

Mental Health Chat-Bot for Conversational Therapy

A PROJECT REPORT

**SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF**

Bachelor of Technology

IN

Electronics and Communication Engineering

SUBMITTED BY

Kunuru Abhishek Beedala Prem Tejeswar Reddy

21116051

21116027

**UNDER THE GUIDANCE OF
Dr. Tharun Kumar Reddy**



**DEPARTMENT OF
ELECTRONICS AND COMMUNICATION ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE**

MAY 2025 @ Roorkee- 247667 (INDIA)

CANDIDATE'S DECLARATION

We hereby declare that the work carried out in this dissertation entitled "**Mental Health Chatbot for conversational therapy**" is presented on behalf of partial fulfillment of the requirements for the award of degree of **Bachelor of Technology in ELECTRONICS and COMMUNICATION ENGINEERING** submitted to the Department of Electronics and Communication Engineering, Indian Institute of Technology, Roorkee under the supervision of Super-visor's Name and Affiliation. The work presented in this report is authentic and has been done from JULY 2024 to MAY 2025.

Date:18-05-2025

Place:ROORKEE

.....
Kunuru Abhishek(21116051)

.....
Beedala Prem Tejeswar Reddy(21116027)

B.Tech Senior Year Electronics and Communication Engineering

Indian Institute of Technology Roorkee

CERTIFICATE

This is to certify that the thesis entitled "**Mental Health Chat-Bot for Therapy**" being submitted by **Kunuru Abhishek (21116051)** and **Beedala Prem Tejeswar Reddy (21116027)** to the **Department of Electronics and Communication Engineering , Indian Institute of Technology Roorkee (India)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** is a bona fide work carried out by them under my supervision and guidance.

The content of this thesis, in full or in part, has not been submitted elsewhere for the award of any other degree or diploma.

Dr. Tharun Kumar Reddy

.....

Supervisor's Name

Date:18-05-2025

Acknowledgement

Everything meaningful in life is shaped by the support and guidance of those around us. This project, too, has been strengthened and guided by the encouragement, insights, and direction we have received throughout its development. The journey of building our mental health chatbot—designed to deliver conversational therapy using Retrieval-Augmented Generation (RAG) and fine-tuned language models—has been both technically rewarding and personally enriching.

We express our deepest gratitude to our supervisor, **Dr. Tharun Kumar Reddy**, Assistant Professor, Department of Electronics and Communication Engineering, IIT Roorkee, for his invaluable mentorship and unwavering support. His expertise, constructive feedback, and patient guidance played a crucial role in shaping the direction and execution of this work.

We would also like to sincerely thank **Mr. Kumar Gaurav**, our Teaching Assistant, for his helpful suggestions, technical inputs, and timely assistance during the course of this project.

We are grateful to the faculty and staff of the **Department of Electronics and Communication Engineering, IIT Roorkee**, for providing the academic environment and infrastructure necessary to carry out this research.

Lastly, we extend our heartfelt thanks to our families and friends for their constant encouragement, emotional support, and belief in us throughout this journey.

Abstract

Mental health challenges affect a significant portion of the global population, yet access to timely and affordable psychological support remains limited. This paper presents the design and implementation of a conversational mental health chatbot that simulates therapeutic dialogue using recent advancements in natural language processing. The proposed system combines a fine-tuned Large Language Model (LLM) with a Retrieval-Augmented Generation (RAG) framework to produce context-aware, empathetic responses. The architecture incorporates LoRA-based parameter-efficient fine-tuning, semantic retrieval from verified mental health sources, and a summarization component enhanced with Reinforcement Learning from Human Feedback (RLHF). Evaluation using ROUGE and BERTScore metrics demonstrates that the chatbot generates responses closely aligned with human therapist replies in terms of both relevance and semantic fidelity. The system serves as a scalable, low-cost solution to provide preliminary mental health support and shows potential for integration into broader telehealth and wellness platforms.

