World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Multilingual AI chatbot using transformers

Payoshni Sanjay Gade [1,*] and S. S. Dhande [2]

[1] *MTech, Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati.*
[2] *Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati.*

World Journal of Advanced Research and Reviews, 2026, 30(01), 2517-2525

Publication history: Received on 16 March 2026; revised on 26 April 2026; accepted on 28 April 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.1.1153

Abstract

The rapid advancement of Natural Language Processing (NLP) has enabled the development of intelligent conversational agents capable of interacting in multiple languages. This paper presents a scalable and efficient multilingual AI chatbot architecture that integrates transformer-based language detection with large language model (LLM)-driven response generation. The proposed system employs a Bidirectional Encoder Representations from Transformers (BERT) model for accurate language identification, followed by a Retrieval-Augmented Generation (RAG) pipeline powered by LLaMA 3 via Groq for response generation. To enhance contextual relevance, LangChain is utilized for orchestration, while Qdrant serves as the vector database for semantic retrieval. Mistral-based embedding models are used to convert textual data into dense vector representations, enabling efficient similarity search across multilingual corpora. The frontend is developed using React and Tailwind CSS, while the backend leverages Python for model integration and API handling.

The system aims to provide accurate, context-aware, and language-specific responses across diverse linguistic inputs. Experimental observations suggest that combining transformer-based language detection with modern LLMs significantly improves chatbot performance in multilingual environments. This architecture is particularly suitable for real-world applications such as customer support, education, and cross-lingual communication systems.

Keywords: Multilingual Chatbot; BERT; Llama 3; Langchain; Qdrant; Mistral Embeddings; Transformer Models; NLP; RAG

1. Introduction

With the increasing globalization of digital platforms, the demand for multilingual conversational systems has grown significantly. Traditional chatbots are limited by rule-based systems or monolingual capabilities, making them inadequate for diverse user bases. Transformer-based models such as BERT have revolutionized NLP by enabling contextual understanding across multiple languages. Multilingual transformer models can generalize across languages due to shared representations and subword tokenization techniques (). Recent advancements in Large Language Models (LLMs), such as LLaMA 3, have further improved conversational AI by enabling human-like dialogue, reasoning, and contextual awareness. Additionally, Retrieval-Augmented Generation (RAG) architectures combine LLMs with external knowledge sources, improving factual accuracy and reducing hallucination.

1.1. Motivation

Despite significant progress, being chatbot systems face several challenges

- Inaccurate language discovery in multilingual surroundings

* Corresponding author: Payoshni Sanjay Gade

- Lack of contextual understanding across languages
- Inefficient reclamation of applicable information
- Poor scalability in real- world operations

Multilingual BERT has shown strong performance in language understanding tasks, particularly in medium- and high-resource languages (). still, standalone models are inadequate for erecting complete conversational systems.

thus, there's a need for an intertwined system that combines

- Accurate language discovery
- environment- apprehensive response generation
- Effective semantic reclamation
- Scalable system armature

This motivates the development of a mongrel armature combining BERT, LLaMA 3, LangChain, and vector databases.

Objective

- To design a multilingual chatbot able of handling multiple languages efficiently
- To apply a BERT- grounded model for high- delicacy language discovery
- To integrate LLaMA 3 for advanced conversational response generation
- To use LangChain for orchestrating RAG channels
- To apply Qdrant as a vector database for semantic hunt
- To employ Mistral embedding models for generating high- quality vector representations
- To develop a scalable full- mound system using Python(backend) and Reply with Headwind CSS (frontend).

2. Literature review

Recent advancements in multilingual conversational systems have been driven by the integration of natural language processing (NLP), transformer-based architectures, and large language models (LLMs). Early work by Vanjani et al. (2019) established the foundational framework for multilingual chatbots, emphasizing the importance of cross-lingual interaction and language adaptability in conversational agents. This study highlighted key challenges such as semantic alignment across languages and maintaining contextual coherence.

Galadima and Lawrence (2024) extended this paradigm by proposing a multilingual conversational AI system for educational consultations, demonstrating the effectiveness of domain-specific dialogue systems in improving user engagement and information accessibility. Similarly, Orosoo et al. (2024) focused on enhancing cross-cultural communication through NLP techniques, emphasizing the role of linguistic diversity and contextual sensitivity in chatbot design.

