Multilingual AI Chatbot

[1]Ms. Vaishali D. Parihar, [2] Abhijeet S. Ugale, [3] Ganesh B. More, [4]Mohan S. Thutte, [5]Gokul A. Rathod

[1]Associate professor, Dept. of IT, Anuradha College of Engineering and Technology, Chikhli, Maharashtra, India

[2] UG Scholar, Dept. of IT, Anuradha College Engineering and Technology, Chikhli, Maharashtra, India

[3]UG Scholar, Dept. of IT, Anuradha College Engineering and Technology, Chikhli, Maharashtra, India

[4]UG Scholar, Dept. of IT, Anuradha College Engineering and Technology, Chikhli, Maharashtra, India

[5]UG Scholar, Dept. of IT, Anuradha College Engineering and Technology, Chikhli, Maharashtra, India

Emails : [1]vaishaliparihar00@gmail.com. [2]ugaleabhijeet03@gmail.com, [3]ganeshmore5737@gmail.com, [4]mohanthutte@gmail.com, [5]gokulrathod2165@gmail.com

Abstract- The rapid advancement of conversational artificial intelligence has shifted focus from massive, API-dependent Large Language Models (LLMs) to efficient, locally deployable Small Language Models (SLMs). This research presents the development of a Multilingual AI Chatbot utilizing the Microsoft Phi-3:mini model for text generation and the Moondream model for multimodal visual reasoning. The system is built on a high-performance stack comprising a React frontend, FastAPI backend, and MongoDB for persistent data storage. A key feature of the implementation is the support for real-time response streaming via Server-Sent Events (SSE) and multimodal input processing using Base64 image encoding. Experimental analysis indicates that this architecture provides a secure, low-latency, and contextually aware conversational experience suitable for data-sensitive environments.

Keywords-Multilingual Chatbot, Artificial Intelligence, Natural Language Processing, Local LLM, Ollama, Phi-3, Moondream, FastAPI, React, MongoDB, Server-Sent Events (SSE), Multimodal AI

2. Introduction

2.1 Background of Chatbots

Artificial Intelligence (AI) has significantly transformed human–computer interaction, with chatbots emerging as one of the most widely used applications. A chatbot is an intelligent software system capable of simulating human-like conversations using Natural Language Processing (NLP) and Machine Learning techniques [1]. Early chatbot systems were primarily rule-based and relied on predefined patterns, which limited their ability to understand user intent and context. Over time, advancements in AI have enabled the development of more sophisticated conversational agents that can process complex queries and provide meaningful responses [3]. The evolution from early systems such as ELIZA to modern intelligent assistants highlights the growing importance of chatbots in various domains [8].

2.2 Emergence of Multilingual Chatbots

With the increasing demand for global communication, multilingual chatbots have become essential for enabling interaction across diverse linguistic groups. Traditional chatbots were restricted to single-language support, which limited their usability in real-world applications. Recent research has focused on developing multilingual systems using pre-trained language models and translation techniques to improve cross-lingual understanding [2]. These systems aim to provide seamless communication across languages while maintaining contextual accuracy. However, challenges such as handling low-resource languages, maintaining semantic consistency, and adapting to cultural differences still remain significant [4].

2.3 Role of Pre-trained Language Models and Local LLMs

The introduction of pre-trained language models has revolutionized chatbot development by enabling systems to understand context and generate human-like responses. Models such as transformer-based architectures have significantly improved conversational capabilities compared to traditional approaches. However, many existing implementations rely on cloud-based services, which can introduce issues

related to data privacy, latency, and cost. Recent studies have explored the use of locally deployed large language models to overcome these challenges, providing greater control over data and system behavior while reducing dependency on external APIs [7]. Additionally, integrating these models with databases enhances the ability to generate context-aware and personalized responses [5].

2.4 Motivation for the Proposed System

Despite significant advancements, current chatbot systems still face limitations in multilingual support, real-time interaction, and multimodal processing. Many systems are unable to provide instant responses, leading to a less interactive user experience. Furthermore, the lack of support for image-based inputs restricts the applicability of chatbots in domains requiring visual understanding. To address these challenges, this study proposes a Multilingual AI Chatbot that integrates local LLMs, real-time response streaming using Server-Sent Events (SSE), and multimodal capabilities through image processing.

