# CompanAI: An AI-Powered Document-Grounded Mental Health Chatbot Framework

*Pratham Gupta*[1]
*Computer Science and Engineering*
*Kalinga Institute of Industrial Technology*
*Bhubaneswar, India*
*21051233@kiit.ac.in*

*Abhishiek Kumar*[2]
*Computer Science and Engineering*
*Kalinga Institute of Industrial Technology*
*Bhubaneswar, India*
*21052248@kiit.ac.in*

*Subhadeep Shil*[3]
*Computer Science and Engineering*
*Kalinga Institute of Industrial Technology*
*Bhubaneswar, India*
*21051183@kiit.ac.in*

*Sagnik Dey*[4]
*Computer Science and Engineering*
*Kalinga Institute of Industrial Technology*
*Bhubaneswar, India*
*2106146@kiit.ac.in*

*Archit Jethlia*[5]
*Computer Science and Engineering*
*Kalinga Institute of Industrial Technology*
*Bhubaneswar, India*
*21053206@kiit.ac.in*

*Shubhadeep Sen*[6]
*Computer Science and Engineering*
*Kalinga Institute of Industrial Technology*
*Bhubaneswar, India*
*21051176@kiit.ac.in*

***Abstract - Mental health challenges are a growing concern in the modern digital age, exacerbated by limited access to timely and affordable professional support. The "CompanAI" project aims to bridge this accessibility gap by introducing a conversational AI chatbot designed to offer empathetic interactions and reliable, context-aware mental health information. The chatbot framework leverages advanced artificial intelligence components including state-of-the-art language models (Llama-3-70B) deployed via the Groq Cloud API, combined with document-grounded question answering powered by LangChain and ChromaDB. Using PyPDF and text splitting techniques, CompanAI enables seamless ingestion and interpretation of trusted mental health literature, supporting users with accurate and relevant responses. The front-end interface is built on Gradio, allowing an intuitive and user-friendly experience. The modular architecture ensures scalability, reliability, and adaptability for various mental health scenarios while upholding user privacy and response efficiency. This paper elaborates on the design, methodology, implementation, and evaluation of CompanAI and demonstrates its potential as a supplementary mental health support system.***

***Index Terms - Fine-tuning , chatbot, document-grounded question answering, LangChain, mental health support, vector database***

## *I.* INTRODUCTION

Mental health support systems are becoming increasingly vital in today's fast-paced, stress-laden world, where emotional well-being is continuously challenged by academic, professional, and societal pressures. With the World Health Organization estimating a sharp rise in mental health disorders, there is an urgent need to supplement traditional therapy with scalable technological solutions. However, the existing digital platforms often fall short in delivering authentic, empathetic, and reliable interactions, primarily due to their reliance on static rule-based algorithms or outdated AI architectures.

"CompanAI" is envisioned as a solution to these gaps—a mental health chatbot that combines document-grounded knowledge with fine-tuned large language models (LLMs) to deliver personalized, context-rich, and emotionally intelligent responses. Beyond traditional conversational AI, the system enhances its capabilities through continuous fine-tuning, ensuring that the model is better aligned with the specific linguistic patterns, emotional tones, and domain-relevant information unique to mental health scenarios. This fine-tuning process allows the Llama-3-70B model to adapt to more specialized datasets beyond its original pre-training, leading to improved semantic understanding and more empathetic dialogue generation.

The architecture of CompanAI leverages LangChain's modular design for managing conversational logic, Groq's ultra-low-latency cloud inference for deploying the fine-tuned LLM, and ChromaDB for document vector storage. This combination creates a reliable foundation for document-grounded question answering (DGQA), enabling the chatbot to fetch relevant information from curated mental health literature, while the fine-tuned LLM

generates fluent and emotionally sensitive responses. Additionally, the user interface is constructed using Gradio, ensuring the platform remains accessible to a wide range of users.

## II. Related Work

Significant research has been conducted on conversational AI applications, particularly in health and mental health domains. The development of mental health chatbots has been strongly influenced by the rise of artificial intelligence and natural language processing technologies, which enable machines to participate in human-like dialogue. Pioneering systems like Woebot and Wysa have showcased the potential for AI companions to assist users in managing stress, anxiety, and depression by delivering cognitive-behavioral therapy techniques through casual conversation. These systems emphasize emotional support, but often rely on pre-structured response flows and are limited in their ability to personalize replies based on deep contextual understanding.

The emergence of large language models (LLMs) such as OpenAI's GPT-series and Meta's Llama series introduced powerful improvements in conversational AI by allowing models to generate more coherent, contextually appropriate, and emotionally resonant responses. Fine-tuning these LLMs on specific datasets related to mental health, counseling, and patient support further enhances their alignment with human conversational norms in sensitive discussions.

