2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) | 979-8-3315-1088-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICHORA65333.2025.11017017

A Comprehensive RAG-Based LLM for AI-Driven Mental Health Chatbot

Anish Ilapaka
Dept. of Engineering and Computing Sciences
Arkansas Tech University
Russellville, Arkansas, USA
ailapaka@atu.edu

Robin Ghosh
Dept. of Engineering and Computing Sciences
Arkansas Tech University
Russellville, Arkansas, USA
rghosh@atu.edu

***Abstract*—Mental health challenges affect millions worldwide, with approximately 970 million people experiencing conditions such as anxiety and depression. In the United States alone, 57.8 million adults struggle with mental illness. Despite the growing need for support, many face barriers such as stigma, limited access to care, and long waiting times, especially in underserved areas with few mental health resources. AI-powered chatbots are an emerging intervention option that offers available, accessible, and confidential services. The study below describes the development of an AI chatbot to support people in distress, using a fine-tuned version of the Llama model, which was trained on actual mental health counseling conversations. The chatbot uses Retrieval-Augmented Generation (RAG) to improve the relevance of its responses and make interactions more personalized. It also leverages LangChain's ConversationBufferMemory to recall past conversations, allowing for more natural and meaningful dialogue. Facebook AI Similarity Search (FAISS) and Sentence Transformer embeddings also allow the retrieval of relevant materials to provide practical, real-time coping strategies. This chatbot uses natural language processing (NLP) and machine learning (ML) to bridge the gaps in mental health care, making it more affordable, accessible, and free of stigma. The current paper discusses the effectiveness of chatbots, the challenges in AI-driven mental health support, and future improvements in deeper personalization, improved response refinement, and integration with multimedia resources for a more holistic user experience.**

***Keywords*—** N LP, Rag, langchain, Llama, FAISS, Sentence Transformers.

I. Introduction

Mental health challenges affect millions of people around the world, yet obtaining the proper support remains a struggle for many. According to the World Health Organization (WHO), nearly 970 million people experience mental health conditions, with anxiety and depression being the most common [1] in the United States; about 57.8 million adults, roughly one in five people, live with a mental illness each year [2]. In underserved communities, access to mental health care is even more limited, leaving people without the help they need [3]. Access to mental health support has always been a challenge for many individuals, often due to stigma, financial constraints, or lack of available professionals, especially in underserved communities. Despite growing awareness about mental health, millions of people still hesitate to seek help, fearing judgment or feeling that their struggles are not significant enough to warrant professional intervention. At the same time, long wait times for therapy, high costs, and an overall shortage of mental health professionals create additional barriers, leaving many without the support they need [4]. In today's fast-paced and often stressful world, people frequently experience anxiety, depression, or emotional distress but may not have someone to share the moment. The feeling of being unheard or struggling alone can intensify mental health challenges, making even small obstacles feel overwhelming. This is where technology is beginning to bridge the gap, providing people with an immediate, judgment-free space to express their thoughts, process their emotions, and find guidance when traditional support systems are unavailable [5]. Digital tools for mental health support are becoming increasingly sophisticated, offering users a safe space to talk, reflect, and access resources. These tools can help individuals track their emotions, recognize patterns in their thoughts, and explore coping strategies tailored to their needs. By guiding users through self-reflection exercises, mindfulness practices, and evidence-based mental health techniques, these platforms empower individuals to manage their emotional well-being actively. The availability of such tools is particularly significant for those who may not be ready or able to seek therapy but still want support in navigating their mental health challenges.

While digital solutions are not a replacement for professional therapy, they are a crucial first step in encouraging individuals to acknowledge their feelings and seek help. For many, these platforms offer a private and accessible way to begin their mental health journey, providing reassurance that they are not alone. By reducing the stigma associated with mental health struggles and making support more widely available, these tools are contributing to a cultural shift, where seeking help is no longer seen as a weakness but as an essential part of self-care and personal growth. The central idea of the research is the creation of a new AI mental health chatbot. The chatbot developed in this study uses a fine-tuned LLaMA 3.2 3b parameters model. It was trained on real mental health conversations from huggingface and refined with Unsloth and Ollama to enhance its performance. It also uses LangChain's ConversationBufferMemory for the user's interactions and FAISS with Sentence Transformer embeddings for resource recommendations. The paper described the chatbot's design, implementation, and effectiveness. It also describes the

