



LLAMA-3 OPTIMIZED RETRIEVAL-AUGMENTED CHATBOT FOR DYNAMIC QUERY HANDLING AND ENTERPRISE INTEGRATION

Juthuka Aruna Devi¹, T. Anusha², Ramesh Makala³, Pala Pooja Ratnam⁴, Krishna Rupendra Singh⁵,
 Rajendra Kumar Ganiya⁶ and Marada Srinivasa Rao*⁷

¹Department of Computer Science and Engineering (DS), Anil Neerukonda Institute of Technology and Sciences (A), Sangivalasa, Visakhapatnam, Andhra Pradesh, India.

²Department of Computer Science and Engineering, GMR Institute of Technology, Andhra Pradesh, India.

³Department of Computer Science and Engineering, Mittapalli College of Engineering, Vizianagaram, Andhra Pradesh, India.

⁴Department of Computer Science and Engineering, Gayatri Vidya Parishad college of engineering (A), Visakhapatnam, Andhra Pradesh, India.

⁵Department of Computer Science and Engineering, Vignans Institute of Engineering for Women (A), Visakhapatnam, Andhra Pradesh, India.

⁶Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

⁷Department of Information Technology, MVGR College of Engineering, Vizianagaram, Andhra Pradesh, India.

¹<https://orcid.org/0009-0004-0003-7257>, ²<https://orcid.org/0009-0001-0534-7690>, ³<https://orcid.org/0000-0002-3789-631X>,

⁴<https://orcid.org/0009-0005-4451-178X>, ⁵<https://orcid.org/0009-0007-6402-9194>, ⁶<https://orcid.org/0000-0002-9959-5985>

⁷<https://orcid.org/0000-0001-6849-0286>

E-mail: * srinivas.marada22@gmail.com

ARTICLE INFO

Article History

Received: February 04, 2026

Reviewed: March 05, 2026

Accepted: April 01, 2026

Published: May 29, 2026

Keywords:

RAG,
 LLaMA-3,
 Chatbots
 Natural Language
 Domain Specific.

ABSTRACT

The research work introduces the development of a Retrieval-Augmented Generation (RAG)-based chatbot using the LLaMA-3 model. Unlike traditional chatbots, which are limited by predefined responses, the RAG-based framework improves the chatbot's ability to retrieve and integrate information from user-provided documents and from a knowledge base, ensuring that responses are accurate, contextually appropriate, and informative. The chatbot has the ability to understand the intricate queries and produce the response after 4 seconds depending on the prompt by utilizing the advanced natural language capabilities of the LLaMA-3. Privacy is achieved by ensuring that all data processing is done in-house so that the data of the user does not leave the system and does not need internet access. This does not only increase security but it also enables those organizations that have stringent data compliance requirements to utilize the system without concerns. Besides strong response generation, chatbot could be specialized in specific domain needs like providing legal advice, academic research, healthcare consultations and customer service automation. The system is multi-turn, contextual and provides module components that can be easily integrated with other existing enterprise systems. The proposed work indicates how the combination of advanced language models and effective document search tools could lead to the development of very productive, smart, and personal chatbots that can be successfully implemented in various real life situations.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

A chatbot can be described as a computer program that can be used to imitate conversation with a human user, particularly in the Internet [1]. The given work outlines the creation of an augmented generation (RA)-based chatbot based on the LLaMA-3 model. In comparison, unlike traditional chatbots where the responses are constrained by the predefined answers, the RAG-based architecture has the benefit of enabling the chatbot to access and merge information provided by the user through documents, as well as through knowledge base to make sure the responses are accurate, contextually relevant, and informative.

This study demonstrates that the development of advanced language models and retrieval techniques can make successful chatbots that are effective in any field [2],[3]. They are utilized to communicate with people, as well as in other spheres such as education, business, e-commerce. Retrieval augmented generation (RAG) is a natural language processing (NLP) algorithm that integrates the two benefits of retrieval and generative-based artificial intelligence models. With the existing capabilities of LLaMA-3 (a large language model (LLM) created by Meta AI) and the RAG framework, the chat robot does not only comprehend complex queries but also provides accurate and relevant responses. As the research demonstrates, the modern AI can be used to improve the interaction process by offering smarter and more responsive conversational agents. Nowadays, most of the online Ai chatbots like chat-GPT, Gemini and others use their user data to train their models which questions the privacy for the users. Current chatbots often suffer from inaccuracies and limitations in their responses, making it difficult in their ability to provide reliable and informative information. This work aims to develop a more effective chatbot by integrating a Retrieval-Augmented Generation (RAG) architecture with the powerful Llama-3 language model.

Chatbots nowadays hallucinate a lot due to their inefficiency in the information storage of their previous knowledge, this leads to the dissatisfaction of the user. Most chatbots come with their predefined responses rather than answering by understanding the user query, so this research helps in solving these issues and improve the chatbot performance. Retrieval-Augmented Generation (RAG)-based chatbots nowadays use vast retrieval techniques to extract information from the user provided documents and then they also retrieve data from the knowledge base that they have used as these methods show promising responses but there are few drawbacks like they response retrieval time and accuracy of the responses. This kind of methodology across other domains is limited by the fact that they are frequently domain-specific. They also doesn't come with a specific interface with a neat design [4]. The task of developing an effective Retrieval-Augmented Generation (RAG)-based chatbot involves overcoming several challenges related to information retrieval, response generation. Before diving into how RAG with LLaMA-3 improves chatbot performance, it's important to understand the conventional methods used in chatbot development and their inherent limitations [5-7]. Most of the chatbots use simple pattern matching mechanisms. These chatbots can be used in the domains like education. Most of the chatbots are also used during the covid-19 for enquiring about the disease well.