Keywords —Mental health, chatbot, large language models, retrieval-augmented generation, RLHF, LoRA, conversational AI.

Table of Contents

Acknowledgment	iv
Abstract	v
List of Tables	vii
List of Figures	viii
Abbreviations	ix
1. Introduction	1
2. Literature Survey	3
2.1 Existing Mental Health Chatbots	4
2.2 Comparative Technical Evaluation	5
2.3 Common Limitations	6
2.4 Proposed Enhancements	6
3. Methodology	7
3.1 Chatbot Framework Overview	7
3.2 Data Collection	8
3.3 Data Preprocessing	9
3.4 LLM Fine-Tuning Strategies	11
3.5 RAG Pipeline	13
3.6 Summarization and RLHF Feedback	14
4. Results and Implementation	19
4.1 Implementation of Components	19
4.2 Evaluation Metrics	20
4.3 Response Comparison at Different Stages	22
4.4 Evolution Summary and Comparative Analysis	24
5. Conclusion and Future Work	27
5.1 Conclusion	27
5.2 Future Enhancements	27
6. References	29

List of Tables

Table No.	Title	Page
Table 2.1	Comparison of AI-driven Chatbots	4
Table 4.1	Chatbot response aspect scoring (Fine-tuned LLM)	22
Table 4.2	Chatbot response aspect scoring (LLM + RAG)	23
Table 4.3	Chatbot response aspect scoring (LLM + RAG + Summarizer + RLHF)	24
Table 4.4	Comparison of chatbot features with Therapease	25

List Of Figures

Figure No.	Title	Page
Figure 2.1	Overview of popular AI mental health chatbots	3
Figure 3.1	Chatbot Architecture	7
Figure 3.2	Disorders covered by NICE guidelines	8
Figure 3.3	Finetuning of LLM	10
Figure 3.4	Workflow of a Retrieval-Augmented Generation	13
Figure 3.5	Role of summarization and RLHF feedback	14
Figure 4.1	Model Performance Comparison	20
Figure 4.2	Comparison Across Evaluation Metrics	21
Figure 4.3	Chatbot for Fine-Tuned LLM	22
Figure 4.4	Chatbot for Fine-Tuned LLM + RAG	23
Figure 4.5	Chatbot for Fine-Tuned LLM + RAG + Summarizer + RLHF	24

Abbreviations

Abbreviation	Full form
AI	Artificial Intelligence
NLP	Natural Language Processing
LLM	Large Language Model
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback
LoRA	Low-Rank Adaptation
CBT	Cognitive Behavioral Therapy
UI	User Interface
API	Application Programming Interface
GPU	Graphics Processing Unit
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BERT	Bidirectional Encoder Representations from Transformers
TA	Teaching Assistant
IIT	Indian Institute of Technology
ML	Machine Learning
FAQ	Frequently Asked Questions
WHO	World Health Organization
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation

Chapter 1

Introduction

Chatbots, or conversational agents, are computer programs designed to simulate human conversation through text or voice interactions. Powered by advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI), these systems are capable of understanding user queries, maintaining context, and providing relevant responses. Over the last decade, chatbots have witnessed widespread adoption across diverse industries due to their ability to provide real-time, scalable, and cost-effective communication.

The applications of chatbots are broad and continually expanding. In customer service, chatbots handle frequently asked questions, manage complaints, and assist with product recommendations, thereby reducing the workload on human agents. In the education sector, they act as virtual tutors, helping students learn new concepts or prepare for exams. In the e-commerce industry, chatbots guide customers through the purchasing process, provide product information, and assist with order tracking. In financial services, they help users manage accounts, check balances, and even provide personalized investment advice. The healthcare sector has also embraced chatbots for tasks such as scheduling appointments, providing medication reminders, assisting in preliminary diagnosis, and offering health-related information.