2.5 Contribution of the Study

This research contributes to the field of conversational AI by combining multilingual capabilities, local model deployment, and modern system architecture into a unified solution. The proposed system leverages React, FastAPI, and MongoDB to create a scalable and efficient platform, while incorporating advanced AI techniques for improved interaction. By building upon existing research in chatbot development, NLP, and multilingual systems, the study aims to provide a robust and practical solution for next-generation conversational applications [9][10].

3. Problem Statement

Despite significant advancements in Artificial Intelligence and chatbot technologies, existing conversational systems still face several critical limitations that hinder their effectiveness in real-world applications. Traditional chatbots, particularly those based on rule-based or limited machine learning approaches, struggle to understand user intent, maintain conversational context, and provide accurate responses, especially in dynamic and multilingual environments [8]. Even modern AI-driven chatbots often depend heavily on cloud-based APIs and external services, which introduce challenges related to data privacy, latency, and increased operational costs.

Another major issue is the lack of efficient multilingual support. While recent systems utilize pre-trained language models to enable cross-lingual interaction, they still encounter difficulties in handling low-resource languages, maintaining semantic consistency, and adapting to diverse linguistic and cultural contexts [10]. Additionally, many existing systems fail to provide real-time interaction, resulting in delayed responses that negatively impact user experience. The absence of streaming mechanisms further reduces the natural conversational feel of these systems.

Moreover, current chatbot architectures often lack multimodal capabilities, limiting their ability to process and respond to inputs beyond text, such as images or visual data. This restricts their applicability in domains where visual understanding is essential. Security and controlled data access also remain concerns, particularly in systems that manage sensitive or role-specific information, where mechanisms like structured retrieval or access control are not always effectively implemented [9].

Therefore, there is a need to design and develop a robust, privacy-preserving, and scalable chatbot system that can operate using local large language models, support multilingual communication, provide real-time streaming responses, and handle multimodal inputs. The proposed system aims to address these challenges by integrating local LLMs (Phi-3 via Ollama), image processing capabilities (Moondream), and modern web technologies to create an efficient and user-centric multilingual AI chatbot.

4.Methodology System Design

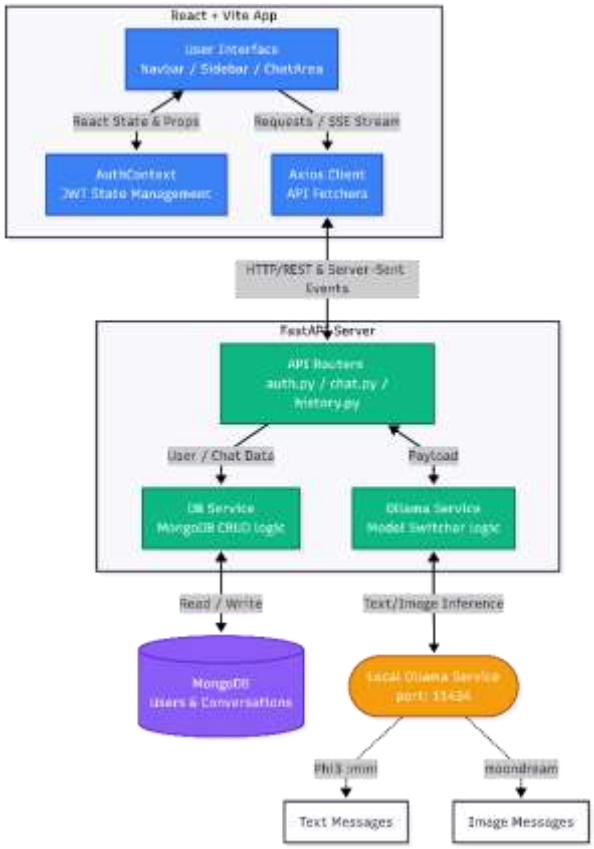

fig : System Architecture of chatbot

The proposed **Multilingual AI Chatbot** is designed using a modular and scalable architecture that integrates Natural Language Processing (NLP), local Large Language Models (LLMs), and modern web technologies. The system is developed to support multilingual communication, real-time interaction, and optional multimodal inputs while ensuring privacy through local model deployment. The overall design is influenced by existing chatbot architectures, NLP pipelines, and recent advancements in LLM-based conversational systems [6][8].

