International Journal On Engineering Technology and Sciences – IJETS
ISSN (P):2349-3968, ISSN (O): 2349-3976 Volume XIII - Issue IV, April - 2026

AI CHATBOT USING LLaMA 3:8B, TinyLLaMA, RAG AND VECTOR DATABASES

Afran Syed P
B.E Student (Third Year)
Department of Computer Science and Engineering
Francis Xavier Engineering College
Tirunelveli, Tamil Nadu, India
afran.ug.23.cs@francisxavier.ac.in

Anwar S
B.E Student (Third Year)
Department of Computer Science and Engineering
Francis Xavier Engineering College
Tirunelveli, Tamil Nadu, India
anwar.ug.23.cs@francisxavier.ac.in

John Amics A
B.E Student (Third Year)
Department of Computer Science and Engineering
Francis Xavier Engineering College
Tirunelveli, Tamil Nadu,
Indiajohn.ug.23.cs@francisxavier.ac.in

Siva Kumar K
Assistant Professor
Department of Computer Science and Engineering
Francis Xavier Engineering College
Tirunelveli, Tamil Nadu,
Indiasivakumar@francisxavier.ac.in

Abstract

This paper proposes an AI intelligent chatbot system using LLaMA 3:8B, TinyLLaMA, RAG, Vector Databases to create a context-aware, human-like and up to date conversation experience. The system utilizes LoRA and QLoRA for model optimization in an effort to boost model efficiency and limit computational costs/memory usage. FastAPI and REST APIs are implemented for a scalable back end and Ngrokallows for safe deployment and testing from afar. It can be used in education, career counselingand customer service due to its efficiency and context awareness.

Keywords: Artificial Intelligence, AI Chatbot, LLaMA 3:8B, TinyLLaMA, Retrieval-Augmented Generation (RAG), Vector Database, LoRA, QLoRA, FastAPI, REST API, Natural Language Processing (NLP), Conversational AI

Introduction

AI and NLP techniques has provided advancements in building intelligent chatbots that can understand human languages and generate human-like response. In this paper, an AI chatbot is designed with the integration of LLaMA 3:8B, TinyLLaMA, Retrieval Augmented Generation (RAG), and Vector Databases to offer intelligent, responsive, real-time and context aware conversational support.LLaMA 3:8B is utilized for advanced reasoning and responses, TinyLLaMA for lightweight and low latency responses. The system also utilizes LoRA and QLoRAmethods to increase efficiency of models and decrease resource usage. For robust backend communications and distributed systems, the FastAPI and REST APIs are employed along with Ngrok for the system to be testable in real remote environment. The intelligent and efficient chatbot offers suitable solution for education, career counselingand customer service basedsystems. Combining RAG with Vector databases ensures efficient semantic search and improve the quality of the

responses using context retrieval before generating the final responses. It focuses on reducing latency, enhance scalability, and achieve cost effective deployment for advanced conversational systems.

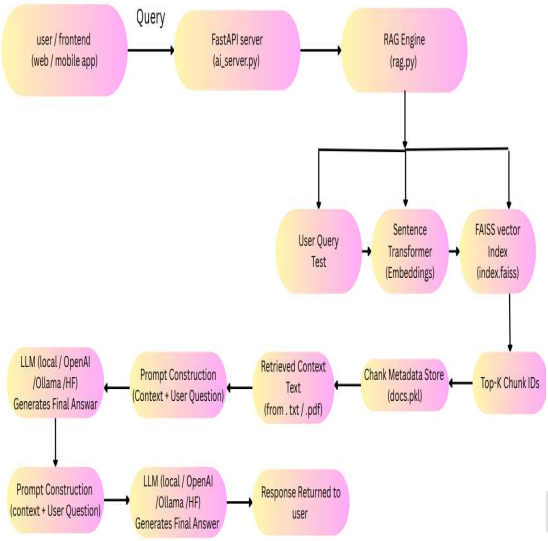

Diagram of the proposed AI Chatbot

Existing System

Currently, most existing AI chatbot systems adopt a rule-based, keyword-based and pattern-based matching mechanism in response to the users. In spite of modern chatbot systems incorporating NLP techniques or LLM for conversing with users, they still suffer from the problem of shallow context-understanding ability, incorrect answers are returned and high computational cost is generated, besides they mostly rely on static knowledge which means they cannot grasp the latest trends. Mostly all the existing chatbot systems have failed to take advantage of RAG and Vector Databases to query the latest contextual information and have led to the problem of returning not related and hallucination answers. Besides, most of them fail to incorporate light-weighted optimization technologies such as LoRA and QLoRA which therefore incurs high cost and requires high performance when deploying the system.