A. Rule – Based Chatbots

Overview: Rule-based chatbots operate on predefined sets of rules and regulations. These systems work by matching user inputs with pre-programmed responses. They are widely used in customer support and FAQ-based systems.

Limitations:

Lack of Flexibility: These bots struggle to handle queries that fall outside their programmed knowledge base.

Scalability: Improving and expanding the rule based chatbots require a significant number of resources and also it is very much difficult in upgrading them and also require a lot of manual effort.

Learning capability is zero: In rule based chatbots, the chatbots do not learn from the user-provided data and often depend upon the predefined data that they have been fed into.

B. Traditional Machine Learning-Based Chatbots

Overview: These chatbots use supervised techniques to generate responses and learn about the use case that they have been trained on. They often train on large sets of data to know their responses and to improve the understanding levels of the user. Machine learning chatbots are better than the traditional rule based chatbots as they understand upto a certain level and provide better responses than the rule based predefined responses. Their responses are not limited unlike the rule based chatbot but they will be learning on how to respond.

Limitations:

Data dependency: These machines learning based chatbots require large amounts of datasets which are often difficult to retrieve and also to maintain, if any new information is needed to be added it is much difficult to train them to respond to their new information.

Less contextual understanding: Even if they are trained on large amounts of data, they do not fully understand the context that the user is talking about and hallucinate a lot.

No dynamic responses: These kinds of chatbots cannot generate responses beyond their trained data and often fail to answer the questions beyond their trained domain. Which makes the chatbots less interactive to the user.

C. Information Retrieval-Based Chatbots

Overview: These chatbots fetch answers from a knowledge base or database using search techniques. They rely on methods such as keyword matching, TF-IDF to retrieve relevant information from documents.

Limitations:

Low understanding: These chatbots often match by keyword matching rather than the full semantic understanding

Poor generating: If a user gives a query rather than the expected in which the system is trained on, these chatbots fail to generate accurate responses.

Non-generative: These chatbots cannot create a new response based on the user query and can often generate responses with the knowledge available in their databases.

D. Generative AI-Based Chatbots (Without RAG)

Overview: Gen Ai chatbots like GPT-3, GPT-4 and other models use pattern-based understanding and can also generate responses like humans making the conversation with them more interactive and comfortable.

Limitations:

Hallucination: These complete generative Ai models sometimes generate incorrect or misleading information due to their limitations in the training data and not identifying the difference between the factual and contextual data.

Do not follow up-to-date knowledge: These Gen Ai models cannot answer about the latest information and cannot actually be up to date.

I.1. HOW RAG ADDRESSES THESE ISSUES

Retrieval-Augmented Generation (RAG) uses the best of both the retrieval mechanisms and also the response generation capability of the LLMs [8]. By using the Retrieval techniques and feeding them to the LLMs will significantly improve the chatbot performance. By integrating RAG with LLaMA-3, the chatbot gains context-aware, factual, and dynamic response capabilities—effectively overcoming the shortcomings of previous methods. These kinds of chatbots can effectively improve customer service in the business sector. These kinds of mechanisms make humans more interactive to chatbots and can make them feel like friendly companions [9][10]. When there is no history or memory updated in the chatbot, it can lead to the looping conversations or make them hallucinate.

A. Objective

The objective of this research is to develop a chatbot which is context aware and provides accurate responses to the user-based queries by combining the power of both the LLM (Large language models) and RAG [11]. The work mainly focuses on

- Improve response accuracy
- Integration of knowledge retrieval
- Develop a user-friendly interface
- To ensure that the system can be adapted to different domains
- Optimize retrieval efficiency

The goal is to make the chatbot useful to all kinds of domains where most chatbots fail to respond to a domain on which there is no knowledge on. This research aims to showcase how AI can be used in future intelligent systems [12].

B. Existing System

The current system of Retrieval-Augmented Generation (RAG) based chatbots combines document retrieval and language generation to provide more relevant and accurate responses to user queries. In these systems, relevant information is first retrieved from large datasets, such as PDFs or knowledge bases, using retrieval techniques. The retrieved data is then passed to a generative model, like LLaMA-2, to produce conversational responses. While effective, the current systems face challenges such as retrieval accuracy, response latency. These systems are often domain-specific, limiting their flexibility across different fields.

C. Proposed System

The proposed system seeks to overcome these limitations by developing a more advanced RAG-based chatbot using the LLaMA-3 model, which can dynamically process user-uploaded documents from any domain and deliver highly relevant, context-aware responses in real-time. By optimizing the retrieval process through fine-tuning the LLaMA-3 model, this system will minimize latency and improve the accuracy of responses. Additionally, the system is designed to be domain-specific and robust solutions that can be applied across various fields, making it adaptable for a wide range of industries and domains.

II. LITERATURE SURVEY

According to [13] has developed a medical chatbot using Llama-2 which uses the Llama-2 model to generate the responses and uses vectorDB as their database. In this study langchain loaders and splitter were used and from this study we have understood that the quantized models can also perform well under the right resources. Although, this study has the drawback of not using the latest models and also there are better embedding models which outperforms the models used in this research. In [14] has developed a RAG based LLM chatbot using Llama-2, this study focuses on using the RAG capabilities to generate a emotional and supportive responses to the victims of the sexual harassment. This study used the langchain text splitters for chunking and used the chromaDB as their knowledge base. It is specifically designed for the survivors of the sexual harassment and cannot go beyond the domain of this topic.