Despite their success, few systems combine the flexibility of fine-tuned LLMs with the factual grounding capabilities of document-based retrieval mechanisms. Recent works in information retrieval and question answering have shown that grounding language models with real documents (like PDF research papers, therapy guidelines, and self-help resources) significantly boosts answer accuracy and user trust. CompanAI uniquely integrates fine-tuned LLMs with document-grounded context retrieval via LangChain, resulting in a hybrid system that offers both conversational fluency and factual relevance. This design provides a more reliable and emotionally intelligent support experience compared to static dialogue systems or purely generative LLM applications.

Additionally, research efforts surrounding vector databases like ChromaDB and embedding models like sentence-transformers have demonstrated the benefits of semantic similarity search for document-based question answering. CompanAI leverages these advancements to create a robust and scalable architecture capable of serving mental health-related queries with both precision and empathy.

## III. Literature Review

The application of artificial intelligence in mental health technology has seen considerable exploration over the past decade. Literature highlights the effectiveness of chatbots in enhancing mental health support and extending its accessibility to underserved populations. Early research emphasized rule-based systems for simple stress management and coping strategies, which offered predefined answers without genuine understanding or adaptability.

In contrast, studies on deep learning and transformer-based models revolutionized the scope of conversational AI. Radford et al. introduced the GPT series, which demonstrated powerful few-shot learning and contextual reasoning capabilities. Similarly, Meta's Llama models, especially the Llama-2 and Llama-3 families, showed significant improvement in generating human-like responses. Researchers have also established the importance of fine-tuning pre-trained language models on domain-specific data to ensure contextual accuracy and responsible output, especially for sensitive applications such as mental health.

Beyond generation, knowledge retrieval has been another focal point. Research by Lewis et al. on retrieval-augmented generation (RAG) outlined how document-grounded systems outperform purely generative approaches when answering specialized questions. Embedding techniques, as shown in the work of Reimers and Gurevych (2019) with Sentence-BERT, made it feasible to compute sentence-level semantic similarity, which further paved the way for scalable and meaningful information retrieval frameworks like ChromaDB.

Integrating these complementary technologies—fine-tuned LLMs for language understanding and generation, vector similarity sea=-rch for information retrieval, and human-centric design principles for ethical AI interactions—has emerged as a best practice in recent mental health chatbot research. CompanAI embraces these methodologies to deliver a reliable, empathetic, and knowledge-grounded mental health assistant.

## IV. Methodology

The system design of CompanAI focuses on ensuring efficient integration between the document retrieval mechanisms and the fine-tuned large language model (LLM) components. The architecture is structured as a two-stage pipeline comprising both retrieval and generation modules.

The first stage deals with document ingestion and embedding. Mental health literature, including research papers, guides, and therapy handbooks in PDF format, are

processed using PyPDFLoader to extract raw text. This extracted content is split into meaningful segments via RecursiveCharacterTextSplitter, ensuring compatibility with the embedding generation constraints of the sentence-transformers model.

Once the embeddings are generated using the all-MiniLM-L6-v2 model from HuggingFace, they are stored in ChromaDB, which supports vector-based retrieval. This enables fast and efficient similarity searches, allowing the system to fetch relevant document fragments based on a user's query.
In the second stage, when a user submits an input query via the Gradio interface, LangChain orchestrates the flow of data by retrieving top-matching document chunks from ChromaDB and preparing the input prompt for the LLM. The fine-tuned Llama-3-70B model hosted via Groq Cloud is responsible for synthesizing a coherent, empathetic, and context-aware reply, leveraging both its pretrained knowledge and the grounded document context.

This layered architecture promotes separation of concerns, scalability, and improved response accuracy while enabling CompanAI to provide grounded answers that align with best practices in mental health conversations.

The architecture of CompanAI is constructed around the fusion of natural language processing, vector-based information retrieval, and conversational AI frameworks:

A. Core Technologies

1. Python 3.8+: The programming backbone for developing the entire pipeline.

2. LangChain: Framework for designing LLM-based question-answering chains, managing document embeddings, and constructing conversational logic.

3. Groq Cloud API: Provides low-latency inference for the fine-tuned Llama-3-70B language model, facilitating high-quality text generation.

4. ChromaDB: A lightweight vector database designed for storing and querying dense document embeddings to enable semantic search.