979-8-3315-1088-6/25/$31.00 © 2025 IEEE

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:48:30 UTC from IEEE Xplore. Restrictions apply.

model fine-tuning and conversation management techniques in technical detail and addresses the ethical issues of using AI in mental health, such as privacy and safety. We also discussed the limitations of AI support and suggested future improvements, such as improved personalization, multimedia integration, and crisis detection. This research attempted to enhance AI-based mental health care by developing a product that can be easily scaled up. It also examined how digital tools, particularly conversational platforms, could transform mental health support by making it more immediate, inclusive, and adaptable to individual needs.

II. Methodology

A. Data Collection and Preparation

To develop the mental health support system, we gathered a dataset of genuine counseling conversations from Hugging Face [6], comprising structured interactions where individuals seek guidance on emotional distress, anxiety, and other mental health concerns. Before using the data set for training, we applied various pre-processing steps to enhance consistency and optimize the learning of the model. First, we cleaned the data by removing duplicate or incomplete records and standardizing text formatting. Then, we tokenized the text, breaking it down into smaller segments for efficient processing. Additionally, we structured conversations to clearly distinguish between patient concerns and counselor responses. Finally, the dataset was split into training (90%) and validation (10%) sets to effectively monitor the model's performance, ensuring that the training material remained structured, relevant, and suitable for developing a high-quality conversational model.

B. Model Development and Training Approach

To build a responsive and context-aware chatbot, we used Unsloth, an optimized framework that enhances the efficiency of training large language models [7]. Unsloth significantly improved speed and computational efficiency, enabling us to train a LLaMA 3.2 3B model on our mental health dataset while minimizing resource consumption. Traditional model fine-tuning methods require substantial computational power and long training times; however, Unsloth streamlined this process through memory optimization and improved training speed. Key advantages included memory efficiency by enabling 4-bit quantization, faster training with gradient checkpointing, and the implementation of Low-Rank Adaptation (LoRA) [8] to fine-tune the model without modifying its core structure, making it more adaptable for mental health applications.

C. Training Process

The training process commenced with initializing the model using FastLanguageModel from Unsloth, which optimized loading and fine-tuning for LLaMA models. We applied LoRA to fine-tune specific layers without altering the entire model, enhancing response generation. A Supervised Fine-Tuning Trainer (SFTTrainer) was employed to refine the model's ability to generate empathetic and contextually relevant responses. The training was conducted with a batch size of 2, a learning rate of 2e-4, and 60 training steps to mitigate overfitting. The dataset was divided into training and validation sets, ensuring the model effectively generalizes to new user inputs.

D. Model Quantization for Efficient Deployment

To optimize the deployment of the mental health AI model, we applied quantization using the method Q5-k-m while saving it in GGUF format. Quantization is a technique that reduces the numerical precision of model weights, lowering memory usage and improving inference speed. By converting high-precision floating-point values (e.g., FP16) to lower-bit representations (e.g., 5-bit), quantization enables efficient execution on resource-constrained hardware. While this process may introduce minor accuracy trade-offs, it significantly enhances computational efficiency, making large-scale models feasible for real-world applications without requiring dedicated GPUs. This optimization is particularly beneficial for real-time mental health AI systems, where responsiveness and accessibility are critical.

E. Retrieval Pipeline and Techniques

Once training was complete, we implemented a retrieval pipeline to enhance the chatbot's ability to provide meaningful and contextually relevant responses. Instead of relying solely on predefined responses, the system dynamically retrieves relevant information from a knowledge base, ensuring that users receive responses based on real mental health guidance rather than generic answers.

F. Retrieval Techniques Used

To improve response accuracy and quality, several retrieval techniques were employed. FAISS [9] was used to accelerate similarity searches, enabling the chatbot to find relevant responses efficiently, even when dealing with large datasets. Sentence Transformer embeddings were implemented to help the chatbot understand the semantic meaning behind user queries instead of merely relying on keyword matching. This approach ensures that retrieved responses genuinely align with the user's concerns. Additionally, we integrated ConversationBufferMemory [10], allowing the chatbot to retain context across multiple interactions, ensuring continuity and coherence in conversations. By incorporating these retrieval techniques, the chatbot delivers relevant, informed, and personalized responses, bridging the gap between AI-driven interactions and real mental health support. This system ensures users receive tailored guidance for their concerns while maintaining a seamless and natural conversational experience [11].