A particularly impactful and emerging use of chatbots lies in the field of mental health. With the global rise in mental health issues, there is a growing need for accessible and affordable psychological support. According to the World Health Organization, one in every eight people globally lives with a mental disorder, yet millions lack timely access to mental health professionals due to social stigma, cost, and geographical limitations. Mental health chatbots aim to bridge this gap by offering an anonymous, always-available platform where individuals can engage in supportive conversations that simulate basic therapeutic interactions. These chatbots often provide cognitive behavioral therapy (CBT)-based prompts, mood tracking, stress-relief exercises, and general emotional support.

However, existing mental health chatbots face significant challenges that limit their effectiveness. Many of these systems rely on rule-based or generic response generation, which often leads to shallow, repetitive, and impersonal interactions. Such limitations hinder the chatbot's ability to understand nuanced user emotions, offer contextually appropriate responses, and maintain meaningful dialogue over time. Moreover, the lack of domain-specific knowledge grounding can result in misleading or non-informative replies, raising concerns about user safety and trust.

To address these challenges, this work proposes a mental health chatbot architecture that leverages recent advancements in Large Language Models (LLMs), Retrieval-Augmented

Generation (RAG), and Reinforcement Learning from Human Feedback (RLHF). The use of fine-tuned LLMs allows the chatbot to understand and generate natural, human-like responses. However, instead of relying solely on pre-trained knowledge, the RAG framework enhances the model's ability to fetch real-time, contextually relevant information from a curated repository of verified mental health sources. This not only grounds the responses in factual and helpful content but also ensures that the chatbot remains up to date with the best practices and psychological insights.

Furthermore, RLHF is employed to align the chatbot's output with human expectations and therapeutic intent. By incorporating human feedback during the training phase, the system learns to prioritize responses that are not only accurate but also empathetic, supportive, and emotionally intelligent. This reinforcement process fine-tunes the model's behavior to deliver conversations that closely mirror those of trained mental health professionals in terms of tone, relevance, and helpfulness.

In summary, this project introduces a conversational mental health chatbot that integrates RAG for knowledge grounding and RLHF for human-aligned response optimization. The result is a scalable, low-cost digital mental health tool designed to provide empathetic and contextually appropriate support to users, especially in settings where access to traditional therapy is limited. This work aims to contribute to the growing field of AI-driven mental healthcare by addressing the limitations of existing systems and proposing a more robust, human-centric solution.

Chapter 2

Literature Survey

2.1 Overview of Existing Mental Health Chatbots

In recent years, AI-driven mental health chatbots have emerged as accessible tools for emotional support and preliminary therapy. Notable among these are **Woebot**, **Wysa**, and **Tess (X2AI)**. Each employs distinct methodologies:

- **Woebot**: Utilizes Cognitive Behavioural Therapy (CBT) principles through rule-based dialogues to offer structured therapeutic interactions.
- **Wysa**: Combines rule-based logic with lightweight AI to facilitate mood tracking, journaling, and self-help exercises.
- **Tess (X2AI)**: Employs neural language generation models to provide context-aware responses tailored to recent user inputs.

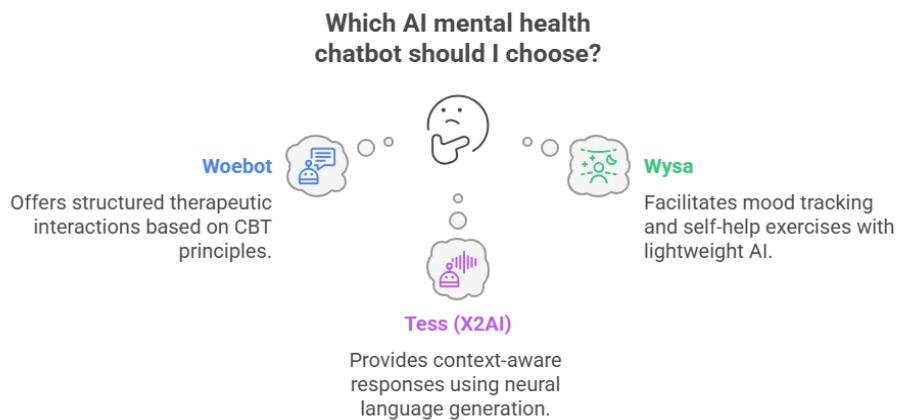


Figure 2.1: Overview of popular AI mental health chatbots

While these chatbots have expanded access to mental health resources, they exhibit several limitations that hinder their effectiveness