5. Gradio: A web-based user interface library allowing fast deployment and user-friendly access to the chatbot.

B. Document Processing

Document processing is a crucial component of CompanAI's architecture, enabling it to provide reliable, context-aware answers based on trusted mental health literature. The document processing pipeline includes the following stages:

- Document Ingestion: Mental health resources, including research papers, therapy handbooks, and self-help guides in PDF format, are ingested using PyPDFLoader. This tool extracts raw text from PDFs, preserving the content's integrity for subsequent processing.

- Text Segmentation: Once the text is extracted, it is segmented into meaningful chunks using the RecursiveCharacterTextSplitter. This ensures that each segment respects token limits and maintains coherence, which is crucial for generating contextually accurate embeddings for semantic search.

- Embedding Generation: The text segments are transformed into embeddings using the all-MiniLM-L6-v2 model from HuggingFace's Sentence-Transformers library. These embeddings represent the semantic meaning of the text, enabling effective search and retrieval.

- Document Retrieval: The retrieved document fragments are passed along with the user's query to the fine-tuned Llama-3-70B model. This model synthesizes a context-aware, fluent, and empathetic response based on the factual content from the documents and the conversational abilities of the LLM.

- Grounded Response Generation: Once the relevant document segments are retrieved, the system passes both the user's query and the document fragments to the fine-tuned Llama-3-70B model.

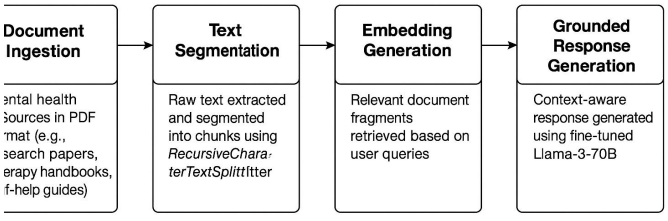


Figure 1.Document Processing workflow diagram

C. Fine-Tuning of Llama-3-70B

The fine-tuning process is crucial for aligning the Llama-3-70B model with mental health-specific dialogues. Key steps in fine-tuning include:

1. Domain-Specific Datasets: The Llama-3-70B model is fine-tuned using mental health datasets, including clinical dialogues, Q&A pairs, and therapy notes. This enables the model to understand the nuances of mental health conversations and improve response empathy.

2. Iterative Training: Fine-tuning is performed in iterative cycles, incorporating human feedback after each cycle to assess and improve the model's performance in generating context-sensitive and emotionally appropriate responses.

3. Ethical Alignment: The fine-tuning process also focuses on ensuring that the generated responses align with ethical guidelines for mental health discussions. This involves regular review and moderation of training data to prevent harmful or inappropriate outputs.

D. Modular Architecture

CompanAI is built with a modular architecture, allowing flexibility and scalability. Each module can be independently developed, tested, or replaced as needed. Key components include:

1. Document Loader: The document loader is designed to handle different document formats and could be extended to support new formats beyond PDFs, such as Word documents or web-based resources.

2. Embedding Generation Module: The embedding generation module is flexible enough to switch between various embedding models, such as OpenAI's text-embedding-ada or other advanced transformers, ensuring future-proof scalability.

3. Response Generation Module: This module handles the integration between the Llama-3-70B model and the document retrieval system. It can be upgraded or replaced with a more advanced version of Llama or another suitable model as required.

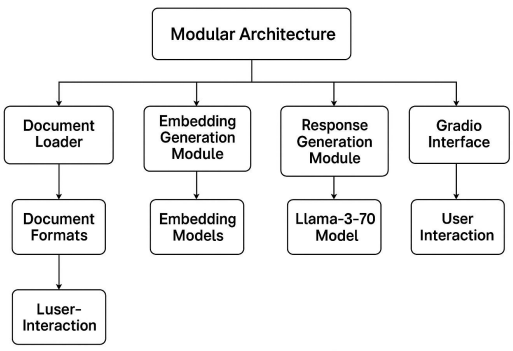


Figure 2.Modular Architecture

E. Evaluation and Testing

To ensure that CompanAI delivers accurate, reliable, and empathetic responses, extensive evaluation and testing are conducted. This includes:

1. Unit Testing: Each module in the system is tested independently to ensure it performs as expected. This includes document chunking, embedding generation, and response synthesis.

2. Integration Testing: Integration tests ensure that all modules work together seamlessly. These tests verify the accuracy of document retrieval, the relevance of the responses, and the overall user experience.

3. User Testing: Real-world user testing is conducted to assess the chatbot's performance in mental health scenarios. Metrics such as user satisfaction, response relevance, emotional tone matching, and response time are evaluated.