III. Implementation

The development of the mental health chatbot follows a structured approach to ensure it provides meaningful and empathetic responses. Combining hybrid retrieval techniques with contextual memory allows the chatbot to understand user input and generate relevant replies. This section outlines the key implementation steps, including how information is retrieved, processed, and used to create responses.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:48:30 UTC from IEEE Xplore. Restrictions apply.

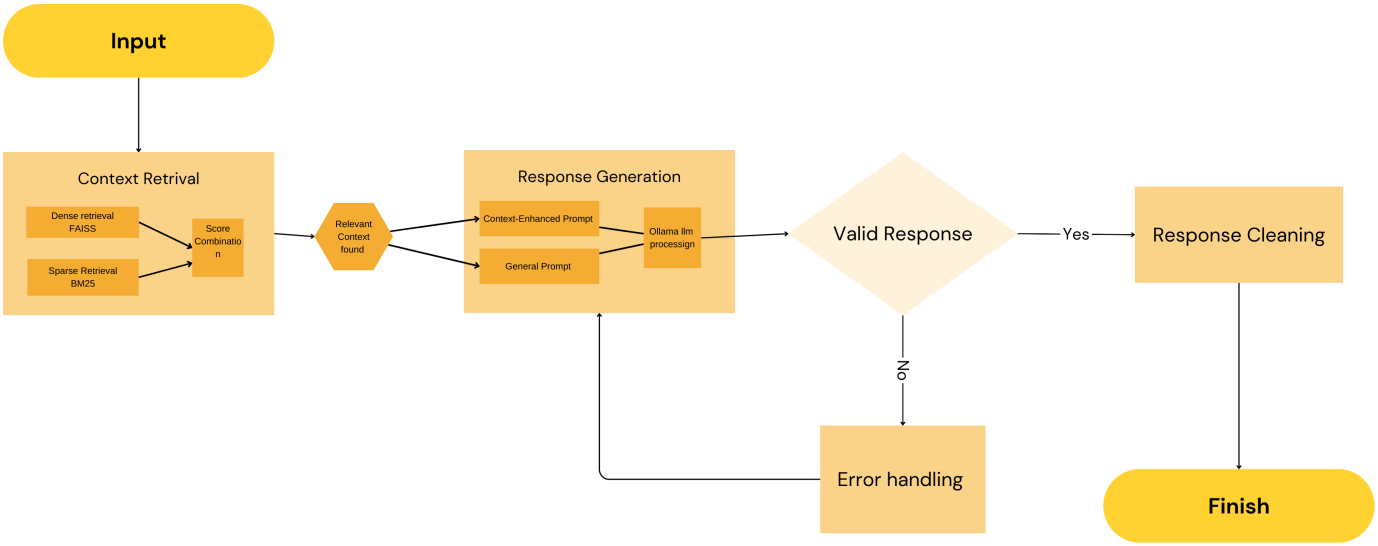

Fig. 1. System architecture flowchart of the mental health AI model.

A. Transforming Text into Searchable Data

Text-based mental health conversations are converted into numerical representations called embeddings to match user queries effectively with relevant responses. These embeddings capture the deeper meanings and relationships between words, allowing the system to recognize context even if the exact words differ. A Sentence Transformer model generates these embeddings stored in a FAISS database. This method ensures that relevant information can be retrieved efficiently when needed.

B. Retrieving Relevant Information

The chatbot uses a hybrid retrieval method, combining two distinct approaches: Semantic Search (Dense Retrieval) identifies responses based on meaning rather than exact word matches. This is instrumental in mental health conversations, where users may phrase their concerns differently. Keyword-based search (Sparse Retrieval with BM25) prioritizes responses containing exact keywords from the user's input. BM25 ranks documents based on how frequently important terms appear [12].

C. Selecting the Best Response

The system aggregates scores from semantic and keyword-based searches to balance both retrieval approaches. It assigns weights to each result and selects responses that align with the meaning and specific words used by the user. This hybrid method ensures that replies are both accurate and contextually relevant. (Retrieval-Augmented Generation for Natural Language Processing: A Survey, 2025) [13].