Recent studies have increasingly leveraged transformer-based architectures for improved performance. Shrivastava et al. (2025) provided a comprehensive overview of NLP techniques for conversational AI, highlighting the transition from traditional sequence models to transformer-based models such as BERT and GPT. Cotfas et al. (2025) further analyzed the evolution of large language models, demonstrating how transformer architectures enable scalable, context-aware, and high-quality text generation.

The introduction of large-scale pretrained models, such as GPT-3 (Brown et al., 2020), marked a significant milestone by enabling few-shot learning capabilities, reducing the dependency on large labeled datasets. Ouyang et al. (2022) further improved these models through Reinforcement Learning from Human Feedback (RLHF), enhancing alignment with human intent and response quality.

In the context of multilingual chatbot applications, Pattanayak et al. (2025) proposed ChatSense, a system designed for multilingual interaction, showcasing the integration of language detection and response generation modules. Munjal et al. (2025) explored the application of multilingual chatbots in healthcare, demonstrating their potential in providing accessible and scalable virtual assistance across diverse linguistic populations.

Additionally, research by Gurioli et al. (2025) and Al Bataineh et al. (2025) addressed challenges related to AI-generated content, including authorship detection and benchmarking of human versus machine-generated text. These studies contribute to the evaluation and reliability of LLM-based systems.

Overall, the literature indicates a clear shift toward transformer-based multilingual conversational systems that integrate retrieval mechanisms, language detection models, and LLMs to achieve high accuracy, scalability, and contextual understanding. However, challenges such as integration complexity, cross-lingual semantic consistency, and evaluation metrics remain active areas of research.

3. Methodology

3.1. Overview

The proposed methodology focuses on designing and implementing a multilingual AI chatbot by integrating transformer-based models with a Retrieval-Augmented Generation (RAG) framework. The system combines language detection, semantic retrieval, and large language model-based response generation to provide accurate and context-aware answers in multiple languages. The methodology is divided into several stages, including data preparation, model training, embedding generation, retrieval, and response generation**.**

3.2. Data Collection and Preprocessing

A multilingual dataset is prepared for the proposed system, consisting of text samples in multiple languages such as English, Hindi, and Marathi. This dataset serves as the foundation for training the language detection model. Before training, the dataset undergoes several preprocessing steps to ensure data quality and consistency. These steps include cleaning the text by removing noise, special characters, and any irrelevant information that may affect model performance.

Furthermore, text normalization is applied, which involves converting all text to lowercase and removing punctuation to maintain uniformity. The data is then tokenized to break down sentences into smaller units suitable for model input. For the purpose of language classification, the language labels are encoded into numerical form using label encoding techniques. Finally, the dataset is divided into training and testing sets to evaluate the model's performance effectively. The processed dataset is subsequently used to train the BERT model for accurate language detection in the multilingual chatbot system.

3.3. Language Detection using BERT

A transformer-based BERT model is utilized in this system to perform language detection on user input. The pre-trained multilingual BERT model is fine-tuned on the prepared dataset to handle the task of language classification effectively. By leveraging its bidirectional context understanding, the model is able to accurately identify the language of a given text.

During execution, the model takes the user query as input and outputs the corresponding predicted language label. This step is crucial for ensuring that the system correctly interprets multilingual queries and processes them appropriately in subsequent stages of the chatbot pipeline.

3.4. Embedding Generation

To enable semantic understanding within the system, both user queries and documents are converted into dense vector representations using Sentence Transformers. These embeddings capture the contextual and semantic meaning of the text, rather than relying solely on traditional keyword-based matching approaches.

By representing text in this vectorized form, the system can effectively measure the similarity between different pieces of information. This step is essential for enabling accurate similarity-based retrieval in the Retrieval-Augmented Generation (RAG) pipeline, thereby improving the relevance and quality of the chatbot's responses.

3.5. Vector Database and Storage

The generated embeddings of documents are stored in a vector database, specifically Qdrant, which is designed for efficient handling of high-dimensional vector data. This database enables fast and scalable storage as well as retrieval of embeddings, making it well-suited for semantic search applications.