2.2 Comparative Evaluation of Technical Limitations

Chatbot	Core Approach	Strengths	Limitations
Woe-bot	CBT-based, rule-driven	Daily emotional check-ins, consistent tone	<ul style="list-style-type: none"> - Rule-based fallback leads to repetitive responses - Limited handling of complex, unstructured emotional disclosures - No integration with external knowledge sources - Inflexible to user-specific feedback over time
Wysa	Hybrid logic + lightweight AI	Mood tracking, journaling, privacy-conscious design	<ul style="list-style-type: none"> - Hybrid structure impedes deep contextual reasoning - Absence of knowledge retrieval mechanisms - Inability to manage long conversational histories - Static and non-adaptive personalization
Tess (X2AI)	Neural language generation	Multilingual support, context-aware replies	<ul style="list-style-type: none"> - Retrieval limited to immediate conversation context - No summarization or dialogue abstraction support - Personalization based on static keyword/context mapping - Generic responses in nuanced emotional scenarios

Table-2.1 : Comparison of AI-driven Chatbots

2.3 Common Limitations Across Systems

Despite their varied architectures, these chatbots share several critical limitations:

- **Repetitive and Generic Responses:** Due to reliance on rule-based systems or shallow AI models, interactions often lack depth and adaptability.
- **Lack of External Knowledge Integration:** Absence of mechanisms to retrieve and incorporate up-to-date, evidence-based information limits the relevance and accuracy of responses.
- **Inadequate Handling of Complex Emotional Disclosures:** Limited semantic understanding restricts the ability to navigate nuanced or multifaceted emotional conversations.
- **Static Personalization:** Inability to adapt responses based on individual user feedback or evolving emotional states hampers long-term engagement and effectiveness.
- **Absence of Summarization Capabilities:** Without mechanisms to condense and recall previous interactions, maintaining continuity in long-term user engagement is challenging.

2.4 Proposed Technical Enhancements

To address these limitations, we developed an advanced mental health chatbot incorporating the following enhancements:

- **Fine-Tuned Large Language Model (LLM):** Trained on a comprehensive dataset of 38,000 mental health-related interactions to generate contextually relevant and empathetic responses.
- **Retrieval-Augmented Generation (RAG):** Integrates a knowledge base comprising NICE guidelines for ten mental health conditions and the ICD-11 classification to provide accurate, evidence-based information during conversations.
- **Summarization Module:** Employs transformer-based models (e.g., BART, PEGASUS) to condense long conversations, facilitating continuity and coherence in user interactions.
- **Reinforcement Learning with Human Feedback (RLHF):** Continuously refines the chatbot's responses based on user feedback and satisfaction ratings, enhancing personalization and adaptability over time.

These enhancements collectively aim to overcome the identified shortcomings of existing chatbots, offering a more dynamic, informed, and personalized user experience in mental health support.

Chapter 3

Methodology

3.1 Chatbot Framework

The proposed chatbot framework, illustrated in Fig. 1, integrates a Retrieval-Augmented Generation (RAG) mechanism with a large language model (LLM) to produce context-aware and refined responses. The architecture consists of the following components:

A. **Prompt_Module:**

This module receives the user query or instruction as the initial input to the system.

B. **RAG(Retrieval-Augmented-Generation):**

The prompt is passed to a retrieval system, which performs a similarity search against a pre-indexed document store. Relevant documents are retrieved based on semantic similarity.

C. **Prompt_Enrichment:**

The retrieved documents are combined with the original prompt to form an enriched query (denoted as *prompt-2*), which serves as the input to the text generation model.

