedules. Although many institutions have shifted towards digital communication methods, students still have trouble accessing accurate information at the right time. Important updates are distributed across websites, emails, student portals, learning management systems, and messaging applications [12].

Because information is spread across different communication platforms, students often become confused about recent notifications and deadlines. Many students repeatedly contact administrative offices to confirm examination schedules, fee payment dates, or procedural instructions. Administrative staff also spend a significant amount of time handling repetitive queries [14].

Earlier chatbot systems mainly depended on predefined rules and fixed response templates [13]. Such systems lacked flexibility and were unable to understand unexpected queries effectively. Recent developments in Large Language Models improved conversational systems through better contextual understanding and natural language generation [7]. However, standalone language models sometimes produce unsupported or inaccurate responses when domain-specific context is unavailable [11]

Retrieval-Augmented Generation improves reliability by combining semantic retrieval with language generation [2]. Instead of producing responses only from pretrained knowledge, the system first retrieves relevant information from institutional documents and then generates context-aware responses. This significantly reduces hallucination and improves factual accuracy.

Another challenge in educational environments is that many important records are available only as scanned PDFs or image-based documents. Optical Character Recognition therefore becomes necessary to convert scanned content into machine-readable text [3]. Multilingual accessibility is also important because students may prefer interacting using regional languages [9].

To address these challenges, this work proposes Language Agnostic Chatbot, a multilingual and voice-enabled academic assistant that integrates OCR, semantic retrieval, and Retrieval-Augmented Generation to provide reliable institutional information.

## **II. LITERATURE SURVEY**

Educational chatbot systems have evolved significantly over the years. Earlier systems mainly used rule-based approaches where responses were predefined using patterns and keywords. Although these systems could handle simple queries, they lacked adaptability and failed to understand language variations effectively [1].

Transformer-based models such as BERT and GPT introduced major improvements in conversational AI through contextual understanding and natural language generation [4], [7]. These models demonstrated strong performance in generating human-like responses and understanding user intent. However, standalone language models frequently generate unsupported information when reliable context is unavailable [11].

Retrieval-Augmented Generation was introduced to improve factual reliability in conversational systems [2]. RAG combines semantic retrieval with language generation so that responses are generated using retrieved documents instead of assumptions [10]. This approach significantly reduces hallucination and improves trustworthiness. Researchers have also explored OCR techniques for digitizing academic records [3].

OCR systems convert scanned documents into searchable text, improving accessibility and document retrieval. Voice-based communication and multilingual interaction are becoming increasingly important in educational systems because they improve accessibility for users who are more comfortable communicating through speech or regional languages [9], [15].

Although many studies focus on OCR, semantic retrieval, or conversational AI individually, only limited work combines all these technologies into a unified academic assistant. Language Agnostic Chatbot addresses this gap through an integrated architecture.

## **III. SYSTEM ARCHITECTURE**

Language Agnostic Chatbot follows a layered architecture consisting of Presentation Layer, Application Layer, Intelligence Layer, and Data Layer.

### **3.1 Architectural Overview**

The overall architecture of Language Agnostic Chatbot is shown in Figure 1. Users interact with the chatbot using a web-based interface that supports both text and voice communication. User requests are processed through backend services where OCR, semantic retrieval, and language generation modules work together to generate responses.

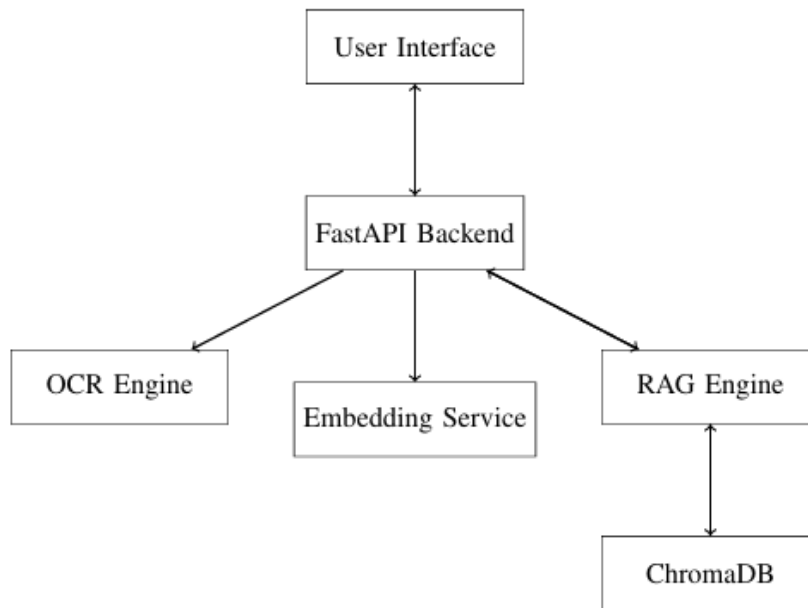


Fig1. Language Agnostic Chatbot System Architecture

**3.2 Presentation Layer**

The presentation layer provides a user-friendly interface where students can interact with the chatbot using text or voice queries. Responses are displayed in both textual and speech formats.