Apart from the limitations in the context information query, many traditional chatbot systems cannot scale on the backend, they fail to perform the real time communication as they do not have effective application with FastAPI, REST APIs, and remote deploying tools like Ngrok. In light of the challenges discussed above, there is an increasing demand for a more efficient and scalable AI chatbot system with better context-understanding ability using LLaMA 3:8B, TinyLLaMA, RAG, Vector Databases etc.

Proposed System

The system, an advanced conversational AI chatbot developed using LLaMA 3:8B, TinyLLaMA, Retrieval-Augmented Generation (RAG), Vector Databases, LoRA, QLoRA, FastAPI, REST APIs, and Ngrok to deliver intelligent and context-aware conversational assistance. The system utilizes LLaMA 3:8B for robust reasoning capabilities and accurate response generation, TinyLLaMA for lightweight and low-latency conversations. RAG coupled with vector databases are used to ensure that the system retrieves relevant contextual data prior to response generation, increasing semantic accuracy while mitigating hallucination issues.

LoRA and QLoRA techniques are used to further fine-tune and reduce the computational cost for greater efficiency. FastAPI, and REST APIs were used to build a fast, asynchronous backend that facilitates communication between frontend, the models, and vector storage. The chatbot can be deployed remotely using Ngrok. Overall the proposed system is capable of delivering highly efficient, scalable, and real-time conversational assistance for educational, career guidance, customer support, or business purposes.

IJETS

Methodology

The proposed methodology follows this process:

1. User Query/Input Collection
2. Text Pre-processing and Embedding Generation
3. Vector Database Retrieval
4. Retrieval-Augmented Generation (RAG) Processing
5. Response Generation using LLaMA 3:8B and TinyLLaMA
6. Response Delivery and Evaluation

A flow has been illustrated above which is required for the generation of an intelligent and contextually correct responses by the proposed AI Chatbot. The full process begins with the collection of user queries and goes through text pre-processing, embeddings generation, contextual information retrieve from Vector Database and the query is being processed through RAG pipeline. The information is retrieved from the vector database, which is then passed to LLaMA 3:8B and TinyLLaMA for the accurate responses generation. The responses are delivered to the user via FastAPI and REST APIs and subsequently performance assessment of the system and analysis of accuracy are done.

User Query/Input Collection

In the initial phase, users provide the AI chatbot with their queries via a web-based chatbot interface. Users are allowed to input questions/requests in a natural language format and interact with the chatbot in real-time. After receiving the user's input query, it will be forwarded to the backend server through the usage of REST APIs (built using FastAPI).

Text Pre-processing and Embedding Generation

In this stage the user query will be clean and processed by NLP (tokenization, normalization etc.) and will be transformed into a vector embedding that represents its meaning, which can then be use to perform a similarity search within the Vector Database.

Vector Database Retrieval

During this stage, the newly created embeddings are compared against the existing embeddings that have been saved to the Vector Database like FAISS or ChromaDB.

The similarity search takes place and the most relevant context information which is related to the query made by the user is retrieved. This helps to improve the accuracy of the generated response by passing the most relevant information to the AI.

Retrieval-Augmented Generation (RAG) Processing

During this phase, the context fetched from Vector Database and the user query is passed to the RAG pipeline. The RAG pipeline augments the context of the chatbot by providing relevant external information before generating response and help improve the contextuality, decrease hallucinated replies and generate more coherent and credible replies.

Response Generation using LLaMA 3:8B and TinyLLaMA

During this phase, the preprocessedquery and fetched context are fed into LLaMA 3:8B and TinyLLaMA models to generate response. LLaMA 3:8B produces verbose and context-aware response.TinyLLaMA produces quick low-latency response to maintain real-time communication with the user. Finallythe response is made ready for being passed to the user through the chatbot.

IJETS

Response Delivery and Evaluation

Finally, we perform performance evaluation of the proposed AI chatbot system to evaluate accuracy and efficiency of conversational responses generation and analysis how effectively the system retrieves the relevant contextual information and generate relevant responses through LLaMA 3:8B, TinyLLaMA and RAG.

1. Response accuracy- the measure of how relevant and contextually correct is the generated response to user prompt.

2. Retrieval accuracy- measure how well is the Vector database used to retrieve relevant contextual information.

3. Response time- measure how fast does the response generated and deliver to the user.

4. User satisfaction- quality, clarity and helpfulness of chatbot response.

Implementation

The above described AI chatbot system can be implemented with LLaMA 3:8B, TinyLLaMA, RAG, Vector Databases, LoRA, QLoRA, FastAPI, REST APIs and Ngrok. User queries are collected through frontend, fed into NLP processing, preprocessed and transformed to embeddings, compared with database embeddings in the Vector Database such as FAISS or ChromaDB, then retrieve related contexts and passed along with user query to RAG. LLMs including LLaMA 3:8B and TinyLLaMA produce responses according to RAG results.