This work shows that the LLMs can generate responses which are emotionally understandable and supportive to the humans, which makes the user more interactive to the chatbot and uses them more. This study used the Llama-2 7b model which is the drawback of this study where Llama-3 outperforms Llama-2 in every aspect of comparison. The responses are also limited to that specific domain and cannot be used to talk to the chatbot about other topics which makes this chatbot less usable to most of the user as most of the people do not go under the sexual harassment. According to [15] proposed "Mr. Dr. Health-Assistant Chatbot," an AI-driven virtual assistant designed to provide preliminary medical guidance to users.

The chatbot is trained on various medical datasets to offer health-related advice, symptom analysis, and disease-related information. The study has indicated that AI is a critical aspect of healthcare, particularly in terms of eliminating the reliance of human medical practitioners in simple medical inquiries. Nevertheless, one of the issues that can be seen in the research is the medical accuracy of the chatbot since some wrong or old information may confuse the users. Further, the use of the chatbot is limited due to the use of fixed databases that are not capable of offering real-time information on changes in medical knowledge. According to [16] investigated the use of chatbots in the context of communicating about the COVID-19 pandemic in the field of public health. Their article describes the successful implementation of chatbots to help with symptom screening, common questions, and ease the burden on healthcare workers since they can automatically respond to the general queries. The study highlights the promise of AI-based tools in the context of handling the health crisis at scale by making sure that the correct information is disseminated in a timely manner. The problems with the study also include the possibility of misinformation because of the outdated knowledge base and limited conversational depth of chatbots that can potentially be poor in solving complicated medical issues, however. According to [17] explored the interaction between users and chatbots and voice assistants and examined the motivations and the perception of AI-based conversational systems. The analysis reveals that although chatbots enhance the levels of customer service accessibility and efficiency, customer engagement mainly hinges on the level of accuracy of responses, personalisation and fluency of conversation. Chatbots have more chances to make the user believe and remain using them when the replies are natural and contextually appropriate. Although these are the advantages, one of the major limitations is the absence of emotional intelligence in chatbots and therefore, the inability to effectively communicate sensitive discussions.

Furthermore, the report indicates that users would rather have human interaction when their questions are more complex meaning that AI-based assistants are most effective in solving more formalized and routine problems than a more subtle dialogue. By [18] discusses the advantages and disadvantages of artificial intelligence chatbots to companies. Although chatbots present the advantages of time-saving and customer service enrichment, they are associated with the emergence of new threats such as prejudice, non-interference with other systems, and data security threats. To help managers to understand whether their organization is ready to use the chatbots created by AI, the research offers a new framework based on the Technology Acceptance Model (TAM). Four very important areas that it contributes towards to give a complete picture of preparedness are social influence (subjective norms), compatibility, available assistance (facilitating conditions), and trust. The framework and evaluation tool assist the managers to be more judgmental when it comes to deciding on whether to implement chatbots in marketing, sales and in customer service. For of author [19] explores the role of chatbots in online services and commerce with particular emphasis on the way they could enhance the quality of customer care. It sets out two main objectives namely, better service performance as well as meeting the expectations of the customers using such features like engagement, entertainment, problem-solving, customisation, and trendiness. It is quite evident that chatbots significantly improve the experience of clients and the outcome of the service.

According to [20] examines the ethical issues associated with the use of chatbots by nephrologists as it brings attention to the fact that special laws should be enforced to protect personal information, integrity of stored information, and the possibility of receiving an informed consent of patients prior to using these tools. Patient information is focused on in the form of secure techniques, data visibility, anonymization and encryption, among others. Regular evaluations of algorithms and diversity programs are proposed as the way to reduce prejudice. The analysis continues to state that maintaining a doctor-patient relationship has to be trustworthy by being open, designed intuitively, culturally sensitive, and finding a happy medium between the human and automated interaction. To enhance patient care by utilizing chatbots in nephrology that was properly and safely used, vast ethical guidelines and continual studies are required. Smart college chatbot by [21] is an academic and administrative chatbot built on machine learning and Python to provide support to students with their academic and administrative questions. The chatbot will be configured to address frequent student issues like enrolling in courses, exam times, and payments, thus it will decrease the workload of the administration. The research shows that chatbots are capable of increasing the engagement of students through the need to provide immediate feedback and increase the availability of institutional information. One such weakness, however, is that the chatbot is less flexible to dynamic or unstructured questions because it relies on pre-created datasets. Besides, the quality of the training data also affects the performance of the chatbot, and it needs periodic updates to be accurate and relevant.

III. METHODOLOGY

The RAG-based chatbot development methodology using LLaMA-3 takes a systematic procedure that is effective to retrieve documents and process them effectively to generate responses. The system will accept user-supplied PDFs, extract the information with the help of embeddings, retrieve the most relevant information with LLaMA-3 and produce the context-based responses. The following is a breakdown of the methodology that will be used in this research work.

A. Preprocessing the Data

a. Data Ingestion and Preprocessing:

Since the chatbot needs to process user-uploaded PDFs, an efficient pipeline is required to handle various document formats, extract textual content, and preprocess the data for retrieval.

- Text Extraction:
 - PDFs are parsed using PyPDF to extract structured and unstructured text.
- Cleaning and Normalization:
 - Unnecessary white spaces, special characters, and formatting issues are removed.
 - The text is split into meaningful chunks to improve retrieval performance.