Each document in the dataset is indexed along with its corresponding vector representation, allowing the system to perform rapid similarity-based searches during query processing. This facilitates the retrieval of the most relevant

documents based on the semantic similarity between the user query and stored data, thereby enhancing the overall performance of the chatbot system.

3.6. Retrieval-Augmented Generation (RAG)

The Retrieval-Augmented Generation (RAG) framework is implemented using LangChain to enhance the quality and accuracy of the generated responses. In this approach, when a user submits a query, it is first transformed into a dense vector representation using an embedding model. This query embedding is then compared with precomputed embeddings stored in the vector database to perform similarity-based retrieval.

Based on similarity scores, the system retrieves the top-*k* most relevant documents, which serve as contextual knowledge for the query. These retrieved documents are subsequently combined with the original user input to form an enriched prompt. This augmented input is then provided to the language model for response generation.

By incorporating relevant external context into the generation process, the RAG approach significantly improves factual accuracy, contextual relevance, and overall response quality of the chatbot system.

3.7. Response Generation using LLaMA 3

The final response in the proposed system is generated using the LLaMA 3 large language model, accessed via the Groq API. The model receives an enriched input consisting of the original user query combined with the retrieved contextual information obtained through the Retrieval-Augmented Generation (RAG) process. By leveraging this augmented input, the model generates responses that are coherent, contextually relevant, and semantically meaningful.

The integration of the Groq API enables high-speed inference, significantly reducing response latency. This ensures efficient real-time performance of the chatbot system, making it well-suited for interactive conversational applications.

3.8. System Integration

All components are integrated into a unified pipeline using Django and LangChain. The workflow of the system is as follows:

- User inputs a query
- BERT detects the language
- Query is converted into embeddings
- Relevant documents are retrieved using RAG
- Context is appended to the query
- LLaMA 3 generates the response
- Response is displayed to the user

This pipeline ensures smooth interaction between different modules and efficient processing of multilingual queries.

3.9. Model Training Environment

The BERT model is trained using the Kaggle platform, which provides GPU support for faster computation. The training process includes optimization using cross-entropy loss and evaluation using performance metrics such as accuracy and F1-score.

4. Implementation

The system processes user input through a structured pipeline in which the query is first analyzed using a BERT-based model to detect the input language and is subsequently preprocessed accordingly. The processed query is then forwarded to the embedding and retrieval module for contextual information extraction, after which the enriched input is passed to the LLaMA 3 model to generate the final response.

Steps:

- User enters query
- BERT detects language
- Query is processed accordingly

- Sent to embedding and retrieval module
- Final response generated by LLaMA 3

4.1. System Architecture Overview

The proposed multilingual AI chatbot system follows a layered, modular architecture integrating transformer-based models with a Retrieval-Augmented Generation (RAG) pipeline. The Client Layer consists of a React-based frontend (TypeScript/Vite) that communicates with the backend via HTTP/REST protocols. Requests are handled by the API Gateway, implemented using the Django REST Framework, which orchestrates routing and manages interactions between system components.

The Processing Layer comprises three core modules: (i) a BERT-based language detection model that identifies the input language, (ii) a RAG service that retrieves relevant contextual information from a vector database using embedding-based similarity search, and (iii) an LLM service (via Groq API) that generates context-aware responses by combining user queries with retrieved knowledge. The language detection output guides the downstream processing to ensure multilingual adaptability.

The Storage Layer includes a relational SQLite database for structured data management, document storage for knowledge base persistence, and a vector embeddings store for efficient semantic retrieval. CRUD and document management operations are supported through dedicated API routes.

The External Services Layer integrates third-party model providers, including Groq-hosted LLMs for inference and Hugging Face models for loading pretrained components such as BERT. Overall, the architecture ensures scalable, language-agnostic, and context-aware conversational capabilities through tight integration of retrieval mechanisms and transformer-based generation.