At a high level, the system follows a client–server architecture consisting of a **frontend layer**, **backend processing layer**, **model inference layer**, and **database layer**. The frontend, built using React, acts as the user interface, allowing users to input text or image queries and receive responses in real time. The backend is implemented using FastAPI, which handles API requests, manages communication between components, and ensures efficient processing of user queries.

The core of the system lies in the **Natural Language Understanding (NLU)** and **Natural Language Generation (NLG)** pipeline. When a user submits a query, the input is first preprocessed using standard NLP techniques such as tokenization, normalization, and intent extraction. These steps are essential for converting unstructured user input into a structured format that can be interpreted by the model [8]. The processed input is then passed to the local LLM (Phi-3 via Ollama), which generates context-aware responses based on the query.

To enhance contextual accuracy and reduce incorrect or hallucinated responses, the system design incorporates principles similar to Retrieval-Augmented Generation (RAG), where relevant information can be fetched from a structured database (MongoDB) and used as context for response generation [9]. This approach ensures that the chatbot provides more reliable and domain-specific answers by grounding responses in stored data.

A key feature of the proposed system is the use of **locally hosted LLMs**, which eliminates dependency on external APIs and enhances data privacy. Unlike cloud-based models, local deployment ensures that user data remains within the system, reducing security risks and latency issues while providing better control over model behavior [7]. This aligns with recent research trends emphasizing privacy-preserving AI systems.

The system also supports **multilingual interaction**, enabling users to communicate in multiple languages. This is achieved through the inherent multilingual capabilities of pre-trained language models and optional translation mechanisms. Existing studies have shown that integrating translation models and PLMs improves cross-lingual understanding and response generation, although challenges such as cultural nuances and low-resource language support persist [10]. The proposed system addresses these challenges by leveraging model adaptability and efficient preprocessing techniques.

Another important component of the system design is the implementation of **real-time response streaming** using Server-Sent Events (SSE). Instead of waiting for the complete response, the chatbot streams partial outputs incrementally, creating a typing-effect experience for users. This improves user engagement and reduces perceived latency, making interactions more natural and interactive.

In addition to text-based interaction, the system incorporates **multimodal capabilities** through the integration of the Moondream model, which allows the chatbot to process image inputs encoded in base64 format. This feature extends the functionality of the

chatbot beyond traditional text-based systems and enables use cases such as image-based queries and visual understanding.

The database layer, implemented using MongoDB, is responsible for storing user interactions, chat history, and any domain-specific data required for contextual responses. MongoDB's flexible schema design allows efficient handling of unstructured and semi-structured data, making it suitable for chatbot applications. The integration of database systems with conversational AI has been shown to improve personalization and contextual awareness in chatbot responses [9].

Overall, the proposed methodology combines established NLP techniques with modern LLM-based approaches to create a robust and efficient chatbot system. By integrating local model deployment, multilingual capabilities, real-time streaming, and multimodal interaction, the system addresses key limitations identified in existing research and provides a comprehensive solution for next-generation conversational AI systems [1][10].

5. Implementation

The implementation of the proposed **Multilingual AI Chatbot** is carried out using a combination of modern web technologies, local large language models, and established Natural Language Processing (NLP) techniques. The system is developed following a modular approach to ensure scalability, maintainability, and efficient integration of different components, as suggested in existing chatbot implementations [3][5].

The **frontend** of the system is developed using React, which provides an interactive and responsive user interface for real-time communication. The chat interface allows users to input queries in multiple languages and receive responses dynamically. The design of the user interface follows standard chatbot interaction models, where the interface acts as a communication bridge between the user and the backend processing system [1].