- **Chunking Strategy:**
 - Fixed-size chunking: Splitting text into uniform-length sections (e.g., 512-1024 tokens).
 - Semantic chunking: Using NLP techniques to split text based on logical divisions (headings, paragraphs, etc.).

b. Embedding Generation & Storage:

To enable efficient retrieval, the extracted text is converted into vector embeddings that capture semantic meaning.

- **Embedding Model:**
 - The model uses Nomic-Embed for vector representation of text.
- **Vector Database (pgVector):**
 - Embeddings are stored in a PostgreSQL + pgVector database for efficient similarity search.
 - Each document chunk is indexed for fast retrieval during user queries.

c. Query Processing & Retrieval:

When a user asks a question, the chatbot retrieves the most relevant sections from the stored document.

- **Query Embedding:**
 - The user's input data which is raw data converted into a unique embedding using the same model used for document embeddings.
- **Similarity Search:**
 - Cosine similarity search is performed in the vector database to find the closest matching chunks.
- **Ranking & Filtering:**
 - Retrieved chunks are ranked based on relevance, and low-confidence matches are filtered out.

d. Response Generation using LLaMA-3:

Once the relevant documents are retrieved, they are passed to the Llama-3 model to generate informative and context aware responses for the user query [22]. Here you can instruct the chatbot with some certain instructions on how to provide the responses or in what way they are needed. Llama-3 is the best open-source model which provides the most accurate and context aware responses [23].

e. System UI and Deployment:

To provide a user-friendly experience, the chatbot is deployed using a Streamlit-based front-end. For Containerization the entire application is packed in the Docker ensuring easy deployment on cloud services [24]. By making this Containerization your system is not affected by the other works that you do and they remain individually separate.

This approach guarantees an effective search of information by the chatbot, the proper output, and the ability to create a continuous user experience. Through LLaMA-3, vector databases, and sophisticated retrieval methods, the system experiences limitations in the conventional chatbots, thus it is scalable, context-sensitive, and highly efficient [25].

System Architecture:

Below Figure 1: shows the architecture of the chatbot and its working mechanism. It is designed to combine efficient information retrieval with powerful generative capabilities to deliver accurate and context-aware responses. The system follows a structured pipeline, as illustrated in the diagram.

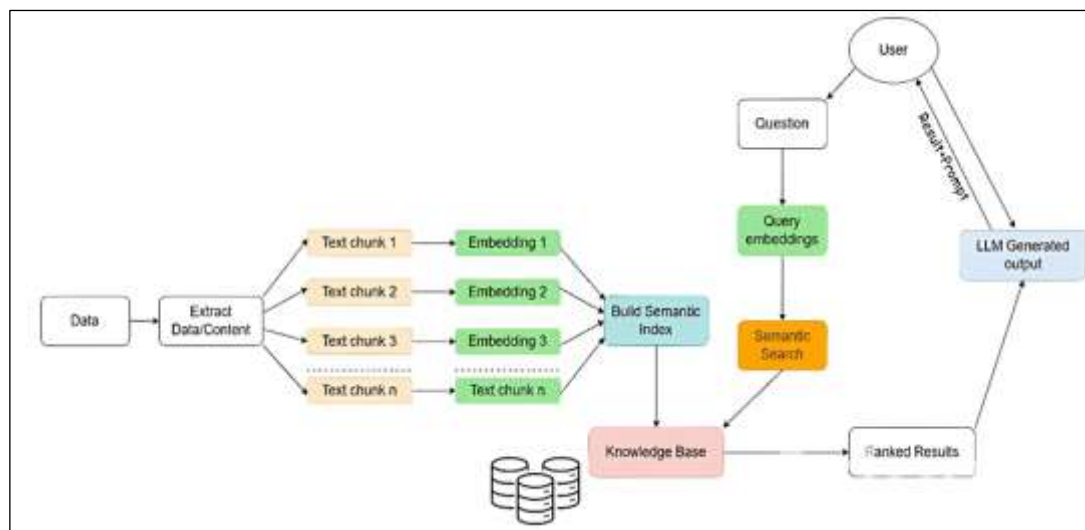


Figure 1: The architecture of the RAG chatbot showing the modules encountered at each.
Source: Authors, (2026).

The data is first extracted from the user provided documents and then it is divided into chunks of data which is changeable, Later the chunked data is sent to create embeddings and then to create a unique semantic index for those embeddings which helps in the retrieval part when there is a query. All of this data is stored into the knowledge base and a similar process will happen to the query to process all those and to understand the context of the question asked by the user. One of the important steps is:

a. Embedding

Nomic-Embed-Text is a text embedding model which converts the low-level textual data into the high-level vector representation. These embeddings further help in the searching process of the information that you have been stored in the database. Table 1 shows that the nomic embed text embedding model used in the research performs better than the other models and Table 2 depicting the Similarity metrics of the different models.

Table 1: Comparison table of different Embedding Models.

Feature	Open Ai Embeddings	Hugging Face Sentence Transformers	Nomic Embed (Used in Work)	FastText
Type of Embedding	Dense, contextual	Dense, contextual	Dense, contextual	Dense, word-level
Input Format	Text (any format)	Text (any format)	Text (any format)	Text
Pretrained Models	Yes (OpenAI proprietary)	Yes (wide variety)	Yes (e.g., nomic-embed-text)	Yes (language-specific)
NLP Context Awareness	High (deep contextual embeddings)	High (contextual sentences)	High	Low (word-level only)
Ease of Use	Easy	Moderate	Moderate	Easy
Speed	Moderate	Moderate	Fast	Fast
Best For	Advanced contextual embeddings	Sentence or document embeddings	Vector-based RAG pipelines	Word-level tasks
Cost	Paid (token-based pricing)	Free/Paid	Free	Free

Source: Authors, (2026).