4.2. System Block Diagram High Level Architecture

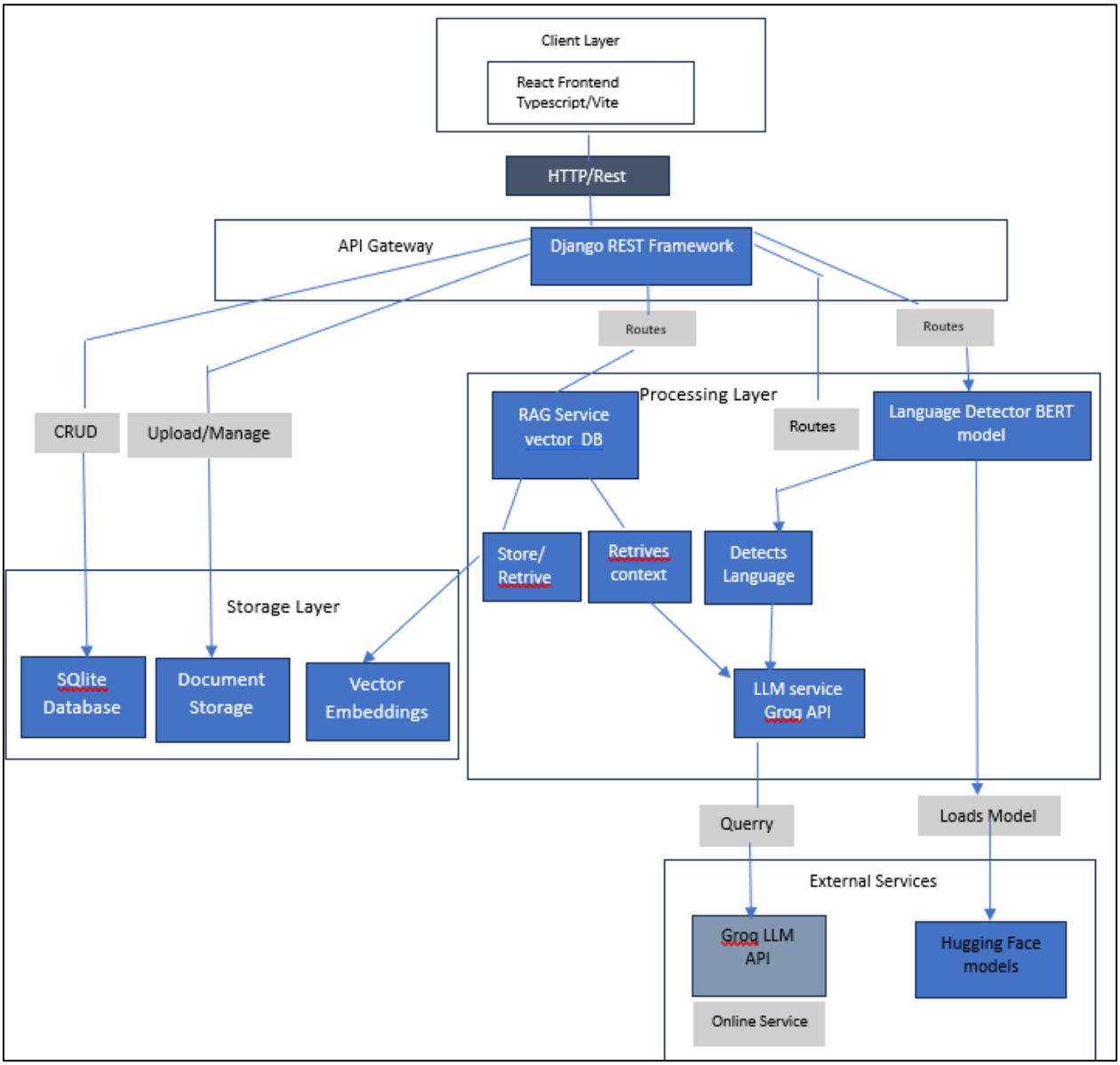

Figure 1 System Block Diagram High Level Architecture

5. Result and Discussion

The proposed multilingual AI chatbot was evaluated on its ability to accurately detect and process queries across five different languages, namely English, Hindi, Marathi, Spanish, and French. The performance of the system was analyzed at multiple stages, including language detection, semantic retrieval, and response generation.

The BERT-based language detection model demonstrated high classification accuracy across all five languages, effectively identifying the input language even in the presence of short and informal queries. The use of a multilingual dataset contributed to improved generalization and robustness of the model.