D. Managing Context with Memory

Conversations often require an understanding of past messages to maintain coherence. The chatbot integrates memory management through LangChain's ConversationBufferMemory, which allows it to track user interactions. This enables the system to recall previous exchanges for a consistent conversation flow, adjust responses based on ongoing discussions, and provide answers that acknowledge the user's emotional state over time. By incorporating memory, the chatbot can offer a more personalized and thoughtful interaction rather than treating each message in isolation.

E. Generating Meaningful Responses

Once the system retrieves relevant information, it constructs a response by incorporating the user's query, previously retrieved context, chat history (if applicable), and a tone that reflects empathy and understanding.

IV. Evaluations and Results

A. Analysis of Results Matrix

The results matrix evaluates the performance of the mental health AI model across four key attributes: Relevance, Empathy, Conciseness, and Context. These metrics were chosen to ensure that the chatbot delivers meaningful, compassionate, and contextually appropriate responses while maintaining brevity for user engagement. Relevance (0.78): This score reflects how well the AI-generated responses align with the user's input. A score of 0.78 suggests that the model effectively understands and addresses user queries but has some room for improvement in tailoring responses more precisely. Empathy (0.84): The chatbot demonstrates strong empathetic responses, scoring the highest among the four attributes. This indicates that the model successfully conveys understanding and emotional support, which is crucial in mental health applications. Conciseness (0.72): The model maintains brevity in its responses, avoiding unnecessary elaboration. However, the score suggests that there is a slight tendency toward either overly brief or slightly verbose answers, which can be refined for a better user experience. Context (0.77): This score measures how well the AI retains and utilizes prior

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:48:30 UTC from IEEE Xplore. Restrictions apply.

conversation history. A score of 0.77 indicates that the chatbot generally maintains context well but may occasionally struggle with long or complex conversations.

TABLE I
EVALUATION RESULTS OF THE MENTAL HEALTH AI MODEL

Attribute	Score
Relevance	0.78
Empathy	0.84
Conciseness	0.72
Context	0.77

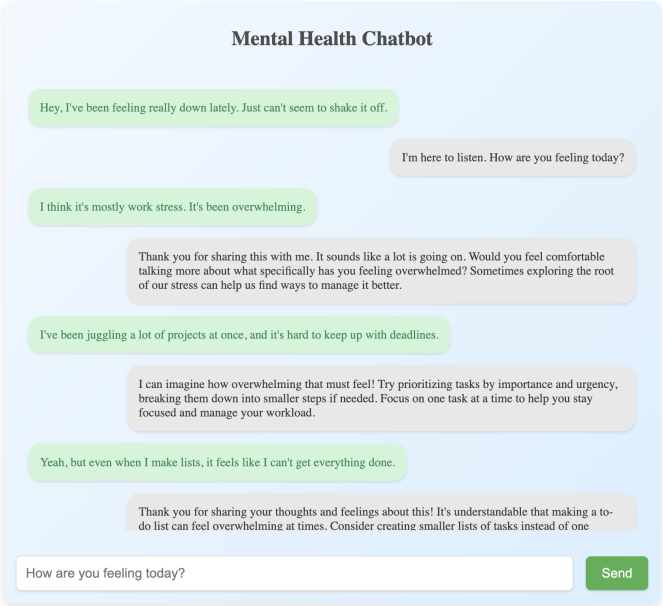

Fig. 2. UI of Mental Health Chat-bot.

V. CHALLENGES & LIMITATIONS

A. Limited Computing Resources

Training a complex language model requires significant processing power, which can be a constraint when working with limited hardware and financial resources. Although Unsloth made it possible to train the model more efficiently on available systems, balancing the size, performance, and cost of the model remained a challenge.

B. Misinformation & Inaccurate Responses

One of the most challenging issues with AI-generated text is the possibility of producing misleading or incorrect information. In mental health support, even minor inaccuracies can have serious consequences. Despite extensive fine-tuning, the chatbot sometimes generated responses that sounded reasonable but were not entirely accurate, making regular oversight and quality control essential [14].