**3.3 Application Layer:**

The application layer handles authentication, request processing, API communication, and document upload. FastAPI is used because of its lightweight architecture and efficient asynchronous request handling [5].

**3.4 Intelligence Layer:**

The intelligence layer performs the core AI operations of the system. It includes:

- OCR Service
- Embedding Generation
- Retrieval Engine
- Language Model Integration

OCR converts scanned documents into machine-readable text. Transformer-based embedding models convert document chunks and queries into semantic vectors [8]. ChromaDB performs similarity search to retrieve the most relevant document content.

**3.5 Data Layer:**

The data layer stores uploaded documents, embeddings, and metadata. SQLite is used for relational storage, while ChromaDB manages vector embeddings for semantic retrieval [6].

**IV. PROPOSED METHODOLOGY**

The proposed system follows multiple stages including document preprocessing, semantic indexing, retrieval, and response generation.

**4.1 Document Processing**

Institutional documents are uploaded in PDF or image format. Text-based PDFs are directly processed, while scanned files undergo OCR processing. Preprocessing includes:

- Removing unnecessary spaces
- Cleaning headers and footers
- Correcting OCR extraction errors

- Unicode normalization

These preprocessing steps improve text quality and retrieval accuracy.

#### 4.2 Text Chunking

Documents are divided into overlapping chunks to preserve contextual continuity. The system uses:

- Chunk size: 400 tokens
- Overlap: 20%

Overlapping improves semantic retrieval performance.

#### 4.3 Embedding Generation

Transformer-based embedding models convert text chunks into semantic vector representations.

$$v_i = f(c_i)$$

where:

- $c_i$  represents document chunks
- $v_i$  represents embeddings

The generated embeddings are stored in ChromaDB along with metadata.

#### 4.4 Query Retrieval

When a user submits a query, the query is converted into a semantic embedding and compared with stored vectors using cosine similarity.

$$\text{Similarity} = (v_q \cdot v_i) / (|v_q| |v_i|)$$

The most relevant document chunks are retrieved and passed to the language model.

#### 4.5 Voice Interaction

The chatbot supports voice interaction using:

- Automatic Speech Recognition (ASR)
- Text-to-Speech (TTS)

Speech input is converted into text, processed through the retrieval pipeline, and converted back into speech output [15].

#### 4.6 Advanced System Workflow

The workflow of Language Agnostic Chatbot is designed to ensure accurate and efficient query processing. The process begins when institutional documents are uploaded into the system. These documents may include examination notifications, fee circulars, placement announcements, scholarship guidelines, and academic schedules. Scanned documents undergo OCR processing to convert image-based content into searchable text. The extracted content is then cleaned and pre-processed to remove formatting artifacts and OCR errors. The cleaned documents are divided into overlapping chunks. Each chunk is converted into semantic vector representations using transformer-based embedding models. These vectors are stored in ChromaDB for efficient semantic retrieval. When users submit queries, the system converts the queries into embeddings and performs cosine similarity search to identify relevant document chunks. The retrieved content is passed to the language model, which generates responses based only on retrieved institutional information. The final response is delivered to the user through text or speech output.

#### 4.7 Retrieval-Augmented Generation Framework

Retrieval-Augmented Generation forms the core intelligence mechanism of Language Agnostic Chatbot. Traditional conversational systems generate responses using only pretrained model knowledge. However, institutional information changes frequently, making static models unreliable for academic environments.

RAG solves this issue by separating document retrieval from language generation. The retrieval module searches institutional documents and identifies relevant information, while the generation module produces responses using retrieved context.

This approach provides several advantages:

- Improved factual reliability
- Reduced hallucination
- Real-time document updates
- Context-aware retrieval
- Scalability for large repositories

The retrieval mechanism uses semantic similarity instead of exact keyword matching, allowing the chatbot to understand different ways of asking the same question.

#### **4.8 Optical Character Recognition Integration**

Many institutional records exist only as scanned PDFs or image-based documents. Traditional search systems cannot retrieve information effectively from such files because the content is not machine-readable. Language Agnostic Chatbot integrates OCR into the document processing pipeline to solve this issue [3]. OCR extracts textual information from scanned academic documents and converts it into searchable digital text.