Table 2: Performance metrics comparison of the Embedding Model.

Model	Embedding Size	Cosine Similarity (Avg)	Euclidean Distance (Avg)	Manhattan Distance (Avg)	Dot Product (Avg)
Hugging Face Sentence Transformers	768	0.82	1.45	2.89	11.76
Nomic Embed (Used in Work)	768	0.88	1.12	2.34	13.21
Fast Text	300	0.78	1.67	3.02	10.45

Source: Authors, (2026).

As we can see that Nomic embed text has performed better in each of those metrics as Higher cosine similarity, Lower Euclidean distance, Lower Manhattan distance, Higher dot product is better. Nomic embed text performed better in each of these performance metrics.

b. Knowledge Base

The knowledge base in this research work is PostgreSQL database with pgVector. The pgvector extension in this database allows searching the embeddings very fast and efficiently. This also helps in retrieving the embeddings from the database. This uses semantic search to search for the matching embeddings. This kind of search is much more ideal for the large-scale ML applications

Average Text Chunk Size: ~512 tokens per chunk

Vector Dimension: 768

Storage Size Per Embedding: ~3-4 KB per vector

The chatbot needs to store the data and retrieve it as quickly as possible. In case of the usage of the normal database it takes a lot of time to store and even to accurately search and retrieve the information. So, in this place the vector database comes which stores the data in the vector format and pgvector further improves the retrieval.

Working mechanism of pgVector:

i. Storing Embeddings:

- After converting text chunks into embeddings using Nomic-Embed-Text, they are stored in a PostgreSQL table with a vector column.

ii. Indexing for fast retrieval:

- A semantic index is created for each of the items for ensuring fast retrieval. This speeds up the search process.

iii. Retrieving Relevant Information:

- When a user asks a question, the input is converted into an embedding and compared to the stored vectors.
- The chatbot retrieves the most similar chunks and uses them as context for LLaMA-3.

c. Semantic Search

Traditional search methods rely on keyword matching, which can be inefficient when dealing with complex queries. Semantic search goes beyond keywords by understanding the meaning of the user's query and retrieving the most relevant results based on contextual similarity.

Vector Search with pgVector:

- This query embedding is compared against the stored document embeddings in pgVector using similarity measures like cosine similarity.

$$\cos \theta = \frac{A \cdot B}{||A|| \ ||B||} \quad (1)$$

- Instead of searching exact words, it retrieves the most contextually similar chunks from your knowledge base.

d. LLM Response Generation

Once semantic search retrieves the most relevant document chunks, they are sent as context to your LLaMA-3 model, along with the user's original question. The LLM then processes this information and generates a natural language response.

How the chatbot handles LLM response generation:

Input to LLaMA-3:

- The user's query: "What are the symptoms of diabetes?"
- Retrieved knowledge base content:
 - Chunk 1: "Diabetes symptoms include frequent urination, frequent thirst, and weight loss."
 - Chunk 2: "Early signs of diabetes also involve fatigue and blurred vision."
- These are formatted into a structured prompt before being sent to LLaMA-3.

Processing the data by Llama-3:

- LLaMA-3 analyzes the provided content and formulates a well-structured response in a conversational manner.
- Instead of hallucinating an answer, it summarizes and reformulates the retrieved information. The main concept of using the RAG is that the chatbots do not hallucinate, they have a predefined knowledge on what they are about to give an answer.

Final Response:

- Example Output: "Diabetes symptoms typically include frequent urination, excessive thirst, weight loss, fatigue, and blurred vision. Early detection is important to manage the condition effectively."

e. Llama-3 and its significance

LLaMA-3 (Large Language Model by Meta AI) is the latest invention of Meta's open-weight large language models, designed for better natural language processing (NLP), retrieval-augmented generation (RAG), and knowledge-based AI applications.

The main reason to select the Llama-3 model is that it is open source and comes with various other uses like

- Efficiency – It offers high performance while using fewer computational resources compared to closed models like GPT-4.
- Open-Source – Unlike proprietary models, LLaMA-3 is freely available for customization and fine-tuning.
- Scalability – Supports a range of applications, from small-scale personal assistants to enterprise-level AI.
- Optimized for Long Context – LLaMA-3 can process up to 16,000 tokens in a single prompt, allowing it to handle long-form documents effectively.

Llama-3 has been trained on over 1.5 trillion tokens from diverse multilingual and domain-specific sources. It can also generate responses within 500 ms - 2s, depending on prompt complexity and hardware.

IV. RESULTS AND DISCUSSIONS

This section analyses about the chatbot data ingestion and its processing of it. The data ingestion usually depends on the size of the document the user has provided. It also discusses the chunking mechanism and how the chatbot thinks in the background of its processing. This also shows the average time taken for each of the response generated by the chatbot and also follows the instructions set previously.

Chatbot UI and Design:

The entire frontend of this research has been made with the use of streamlit which is an open-source Python framework. It is mainly divided into two sections, The chat interface for the user to chat with the chatbot and the customization section where the user can upload the documents and change the Llama model, change the embedding or clear the knowledge base. The entire interface looks like the Figure 2.