The embedding generation using Sentence Transformers enabled efficient semantic representation of both queries and documents. This, in combination with the Qdrant vector database, resulted in accurate retrieval of contextually relevant documents. The similarity-based retrieval mechanism consistently returned top-*k* relevant results, which significantly enhanced the contextual grounding of responses.

The integration of the Retrieval-Augmented Generation (RAG) framework further improved the factual accuracy and relevance of the generated outputs. By incorporating retrieved context into the input prompt, the system minimized generic or hallucinated responses commonly observed in standalone language models.

The LLaMA 3 model, accessed via the Groq API, generated coherent, context-aware, and linguistically appropriate responses across all supported languages. The system demonstrated low latency due to fast inference provided by the Groq API, making it suitable for real-time applications.

Overall, the experimental results indicate that the proposed system effectively handles multilingual queries and generates accurate, contextually relevant responses. The combination of BERT for language detection, Sentence Transformers for embeddings, Qdrant for retrieval, and LLaMA 3 for response generation contributes to the robustness and efficiency of the chatbot.

5.1. Performance Evaluation

To evaluate the performance of the proposed multilingual chatbot system, the accuracy of the BERT-based language detection model was measured across five supported languages.

Table 1 Language Detection Accuracy

Language	Accuracy (%)	Precision	Recall	F1-Score
English	99.20%	0.989	0.99	0.99
Hindi	99.80%	0.99	0.99	0.99
Marathi	96.80%	0.96	0.97	0.96
Spanish	98.50%	0.98	0.98	0.98
French	98.10%	0.97	0.98	0.96
Overall	98.48%	0.97	0.98	0.97

5.2. Results

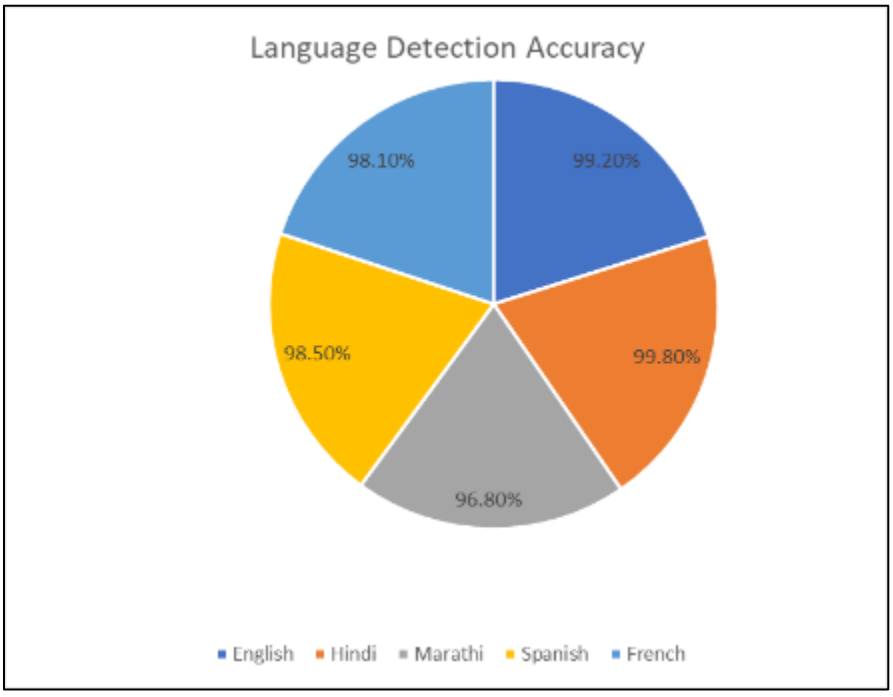

Figure 2 Language Detection Accuracy Graph

5.3. Analysis

The experimental evaluation of the proposed BERT-based multilingual AI chatbot model was conducted across five languages: English, Hindi, Marathi, Spanish, and French. The performance was assessed using standard classification metrics, namely Accuracy, Precision, Recall, and F1-score.

The results indicate that the proposed model achieved high classification performance across all supported languages, with an overall accuracy of 98.48%, demonstrating the effectiveness of the transformer-based architecture for multilingual text understanding. Among all languages, Hindi exhibited the highest performance, achieving 99.80% accuracy, along with 0.99 precision, 0.99 recall, and 0.99 F1-score. This reflects the model's strong capability in identifying and processing Hindi language inputs with minimal misclassification.