- Image preprocessing
- Noise reduction
- Character extraction
- Text normalization
- Error correction

The extracted content is validated before being added to the semantic retrieval system.

#### **4.9 Semantic Embedding and Vector Search**

Semantic retrieval is one of the most important components of Language Agnostic Chatbot. Traditional keyword-based search systems often fail when users phrase queries differently. Transformer-based embedding models convert text into semantic vectors [8]. These embeddings capture contextual meaning instead of exact words. For user query  $q$ :

$$v_q = f(q)$$

Cosine similarity is used to measure semantic relevance:

$$\text{Similarity} = (v_q \cdot v_i) / (|v_q| |v_i|)$$

Higher similarity values indicate stronger contextual relevance between user queries and document content.

#### **4.10 Voice-Based Interaction**

Voice interaction improves accessibility and usability in educational environments. Many students prefer speech-based interaction while using mobile devices. Language Agnostic Chatbot integrates both Automatic Speech Recognition and Text-to-Speech technologies [15].

The workflow includes:

- Capturing speech input
- Converting speech into text
- Processing semantic retrieval
- Generating responses
- Converting responses back into speech

Voice support improves accessibility for users with limited typing convenience.

### **V. RESULTS AND DISCUSSION**

The proposed system was evaluated using response accuracy, latency, multilingual support, and scalability metrics.

#### **5.1 Experimental Setup**

Experiments were conducted on a local system with:

- Intel Core i5 Processor
- 16GB RAM

A total of 300 queries were tested

- 60% English
- 20% Hindi
- 20% Kannada

## 5.2 Performance Metrics

TABLE I  
SYSTEM PERFORMANCE METRICS

| Operation              | Average Time |
|------------------------|--------------|
| Login Authentication   | 150 ms       |
| Text Query Processing  | 3.2 s        |
| Voice Query Processing | 5.5 s        |
| OCR Processing         | 12 s         |

The system maintained stable performance during concurrent access and generated responses within acceptable latency limits.

Out of 300 test queries, 282 queries successfully retrieved relevant context, resulting in a retrieval accuracy of 94%. Compared with standalone language models, the proposed RAG-based system significantly reduced unsupported responses by generating answers only from retrieved institutional documents [11]. The chatbot successfully processed English, Hindi, and Kannada queries. The multilingual embedding model improved semantic understanding across different languages. The system successfully handled concurrent requests without significant performance degradation. FastAPI asynchronous request handling improved concurrency management. The results demonstrate that Language Agnostic Chatbot provides reliable and efficient academic assistance. The integration of OCR, semantic retrieval, multilingual support, and voice interaction makes the system suitable for educational institutions.

## VI. CONCLUSION

This paper presented Language Agnostic Chatbot, a language-agnostic academic assistant developed using Retrieval-Augmented Generation, OCR, and voice-based interaction techniques. The proposed system retrieves information directly from institutional documents and generates context-aware responses. By grounding responses in retrieved content, the chatbot reduces unsupported answers and improves reliability.

Experimental evaluation demonstrated strong retrieval accuracy, multilingual support, and stable response performance. The system also reduced repetitive administrative workload and improved accessibility for students. Future improvements may include integration with Student Information Systems, analytics dashboards, messaging platforms such as WhatsApp and Telegram, and enhanced multilingual optimization. Overall, Language Agnostic Chatbot provides a practical and scalable solution for academic information management in modern educational environments.

## REFERENCES

- [1] Wollny et al., "Chatbots in Education: A Systematic Review," IEEE Access, 2021.
- [2] Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
- [3] Singh and Kaur, "Optical Character Recognition in Document Digitization," International Journal of Computer Applications, 2022.
- [4] Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
- [5] FastAPI Documentation, 2024.
- [6] ChromaDB Documentation, 2024.
- [7] Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.
- [8] Reimers and Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," EMNLP, 2019.
- [9] Zhang et al., "Multilingual Neural Machine Translation," ACL, 2020.
- [10] Chen et al., "A Survey on Retrieval-Augmented Text Generation," arXiv, 2023.
- [11] Huang et al., "Reducing Hallucination in Large Language Models via Retrieval," ACL Findings, 2022.
- [12] Kumar et al., "Smart Campus Systems: Architecture and Applications," IEEE Internet of Things Journal, 2021.
- [13] Al-Zubaidi and Issa, "OntBot: Ontology Based Chatbot," ISCA, 2011.
- [14] Adamopoulou and Moussiades, "Chatbots: History, Technology, and Applications," Machine Learning with Applications, 2020.
- [15] Raj et al., "Voice-Enabled Conversational Agents for Higher Education," IEEE EDUCON, 2023.



# International Journal of Advanced Research in Education and Technology

**ISSN: 2394-2975**

**Impact Factor: 9.971**