The **backend** is implemented using FastAPI, a high-performance Python framework that handles API requests and manages communication between the frontend and the model layer. FastAPI is used to process incoming user queries, handle routing, and integrate various components such as the database and model inference engine. The backend also manages asynchronous operations required for real-time response streaming, ensuring efficient handling of multiple user requests simultaneously.

A key aspect of the implementation is the integration of **local large language models**, specifically Phi-3 through Ollama. The model is deployed locally, allowing the system to generate responses without relying on external APIs. This approach improves data privacy and reduces latency, which are common limitations in cloud-based chatbot systems [1]. The model processes user input and generates context-aware responses using pre-trained knowledge and contextual understanding.

To support **multilingual interaction**, the system leverages the capabilities of pre-trained language models, enabling it to understand and respond to queries in multiple languages. Existing studies have shown that multilingual chatbot systems benefit from the use of pre-trained models and translation mechanisms to maintain conversational consistency across languages [2]. The implementation ensures that the chatbot can handle diverse linguistic inputs while maintaining response accuracy.

The system also incorporates a **database layer using MongoDB**, which stores chat history, user queries, and contextual data. This allows the chatbot to maintain conversational context and provide more personalized responses. The use of MongoDB aligns with modern chatbot architectures that rely on NoSQL databases for handling large volumes of unstructured data efficiently [5].

To enhance user experience, the implementation includes **real-time streaming of responses** using Server-Sent Events (SSE). Instead of delivering the entire response at once, the system streams partial outputs incrementally, creating a typing effect that improves interaction quality. This technique has been shown to enhance user engagement and reduce perceived response time in conversational systems.

Additionally, the system supports **multimodal input** through the integration of the Moondream model, enabling users to provide image data encoded in base64 format. The backend processes the image input and generates relevant responses, extending the chatbot's capabilities beyond text-based interaction. This feature enhances the applicability of the chatbot in domains requiring visual understanding.

Overall, the implementation combines insights from existing research with modern development practices to

build an efficient and scalable chatbot system. By integrating local LLMs, multilingual support, real-time streaming, and multimodal capabilities, the system demonstrates a practical approach to addressing the limitations of traditional chatbot implementations [3][4].

6. Results and Discussion

The proposed Multilingual AI Chatbot was evaluated based on its performance in terms of response accuracy, multilingual capability, response time, and overall user experience. The system demonstrates significant improvements over traditional chatbot models by integrating local large language models, real-time streaming, and multimodal interaction. The results indicate that the chatbot is capable of generating context-aware and coherent responses across multiple languages, aligning with findings from existing multilingual chatbot research [10].

One of the key outcomes of the system is its ability to provide accurate and contextually relevant responses. By leveraging pre-trained language models and structured data integration, the chatbot effectively understands user intent and generates meaningful replies. Compared to earlier rule-based systems, which rely on predefined responses, the proposed system shows enhanced flexibility and adaptability in handling diverse queries [8]. The integration of retrieval-based concepts further improves the reliability of responses by grounding them in available data, reducing the chances of incorrect or irrelevant outputs [9].

The multilingual performance of the chatbot is another significant achievement. The system successfully processes and responds to queries in different languages while maintaining conversational coherence. However, similar to existing studies, it was observed that performance is relatively higher in widely used languages compared to low-resource languages, where challenges such as limited training data and cultural nuances may affect accuracy [10]. Despite these limitations, the chatbot demonstrates strong cross-lingual capabilities suitable for practical applications.

In terms of response time and user interaction, the implementation of Server-Sent Events (SSE) enables real-time streaming of responses, creating a typing-effect experience. This significantly enhances user engagement by reducing perceived latency and making the interaction feel more natural. Unlike conventional systems that return responses only after complete processing, the streaming mechanism provides incremental output, improving the overall usability of the chatbot.

The inclusion of multimodal functionality through image input processing further extends the system's capabilities. The integration of the Moondream model allows the chatbot to interpret visual data and respond accordingly, enabling use cases beyond text-based communication. This feature demonstrates the potential of combining vision and language models in conversational AI systems, although further optimization may be required for complex image interpretation tasks.