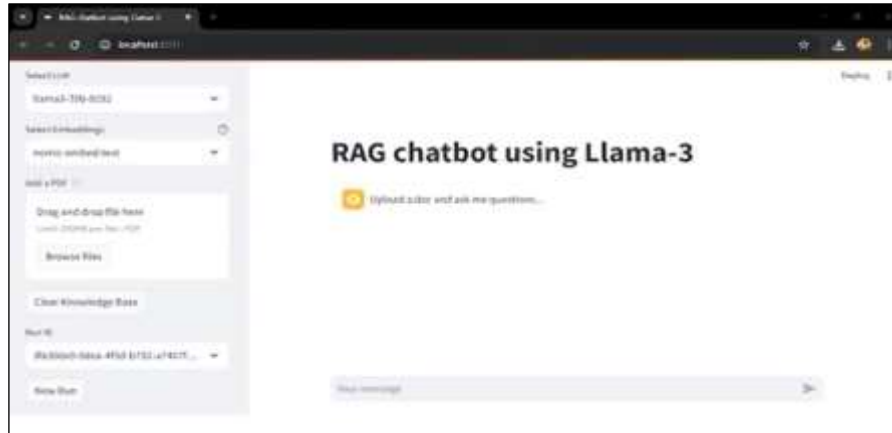


Figure 2: User Interface of the chatbot showing the different options available.
Source: Authors, (2026).

a. LLM Selection

In this proposed work there are 3 types of models from which you can select one on which your chatbot works and generates responses, The models are

i. Llama3-70b-8192

- 70B: It has 70 billion parameters, making it highly capable for complex reasoning and generation.
- 8192: It refers to the context window size (8,192 tokens), which determines how much text it can process in one go.

ii. Llama3-8b-8192

- 8B: A smaller version of LLaMA 3 with 8 billion parameters.
- 8192 tokens: Same context window as the 70B version.

iii. Mixtral-8x7b-32768

- Mixtral (Mistral AI): A mixture of experts (MoE) model developed by Mistral AI.
- 8×7B: Uses 8 experts, each with 7 billion parameters, but only 2 experts are active per inference, meaning 14B active parameters at a time.
- 32768 tokens: Extremely large context window, allowing it to process long documents or conversations.

We have also put the Mixtral model for testing purposes and to compare the results of the Llama-3 with this model. The model can be selected from the drop down which looks like the Figure 3.



Figure 3: Showing the options available for the selection of the different models
Source: Authors, (2026).

b. Uploading the document:

The user can upload a file from the drag and drop section and the chatbot takes the input and processes them further for chunking and embedding.



Figure 4: Drag and drop section for the files to be uploaded by the use.
Source: Authors, (2026).

After uploading the document, the chatbot processes the text inside it and divides them into several chunks. This process can be checked in the step in the terminal that you have used to start, it looks like the Figure 5.

```
DEBUG Upserted document: 20-health-tips_1_1 | 20-health-tips | {'page': 1,
DEBUG 'chunk': 1, 'chunk_size': 512}
DEBUG Upserted document: 20-health-tips_1_2 | 20-health-tips | {'page': 1,
DEBUG 'chunk': 2, 'chunk_size': 505}
DEBUG Upserted document: 20-health-tips_1_3 | 20-health-tips | {'page': 1,
DEBUG 'chunk': 3, 'chunk_size': 505}
DEBUG Upserted document: 20-health-tips_1_4 | 20-health-tips | {'page': 1,
DEBUG 'chunk': 4, 'chunk_size': 511}
DEBUG Upserted document: 20-health-tips_1_5 | 20-health-tips | {'page': 1,
DEBUG 'chunk': 5, 'chunk_size': 501}
DEBUG Upserted document: 20-health-tips_1_6 | 20-health-tips | {'page': 1,
DEBUG 'chunk': 6, 'chunk_size': 170}
DEBUG Upserted document: 20-health-tips_2_1 | 20-health-tips | {'page': 2,
DEBUG 'chunk': 1, 'chunk_size': 509}
DEBUG Upserted document: 20-health-tips_2_2 | 20-health-tips | {'page': 2,
DEBUG 'chunk': 2, 'chunk_size': 495}
DEBUG Upserted document: 20-health-tips_2_3 | 20-health-tips | {'page': 2,
DEBUG 'chunk': 3, 'chunk_size': 511}
```

Figure 5: Division of text into chunks of different sizes.
Source: Authors, (2026).

Outputs:

The below figure shows about how the chatbot has generated a response about the document that the user has uploaded which is “20 health tips” which consists of simple 20 health tips which we need to follow.



Figure 6: Response generated by the chatbot when the user prompts it regarding a document that has been uploaded.
Source: Authors, (2026).

b. Response Time

The time taken to generate response for this certain prompt is 1.2404 seconds, as the time may change depending upon the prompt complexity, data retrieval time and the processing time.

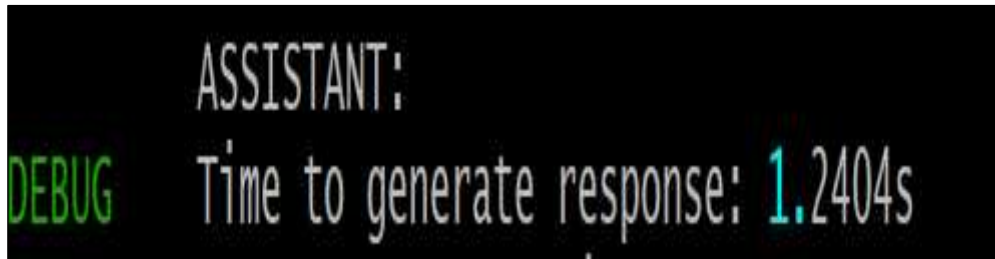


Figure 7: Response time taken by the chatbot.