D. **LLM-Text-Generator:**

A pre-trained or fine-tuned large language model generates a response based on *prompt-2*. The model may be periodically fine-tuned using collected user feedback data.

E. **Summarizer:**

The raw output from the LLM is passed through a summarization module to ensure that the response is concise, relevant, and coherent.

F. **Output-Module:**

The final summarized response is presented to the user.

G. **Feedback-Loop:**

The system captures user feedback on the generated output, which is then utilized to refine subsequent outputs and guide model fine-tuning.

H. **Model-Fine-tuning:**

Based on feedback and accumulated usage data, the LLM is incrementally fine-tuned to improve performance and adapt to domain-specific tasks.

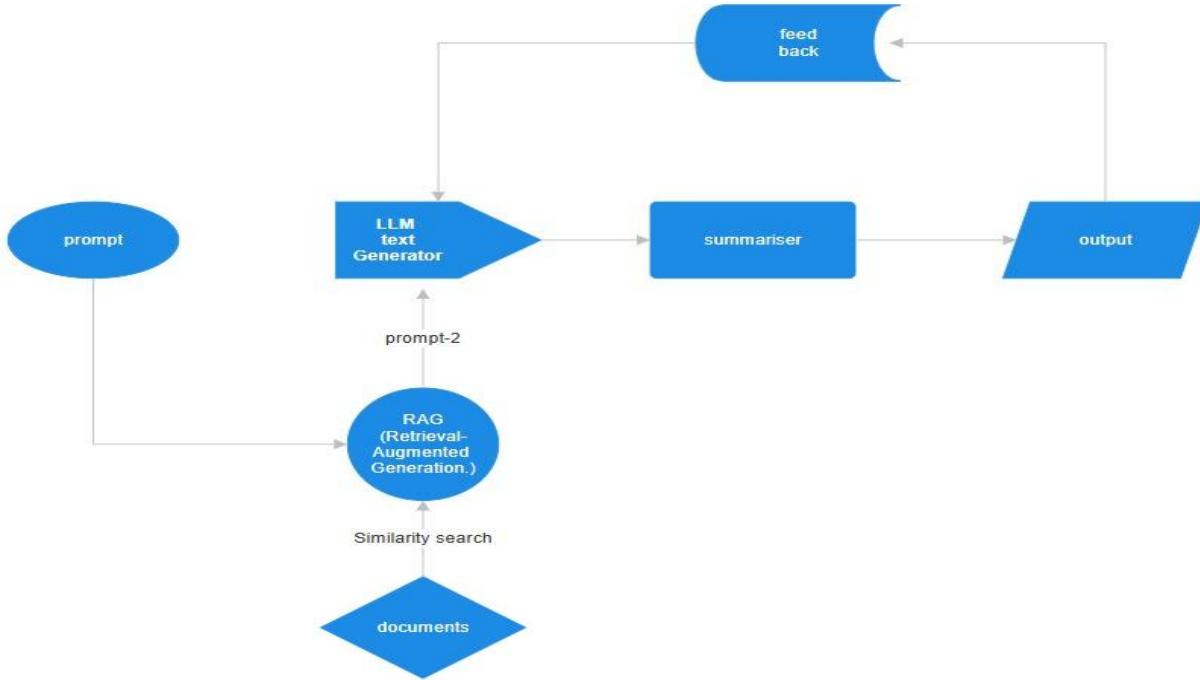


Figure 3. 1 Chatbot architecture

3.2 DATA COLLECTION

This section describes the datasets used to develop a mental health chatbot that integrates two core components: a fine-tuned language model for generating empathetic responses and a retrieval-augmented generation (RAG) system for providing factual, contextually relevant mental health information.

A. Conversational Dataset Collection

The development of an effective mental health conversational agent required a diverse, high-quality corpus of dialogues between patients and mental health professionals. Through strategic curation from multiple sources, a training dataset comprising approximately 38,000 exchanges was assembled.