REST APIs and FastAPI is used to handle communication among frontend, backend and vector storage; Ngrok could be employed to facilitate remote deployment and testing of prototype in early phase. The presented system can therefore support responsive, scalable and real time conversational service while achieve more accurate answer.

Result and Discussion

The AI chatbot system we propose is able to provide highly efficient and contextual answers due to the use of LLaMA 3:8B, TinyLLaMA, RAG, Vector Database, LoRA, QLoRA, FastAPI, REST APIs and Ngrok.

The system was able to provide contextual and human-like responses for a variety of questions asked by the user with low latency and good semantic ability. RAG with the Vector Database provides higher accuracy as the retrieved context is provided as input prior to response generation, therefore limiting irrelevant and hallucinated answers.

LLaMA 3:8B provided a thorough and knowledgeable answer for complex answers whilst the use of TinyLLaMA is good for lightweight answers as it requires a faster processing time and more computational power to answer. The addition of LoRA and QLoRA reduce the computational power required and memory use therefore providing a more cost-effective and scalable system.

Conclusion:

The LLaMA 3:8B, TinyLLaMA, RAG, Vector Databases, LoRA, QLoRA, FastAPI, REST APIs, and Ngrok-based AI chatbot system proposed was able to produce contextually relevant and smart responses for its users. Response accuracy was improved and the system was quick, interactive and could handle the conversations in real time. The described chatbot is a reliable, inexpensive and scalable solution for career, education, and customer support purposes.

IJETS

Future work:

To further develop the system of the suggested AI chatbot in the future, we could try to implement voice interaction feature, multilingual function and more personalized communication strategy to improve it. By adopting more advanced language models, with the support of cloud computing technologies and deployment, we could achieve higher accuracy of response, greater scalability and more efficient real-time performance for big data application.

References:

[1] D. Priyadharshini and R. Ravi, "Deep learning: a survey and techniques for language processing, image, speech and text",Francis Xavier Journal of Science Engineering and Management, vol.1, no.1, pp.11-14, 2020.

[2] A. Jenefa, R. Ravi, and H. Manimala, "A machine doctor that diagnosing ophthalmology problems using Neural Networks",International Journal of Advanced Research in Computer Engineering & Technology, vol.3, no.2, pp.528-533, 2014.

[3] F. Ajesh and R. Ravi, "Hybrid features and optimization-driven recurrent neural network for glaucoma detection",International Journal of Imaging Systems and Technology, vol.30, no.4, pp.1143-1161, 2020.

[4] S. Edwin Raja and Dr. R. Ravi, "An Efficient Detection and Isolation of Phishing Attacks using Customized Hidden Markov Model based False Prediction",
Caribbean Journal of Science, vol.53, no.2, pp.2218-2225, 2019.

[5]D.Priyadharshini,R.Malliga@pandeeswari, S. Shargunam, and R. Ravi, "Data science: a comprehensive surveyand perspective on recent works", Francis Xavier Journal of Science Engineering and Management, vol.1, no.1, pp.7-10, 2020.

[6] J. Johnson, M. Douze, & H. Jgou (2019). Billion-Scale Similarity Search with FAISS. IEEE Transactions on Big Data, 7(3), 535–547.

[7] T. Wolf et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 38–45.

[8] S. Ramrez& A. Gupta (2024). FastAPI for High-Performance AI Applications. Journal of Web Engineering, 23(2), 120–132.

[9] LangChain Documentation (2025). Building RAG Applications with Vector Databases.

[10] Hugging Face (2025). TinyLLaMA: Lightweight Open-Source Language Models for Efficient AI Systems.

[11] OpenAI (2024). Advancements in Conversational AI and Large Language Models. Artificial Intelligence Review, 58(4), 1–15.

[12] S. Zhang & Y. Liu (2023). Natural Language Processing Techniques for Intelligent Chatbot Systems. International Journal of AI Research, 12(1), 45–58.

[13] M. Chen et al. (2023). Semantic Search Using Vector Databases in AI Applications. IEEE Access, 11, 78521–78535.

[14] A. Kumar & R. Singh (2025). AI Chatbot Systems Using Retrieval-Augmented Generation and Large Language Models. Proceedings of the International Conference on Artificial Intelligence and Data Science, 1–6.

IJETS