Source: Authors, (2026).

For deeper and lengthy responses, the time taken to generate may increase but it is necessary for the chatbot to not hallucinate and to provide a context aware response. Table 3 depicts the features of the existing work with our current work and showcasing the advancements made.

Table 3: Comparison of Proposed work with existing work.

Feature	Existing Work	Proposed work	Improvement
LLM Used	LLaMA-2	LLaMA-3	LLaMA-3 is more accurate, efficient, and better at reasoning
Knowledge Retrieval	Uses FAISS for vector search	Hybrid retrieval (Vector + Keyword search)	Better search accuracy and relevant document retrieval
Fine-tuning	Not fine-tuned, uses base model	Can be fine-tuned or adapted for different domains	More flexibility for different applications
Embedding Model	Sentence Transformer	Nomic-embed-text (More optimized for retrieval)	Better document understanding and query matching
Context Retention	Limited memory of past interactions	Implements vector-based memory for multi-turn conversations	Better long-term conversation handling
Real-Time Updates	Static knowledge base (Pre-fed data)	Can dynamically update knowledge base from new documents	Supports real-time knowledge ingestion
Deployment Flexibility	Works on CPU but limited efficiency	Supports both CPU and GPU inference with optimizations	Better scalability
User Interface	ChainLit-based UI	Streamlit	More customizable and modern UI options
Domain-Specific Adaptation	Focused on medical domain	Domain-agnostic (Can adapt to finance, legal, etc.)	More use cases beyond healthcare
Embedding Model	Nomic-embed-text (assumed)	msmarco-bert-base-dot-v5	Potentially More Advanced

Source: Authors, (2026).

Table 4 shows the comparison of the different metrics of the proposed system with the existing system.

Table 4: Numerical comparisons of existing work with the current work.

Feature	Proposed system	Existing system	Improvement
LLM Model	LLaMA-3	LLaMA-2-7B	Upgraded Model
Context Length (tokens)	8000	4096	+95%
CPU Usage (%)	~42.5%	~75%	43.33%
Vector Dimensions	768	68	1030%
Response time (avg)	~ 4 sec	~ 9 sec	57%

Source: Authors, (2026).

V. CONCLUSIONS

In this research, we successfully developed a Retrieval-Augmented Generation (RAG)-based chatbot using the LLaMA-3 model. By integrating retrieval mechanisms with a powerful language model, the chatbot delivers contextually relevant and accurate responses, significantly improving user interaction. With various optimizations we ensured the system maintains efficiency and responsiveness. The use of vector databases and embedding techniques further strengthened its retrieval capabilities, making the chatbot more reliable in providing precise information. Overall, this chatbot showcases the potential of RAG models in bridging the gap between static knowledge bases and dynamic conversational AI.

V.1 LIMITATIONS

Present chatbot has few limitations like if the document contains any images without the text then it cannot process them as the chatbot is built only to understand the textual data. Current chatbot is not a summarizer of documents but it is just an information ingesting agent that accepts the data for storing them into the knowledge base, the knowledge you feed is just like the training data. So it is a drawback on what you feed the data. Llama-3 has a limit of context window, due to which it is set to a fixed token limit. Due to which it cannot process a large number of documents at once. There is no real time updating of the data in the knowledge base of the chatbot, it completely relies on the knowledge provided by the user.

V.2 FUTURE WORK

The future work of our research includes understanding the images, as the proposed work only processes the text and stores them in the knowledge base. Improving the chatbot to understand about the images will help improve the performance of the chatbot. Another most effective work is to implement continuous learning with the online sources which helps improve the chatbots knowledge base and keep in up to date with the real life updated information. Implementing text-to-speech and speech-to-text for voice based chat will significantly improve the chatbot performance and also makes the user experience better.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Juthuka Aruna Devi, T. Anusha, Ramesh Makala, Pala Pooja Ratnam and Marada Srinivasa Rao.

Methodology: Juthuka Aruna Devi, Krishna rupendra Singh, Rajendra Kumar Ganiya and Marada Srinivasa Rao.

Investigation: Krishna rupendra Singh, Rajendra Kumar Ganiya and Marada Srinivasa Rao.

Discussion of results: Juthuka Aruna Devi, Krishna rupendra Singh, Rajendra Kumar Ganiya and Marada Srinivasa Rao.

Writing – Original Draft: Pala Pooja Ratnam, Krishna rupendra Singh, Rajendra Kumar Ganiya and Marada Srinivasa Rao.

Writing – Review and Editing: Juthuka Aruna Devi, Krishna rupendra Singh, Rajendra Kumar Ganiya and Marada Srinivasa Rao.

Resources: Juthuka Aruna devi, T. Anusha, Pala Pooja Ratnam, Krishna rupendra Singh and Marada Srinivasa Rao.

Supervision: T. Anusha, Ramesh Makala, Pala Pooja Ratnam, Krishna rupendra Singh and Marada Srinivasa Rao.

Approval of the final text: Juthuka Aruna Devi, Krishna rupendra Singh and Marada Srinivasa Rao.