Similarly, English also showed excellent results, with 99.20% accuracy and balanced precision, recall, and F1-score values close to 0.99, indicating highly reliable classification performance. For Spanish and French, the model maintained consistently strong performance, with accuracies of 98.50% and 98.10%, respectively. Their precision and recall values demonstrate robust multilingual generalization of the BERT model across non-native and cross-lingual datasets.

The comparatively lower performance was observed for Marathi, with an accuracy of 96.80%, precision of 0.96, recall of 0.97, and F1-score of 0.96. Although slightly lower than the other languages, these values still indicate strong predictive capability. The minor reduction may be attributed to factors such as:

- Limited dataset size,
- Code-mixed or colloquial expressions,
- Spelling variations in regional language text,
- Comparatively fewer pretrained contextual representations.

The overall precision (0.97), recall (0.98), and F1-score (0.97) further confirm that the proposed BERT-based model maintains a strong balance between correct positive predictions and comprehensive class coverage, making it suitable for deployment in multilingual chatbot systems. Overall, the results validate that BERT provides highly effective contextual embeddings for multilingual language detection and intent understanding, significantly improving chatbot response accuracy across multiple languages.

The experimental results demonstrate that the proposed BERT-based multilingual chatbot model achieves robust and consistent performance across five languages, with an overall accuracy of 98.48%, confirming its effectiveness for multilingual conversational AI applications.

6. Conclusion

The experimental evaluation of the proposed BERT-based multilingual AI chatbot model demonstrates its high effectiveness in multilingual language detection and response processing. The model achieved an overall accuracy of 98.48%, along with 0.97 precision, 0.98 recall, and 0.97 F1-score, indicating strong classification performance across all five supported languages. The results confirm that the transformer-based BERT architecture efficiently captures contextual and semantic features of multilingual text inputs. Although a slight performance variation was observed for Marathi, the overall metrics validate the robustness, reliability, and scalability of the proposed system for real-world multilingual conversational AI applications.

References

[1] Vanjani, M., Aiken, M., and Park, M.(2019). Chatbots for Multilingual exchanges. Journal of Management Science and Business Intelligence.

[2] Galadima, K. R., and Lawrence, E.(2024). AMulti- Lingual Conversational AI Chatbot for Effective Educational Consultations A Study of ACE- DS.

[3] Shrivastava, N., Tewari, P., Sujatha, S., Rao, S. B., Varshney, N., and Sharma, V.(2025). Natural Language Processing for Conversational AI Chatbots and Virtual assistants.

[4] Cotfas, L.- A., Sandu, A., Delcea, C., Diaconu, P., Fra˘sineanu, C., and Sta˘nescu, A.(2025). From Mills to ChatGPT An Analysis of Large Language Models Research. IEEE Access.

[5] Gurioli, A., Gabbrielli, M., and Zacchiroli, S.(2025). recognizing AI- written Programs with Multilingual Code Stylometry. IEEE.

[6] Al Bataineh, A., Sickler, R., Kurcz, K., and Pedersen, K.(2025). AI- Generated Versus Human Text Introducing a New Dataset for Benchmarking and Analysis. IEEE Deals on AI.

[7] Pattanayak, S., Mohammed, A., et al.(2025). ChatSense – A Multilingual Chatbot. International Journal of Scientific Research in Science and Technology.

[8] Orosoo, M., Goswami, I., Alphonse, F. R., Fatma, G., Rengarajan, M., and Bala, B. K.(2024). Enhancing NLP in Multilingual Chatbots forCross- Cultural Communication. IEEE ICICV.

[9] Munjal, G., Agarwal, P., Goyal, L., and Samiran, N.(2025). Multilingual Virtual Healthcare Assistant. Health Care Science.

[10] Brown et al.(2020), Language Models are numerous- Shot Learners(GPT- 3).

[11] Ouyang et al.(2022), Training Language Models to Follow Instructions with mortal Feedback(InstructGPT/ RLHF).

[12] LLM Selection and Vector Database Tuning A Methodology for Enhancing RAG Systems, Lukasz Pawlik(2025).