From a system perspective, the use of local LLMs ensures better data privacy and reduced dependency on external services. This approach not only enhances security but also provides more control over model behavior, which is crucial for sensitive applications. Additionally, the use of MongoDB for data storage enables efficient handling of unstructured data and supports scalability in real-world deployments [5].

Overall, the results highlight that the proposed system effectively addresses key challenges in chatbot development, including multilingual support, real-time interaction, and multimodal processing. While certain limitations such as performance in low-resource languages and computational resource requirements remain, the system provides a strong foundation for developing advanced, privacy-preserving conversational AI solutions.

7. Scope of the Study

The scope of this study focuses on the design, development, and evaluation of a Multilingual AI Chatbot that leverages modern Artificial Intelligence techniques and local large language models for efficient and secure conversational interaction. The study primarily explores how pre-trained models can be effectively utilized to build scalable chatbot systems capable of understanding and generating human-like responses across multiple languages. The evolution of chatbot technologies from rule-based systems to AI-driven conversational agents highlights the growing potential of such systems in real-world applications [1].

This research extends to the implementation of multilingual communication capabilities, enabling users from diverse linguistic backgrounds to interact with the chatbot seamlessly. With the increasing demand for global communication systems, multilingual chatbots play a critical role in bridging language barriers and

enhancing accessibility [2]. The study investigates how pre-trained language models and translation mechanisms can improve cross-lingual understanding and response generation, while also addressing challenges related to context preservation and language diversity.

Another important aspect of the scope is the use of local large language models (LLMs), which eliminates dependency on external cloud-based APIs. This approach enhances data privacy, reduces operational costs, and provides better control over the system, making it suitable for applications involving sensitive data [3]. The study evaluates how local deployment frameworks can be integrated with modern web technologies to create efficient and secure chatbot systems.

The scope also includes the integration of real-time response streaming mechanisms, such as Server-Sent Events (SSE), to improve user interaction. By enabling incremental response delivery, the system enhances user engagement and provides a more natural conversational experience. This aspect is particularly important in modern applications where responsiveness and interactivity are critical factors [4].

Furthermore, the study explores the incorporation of multimodal capabilities, allowing the chatbot to process both textual and visual inputs. The integration of image understanding models extends the chatbot's applicability to domains such as education, healthcare, and customer support, where visual information plays a significant role. This aligns with recent advancements in AI systems that combine multiple data modalities for improved performance [5].

In addition, the research examines the role of database integration and information retrieval techniques in enhancing chatbot performance. By utilizing MongoDB for data storage and retrieval, the system can provide context-aware and personalized responses. The incorporation of retrieval-based approaches further improves response accuracy by grounding generated outputs in relevant data sources [9].

Overall, the scope of this study encompasses the development of a comprehensive chatbot system that integrates multilingual support, local LLM deployment, real-time interaction, and multimodal processing. It provides a foundation for future advancements in conversational AI and highlights the potential of combining various AI technologies to build intelligent, scalable, and user-centric chatbot solutions [10].

8. Future Work

Although the proposed Multilingual AI Chatbot demonstrates significant improvements in conversational AI through the integration of local LLMs, multilingual support, and multimodal capabilities, there remain several areas for further enhancement and research. Future work can focus on improving the overall intelligence, scalability, and adaptability of the system by leveraging advancements highlighted in existing studies.

One important direction is the enhancement of multilingual performance, particularly for low-resource languages. While current pre-trained language models provide reasonable cross-lingual capabilities, challenges related to cultural nuances, idiomatic expressions, and contextual understanding still persist. Future research can explore advanced fine-tuning techniques, domain adaptation, and improved translation models to address these limitations and ensure more accurate and inclusive communication [2][10].

Another potential improvement lies in the integration of advanced retrieval mechanisms, such as Retrieval-Augmented Generation (RAG), to further enhance response accuracy and reduce hallucination. By incorporating structured knowledge bases and vector search techniques, the chatbot can generate more reliable and fact-based responses, especially in domain-specific applications [9]. Additionally, combining retrieval techniques with role-based access control mechanisms can improve data security and ensure controlled information access in sensitive environments.