C. Addressing Bias in Responses

Every AI system learns from the data it is trained on, which means inherent biases in the dataset can influence its responses. Mental health conversations are profoundly personal and vary across cultures, communities, and individuals. Efforts have been made to use diverse datasets and fine-tune the model to be more inclusive, but eliminating bias is an ongoing process (Model Cards for Model Reporting, 2019) [15].

D. Maintaining Conversation Context

A significant challenge in chatbot interactions is ensuring the AI remembers past exchanges within a conversation. Although LangChain's memory management helped retain context in short dialogues, longer conversations occasionally led to inconsistencies. This was addressed by implementing retrieval-augmented generation (RAG) techniques, but further improvements are needed for seamless, long-term interactions [16].

CONCLUSION (FUTURE WORK & IMPROVEMENTS)

As technology advances, the future of AI-driven mental health support looks promising. Moving forward, there is a need to enhance the chatbot's ability to remember past conversations and provide long-term, contextually relevant responses. Another area for growth is expanding its capabilities to interpret and respond to multiple forms of communication, such as voice or visual cues, which could allow the system to understand better and engage with users. Additionally, improving the chatbot's ability to provide factually accurate answers and reducing instances of generating incorrect or misleading information will be crucial for building trust and reliability. Personalization is also essential, as the chatbot could be further developed to adapt its responses based on each user's unique preferences and needs. Expanding language support will ensure the chatbot can cater to diverse communities, offering assistance in various languages and cultures. Lastly, enhancing privacy measures and ensuring that ethical guidelines are followed will be essential to safeguarding users' well-being and trust in the system.

REFERENCES

[1] (2025, February 10). "Mental Health," World Health Organization. Retrieved from https://www.who.int/health-topics/mental-health#tab=tab_1.

[2] NAMI, (2021, January 20). "NAMI Research News," NAMI. Retrieved from https://www.nami.org/about-mental-illness/research/research-news/2021-2/.

[3] (2025, February 10). "World Mental Health Report," World Health Organization. Retrieved from https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report.

[4] H, M., (2023, March 30). "What's the Value in Value-Based Care?" Exploring Barriers to Mental Health Care in the U.S. Retrieved from https://doi.org/10.15766/rai_a3ewcf9p.

[5] (2023, March 01). "Digital Technologies for Mental Health Improvements in the COVID-19 Pandemic: A Scoping Review," BMC Public Health. Retrieved from https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-15302-w?utm_source=chatgpt.com.

[6] (n.d.). "Mental Health Counseling Conversations," HuggingFace. Retrieved from https://huggingface.co/datasets/Amod/mental_health_counseling_conversations.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:48:30 UTC from IEEE Xplore. Restrictions apply.

[7] (n.d.). “Unsloth,” GitHub. Retrieved from https://github.com/unslothai/unsloth.
[8] (2021, October 16). “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv.org. Retrieved from https://doi.org/10.48550/arXiv.2106.09685.
[9] Douze, M., (2024, September 06). “The Faiss Library,” arXiv.org. Retrieved from https://doi.org/10.48550/arXiv.2401.08281.
[10] (n.d.). “Langchain: Memory Buffer,” Langchain. Retrieved from https://python.langchain.com/v0.1/docs/modules/memory/types/buffer/.
[11] (2022, December 28). “Artificial Intelligence-Enabled Chatbots in Mental Health: A Systematic Review,” Tech Science Press. Retrieved from https://doi.org/10.32604/cmc.2023.034655.
[12] (2009, December 15). “The Probabilistic Relevance Framework: BM25 and Beyond,” http://dx.doi.org/10.1561/1500000019.
[13] (2025, February 11). “Retrieval-Augmented Generation for Natural Language Processing: A Survey,” arXiv. Retrieved from https://arxiv.org/html/2407.13193v1.
[14] (2025, February 11). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big,” ACM Digital Library. Retrieved from https://doi.org/10.1145/3442188.3445922.
[15] (2019, January 14). “Model Cards for Model Reporting,” arXiv. Retrieved from https://arxiv.org/abs/1810.03993.
[16] (2025, February 11). “Retrieval-Augmented Generation for Knowledge-Intensive Tasks,” NeurIPS Proceedings. Retrieved from https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:48:30 UTC from IEEE Xplore. Restrictions apply.