VII. REFERENCES

- [1] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology*, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham: Springer, DOI: 10.1007/978-3-030-49186-4_31, 2020.
- [2] H. Bansal and R. Khan, "A review paper on human-computer interaction," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 8, p. 53, 2018.
- [3] B. A. Abu Shawar and E. S. Atwell, "Chatbots: are they really useful?," *J. Lang. Technol. Comput. Linguist.*, vol. 22, pp. 29–49, 2007.
- [4] L. Klopfenstein, S. Delpriori, S. Malatini, and A. Bogliolo, "The rise of bots: a survey of conversational interfaces, patterns, and paradigms," in *Proc. 2017 Conf. Designing Interact. Syst., Assoc. Comput. Mach.*, pp. 555–565, 2017.
- [5] P. B. Brandtzaeg and A. Folstad, "Why people use chatbots," in *Internet Science*, I. Kompatsiaris et al., Eds. Cham: Springer, pp. 377–392, 2017.
- [6] G. Molnár and S. Zoltán, "The role of chatbots in formal education," Presented at the 15 September, 2018.
- [7] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university-related FAQs," in *Proc. 2017 Int. Conf. Adv. Comput. Commun. Informatics (ICACCI)*, Udipi, pp. 1525–1530, 2017.
- [8] P. C. F. da Costa, "Conversing with personal digital assistants: on gender and artificial intelligence," *J. Sci. Technol. Arts*, vol. 10, pp. 59–72, 2018.
- [9] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A survey of design techniques for conversational agents," in *ICICCT 2017, Commun. Comput. Inf. Sci. (CCIS)*, vol. 750, S. Kaushik, D. Gupta, L. Kharb, and D. Chahal, Eds. Singapore: Springer, pp. 336–350, 2017.
- [10] K. Nimavat and T. Champaneria, "Chatbots: an overview, types, architecture, tools, and future possibilities," *Int. J. Sci. Res. Dev.*, vol. 5, pp. 1019–1024, 2017.
- [11] A. Khanna, B. Pandey, K. Vashishta, K. Kalia, P. Bhale, and T. Das, "A study of today's A.I. through chatbots and rediscovery of machine intelligence," *Int. J. u-e-Serv. Sci. Technol.*, vol. 8, pp. 277–284, 2015.
- [12] A. Kumar, P. K. Meena, D. Panda, and M. Sangeetha, "Chatbot in Python," *Int. Res. J. Eng. Technol.*, vol. 6, no. 11, 2019.
- [13] A. P. S. Sasan, A. Kumar, A. K. Singh, and A. Baruah, "A research paper on a medical chatbot using LLaMA 2," *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)*, vol. 12, no. IV, DOI: 10.22214/ijraset.2024.59868, 2024.
- [14] S. Vakayil, J. Anitha, D. S. Juliet, and S. Vakayil, "RAG-based LLM chatbot using LLaMA-2," in *Proc. 7th Int. Conf. Devices, Circuits, Syst. (ICDCS)*, DOI: 10.1109/ICDCS59278.2024.10561020, 2024.

- [15] M. M. Hossain, S. K. Pillai, S. E. Dansy, and A. A. Bilong, “Mr. Dr. Health-assistant chatbot,” *Int. J. Artif. Intell.*, vol. 8, no. 2, pp. 58–73, Dec. 2021.
- [16] P. Amiri and E. Karahanna, “Chatbot use cases in the COVID-19 public health response,” *J. Amer. Med. Inform. Assoc.*, vol. 29, no. 5, pp. 1000–1010, Apr. 2022.
- [17] P. Weber and T. Ludwig, “(Non-) interacting with conversational agents: Perceptions and motivations of using chatbots and voice assistants,” in *Proc. Conf. Mensch Comput.*, vol. 1, pp. 321–331, Sep. 2020.
- [18] R. Urbani, C. Ferreira, and J. Lam, “Managerial framework for evaluating AI chatbot integration: Bridging organizational readiness and technological challenges,” *Business Horizons*, vol. 67, no. 5, pp. 595–606, May 2024, doi: 10.1016/j.bushor.2024.05.004.
- [19] C. V. Mischia, F. Poecze, and C. Strauss, “Chatbots in customer service: Their relevance and impact on service quality,” *Procedia Computer Science*, vol. 201, pp. 421–428, Jan. 2022, doi: 10.1016/j.procs.2022.03.055.
- [20] O. a. G. Valencia et al., “Ethical implications of chatbot utilization in nephrology,” *Journal of Personalized Medicine*, vol. 13, no. 9, p. 1363, Sep. 2023, doi: 10.3390/jpm13091363.
- [21] K. H. Koundinya, A. K. Palakurthi, V. Putnala, and K. A. Kumar, “Smart college chatbot using ML and Python,” in *Proc. Int. Conf. Syst., Comput., Automat. Netw. (ICSCAN)*, pp. 1–5, Jul. 2020.
- [22] B. A. Shawar and E. Atwell, “Using dialogue corpora to train a chatbot,” in *Proc. Corpus Linguistics Conf.*, pp. 681–690, Mar. 2003.
- [23] G. Battineni, N. Chintalapudi, and F. Amenta, “AI chatbot design during an epidemic like the novel coronavirus,” *Healthcare*, vol. 8, no. 2, pp. 1–8, Jun. 2020.
- [24] S. Bhattacharya and B. Singla, “Exploring the secrets behind chatbot success in modern banking,” in *Design and Development of Emerging Chatbot Technology*, pp. 297–316, DOI: 10.4018/979-8-3693-1830-0.ch017, Mar. 2024.
- [25] N. Rosruen and T. Samanchuen, “Chatbot utilization for medical consultant system,” in *Proc. 3rd Technol. Innov. Manag. Eng. Sci. Int. Conf. (TIMES-iCON)*, pp. 1–5, Dec. 2018.