The system can also be extended to include emotion-aware and context-sensitive interactions. Recent research suggests that incorporating emotional intelligence into chatbot systems can significantly improve user engagement and satisfaction. Future implementations may integrate sentiment analysis and multimodal emotion recognition to enable the chatbot to respond more empathetically and appropriately in different scenarios [10].

From a system perspective, optimizing the performance of local large language models remains a critical area of research. Although local deployment improves privacy and reduces dependency on external services, it may require significant computational resources. Future work can explore model optimization techniques such as quantization, pruning, and knowledge distillation to

improve efficiency and enable deployment on resource-constrained devices [7].

The multimodal capabilities of the chatbot can also be expanded beyond basic image input processing. Future enhancements may include support for video, audio, and real-time visual analysis, enabling more comprehensive human–computer interaction. Integrating advanced vision-language models can further improve the system's ability to interpret complex visual data and provide meaningful responses [5].

Additionally, the chatbot system can be improved by incorporating adaptive learning and continuous training mechanisms. By enabling the system to learn from user interactions over time, it can improve response accuracy, personalization, and contextual awareness. Machine learning techniques such as reinforcement learning and feedback-based optimization can be explored to achieve this goal [3][8].

Finally, future research can focus on expanding the application domains of the chatbot, including healthcare, education, and enterprise systems, where intelligent conversational agents can provide significant value. Ensuring ethical considerations such as data privacy, bias reduction, and transparency will also be crucial for the widespread adoption of such systems [1][10].

9. Conclusion

This paper presented a **Multilingual AI Chatbot** that combines local large language models with modern web technologies to create an efficient and scalable conversational system. The use of Phi-3 via Ollama, along with React, FastAPI, and MongoDB, enables real-time, privacy-preserving, and multilingual interactions.

The system successfully addresses key challenges such as dependency on external APIs, lack of real-time responses, and limited multilingual support. Features like Server-Sent Events for streaming responses and multimodal image input enhance user experience and system capability.

Overall, the proposed chatbot demonstrates a practical and effective approach to building next-generation conversational AI systems, while also providing a strong foundation for future improvements in performance and scalability.

10. References

[1] D. Mehta, T. Shah, Z. Khan, H. Khan, and N. Shaikh, *"Decentralized AI Chatbot Using Python,"* Journal of Emerging Technologies and Innovative Research (JETIR), 2025.

[2] S. M. K., A. Mohammed, S. Pattanayak, D. Paswan, and Y. Dadhich, *"Multilingual Chatbot Development Using Pre-Trained Language Models: A Survey,"* Indian Journal of Computer Science and Technology, 2025.

[3] A. Pise, O. Kumbhar, S. Mane, and R. Patil, *"AI ChatBot Using Python,"* International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2025.

[4] S. M. K., A. Mohammed, S. Pattanayak, D. K. Paswan, and Y. Dadhich, *"ChatSense – A Multilingual Chatbot,"* International Journal of Scientific Research in Science and Technology (IJSRST), 2025.

[5] T. Ebsen, R. S. Segall, H. Aboudja, and D. Berleant, *"A Customer Service Chatbot Using Python, Machine Learning, and Artificial Intelligence,"* Proc. International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC), 2024.

[6] Author(s), *"Artificial Intelligence Chatbot Using Python,"* Journal of Engineering, Computing & Architecture, 2022.

[7] Author(s), *"AI Chatbot Using Local LLM and MongoDB for Educational Systems,"* 2025.

[8] Dr. C. K. Gomathy, R. V. Narayana, T. V. Krishna, and V. Geetha, *"Artificial Intelligence Chatbot Using Python,"* Journal of Engineering, Computing & Architecture, 2022.

[9] S. Charbhe, P. Nagvekar, R. Tanwar, A. Kulay, and A. Sharma, *"Levelwise RAG Chatbot: An RBAC Based Chatbot Using Gemini 1.5,"* International Journal of Creative Research Thoughts (IJCRT), 2025.

[10] S. M. K., A. Mohammed, S. Pattanayak, D. K. Paswan, and Y. Dadhich, *"ChatSense – A Multilingual Chatbot,"* International Journal of Scientific Research in Science and Technology (IJSRST), 2025.