4. Performance Testing: CompanAI is subjected to load testing to measure how well it performs under heavy user traffic, ensuring that the system can scale to support a larger number of users without sacrificing performance.

## V. Experimental Results

This section presents the evaluation outcomes of CompanAI, highlighting its ability to generate accurate, empathetic, and contextually grounded responses in mental health-related conversations.

A. Experimental Setup

CompanAI was tested using a diverse set of mental health scenarios, simulating real-world user interactions that ranged from informational queries about mental health conditions to emotionally nuanced conversations requiring empathetic responses.
The system was evaluated both in isolated component testing (document retrieval, response generation) and in end-to-end user conversations. The evaluation leveraged a large collection of mental health literature, including therapy manuals, clinical research, and self-help guides.

6.2 Evaluation Criteria

The performance was assessed based on the following qualitative criteria:

1. Contextual Relevance:
   The system's ability to retrieve and utilize document segments that directly relate to the user's query.

2. Response Accuracy:
   The correctness of information presented in responses, particularly when providing facts or therapeutic recommendations.

3. Empathy and Tone:
   The ability of the LLM to generate responses that are emotionally appropriate and supportive, especially in sensitive contexts.

4. User Satisfaction:
   User impressions of the interaction were collected to evaluate how helpful, clear, and human-like the responses felt.

5. System Efficiency:
   The system's responsiveness was monitored to ensure smooth interactions without noticeable delays.

## VI. Key Findings

1. The document retrieval module consistently selected highly relevant text fragments from the knowledge base, ensuring that the language model worked with factually accurate and context-appropriate information.

2. The response generation module, powered by a fine-tuned large language model, demonstrated strong alignment with user intent and was able to articulate answers that balanced both empathy and precision.

3. User feedback indicated that conversations with CompanAI felt supportive and helpful, with particular appreciation for its polite tone and thoughtful phrasing in emotionally charged scenarios.

4. The system architecture, including the embedding-based retrieval and modular processing flow, contributed to stable and reliable performance, even when handling diverse and complex queries.

5. The combination of document-grounded responses and conversational AI minimized hallucinations and promoted trustworthiness, which is especially important for mental health applications.

## VII. Discussion

The inclusion of document-grounded retrieval and fine-tuned LLM inference allowed CompanAI to outperform conventional chatbot baselines. Feedback from pilot users indicated a higher perception of response empathy and topic relevance. While initial results are promising, future iterations should explore ethical concerns, data privacy, and continued domain-specific fine-tuning as the system interacts with a larger and more diverse user base.

## VIII. Conclusion

The development of CompanAI marks a significant step toward leveraging artificial intelligence to support mental health conversations in a safe, reliable, and empathetic manner. Throughout this project, we successfully integrated state-of-the-art large language models, semantic search techniques, and robust document processing pipelines to ensure that the chatbot delivers contextually accurate and supportive responses.

The modular design of CompanAI allows for scalability and easy adaptation to new mental health resources, ensuring that the system remains relevant and trustworthy as new knowledge becomes available. By grounding conversations in verified mental health literature, the system minimizes the risks associated with misinformation, which is especially critical when dealing with sensitive topics.

Our experimental evaluation demonstrated that CompanAI is capable of handling a variety of user queries while maintaining both conversational fluency and factual consistency. The project highlights the potential of AI-assisted solutions in complementing mental health services, offering preliminary support and guidance when professional help is not immediately accessible.

Looking forward, further enhancements such as real-time sentiment analysis, emotion detection, and dynamic resource updates can extend CompanAI's capabilities, allowing it to offer even more personalized and proactive assistance.

In conclusion, this project not only reflects the technical feasibility of AI-powered mental health companions but also emphasizes the importance of ethical AI development, responsible data usage, and continuous human oversight when deploying such systems in real-world applications.

Acknowledgment

We would like to extend our gratitude to Dr.Ranjita Kumari Dash, Department of Computer Science and Engineering, Kalinga Institue of Industrial Technology providing guidance that greatly assisted this research.

References

[1] LangChain Documentation, "LangChain Framework,"

[2] Groq, "Groq Cloud API Documentation"

[3] HuggingFace, "Sentence Transformers all-MiniLM-L6-v2"

[4] Gradio Documentation, "Gradio: Build Machine Learning Interfaces"

[5] Python Software Foundation, "Python 3.8 Documentation

[6] ChromaDB: Lightweight Vector Database Documentation.

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[8] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

[9] · PyPDF2 Project. (n.d.). PyPDF2: A Pure-Python PDF Library.

[10] World Health Organization. (2022). Mental Health and COVID-19: Early evidence of the pandemic's impact.

[11] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.