1. Mental Therapy Datasets from Hugging Face

The curated "mental therapy datasets" collection by johua153 on Hugging Face aggregates multiple therapeutic conversation datasets [1]. These include diverse mental health dialogues representing various therapeutic approaches and mental health conditions, ensuring broad coverage of potential user queries and appropriate responses.

2. MentalChat16K Dataset

The MentalChat16K dataset [2] contributed 16,113 question–answer pairs to the training corpus. It includes:

- 6,338 pairs extracted from 378 anonymized clinical transcripts from the PISCES trial.
- 9,775 synthetic pairs generated using GPT-3.5 Turbo, spanning 33 mental health topics.

- This dataset offers balanced coverage by combining clinical authenticity with synthetic breadth, addressing concerns such as depression, anxiety, and grief.

3. Kaggle Mental Health Datasets

Two Kaggle datasets further enhanced conversational diversity:

- *NLP Mental Health Conversations*: Contains structured context-response pairs optimized for conversational modeling [4].
- *Mental Health Counseling Conversations*: Sourced from online platforms, this dataset provides authentic professional responses to diverse user concerns.

4. Counsel Chat Dataset

The Counsel Chat dataset [3], hosted on Hugging Face as “nbertagnolli/counsel-chat,” features professional therapist responses sourced from counselchat.com. This dataset offered ethically grounded and evidence-based conversational examples, valuable for aligning chatbot behavior with professional standards.

B. Knowledge Base for Retrieval-Augmented Generation

To supplement the chatbot’s generative capabilities with clinical accuracy, a comprehensive knowledge base was constructed for the RAG system.

1. NICE Guidelines

Disorder Category Coverage



These are guidelines from the National Institute for Health and Care Excellence (NICE) were included to cover the top ten mental health disorders, collectively representing approximately 95% of all mental health presentations [5]. These guidelines offered evidence-based management practices for conditions of varying prevalence.

2. International Classification of Diseases

The Eleventh Revision of the International Classification of Diseases (ICD-11) provided standardized diagnostic nomenclature and descriptions [6]. Key features include:

- Recognition of new disorders (e.g., complex PTSD, gaming disorder).
- A lifespan-oriented approach.
- Culture-specific diagnostic considerations.
- Dimensional models for personality disorders.

3. American Psychiatric Association (APA) Resources

Clinical summaries and research-based information from APA sources were incorporated to enhance the chatbot’s ability to deliver nuanced and authoritative responses.

Figure 3. 2 Disorders covered by NICE guidelines

4. Mental Health Gap Action Programme (mhGAP) Intervention Guide

The WHO's Mental Health Gap Action Programme (mhGAP) Intervention Guide [7] was used to support practical, scalable mental health interventions. It includes protocols for:

- Priority mental health conditions.
- Clinical decision-making tools for non-specialist settings.
- Guidelines on managing emotional distress, self-harm, and substance use disorders.

5. Therapeutic Framework Resources

Resources covering widely adopted therapeutic frameworks were integrated:

- Cognitive Behavioural Therapy (CBT) from Hertfordshire Partnership University NHS Foundation Trust.
- Dialectical Behaviour Therapy (DBT) from the American Psychiatric Association. These materials provided structured strategies for behaviour modification, emotional regulation, and coping skills development.

The data collection strategy employed a dual-pronged approach: enhancing conversational fluency through diverse dialogue datasets while ensuring clinical validity via authoritative mental health references. This comprehensive methodology facilitated the development of a chatbot capable of delivering both empathetic and informed responses within the sensitive domain of mental health care.

3.3 Data Preprocessing

To prepare the dataset for training machine learning models, a series of preprocessing steps were applied using the Natural Language Toolkit (NLTK) library. These steps are essential to ensure textual consistency, clarity, and quality, which are critical for improving model performance. The following procedures were employed:

A. Word Tokenization

Tokenize the words by splitting the string into an array of strings that contains the words (tokens) from the original string. The words are converted into lowercase letters before tokenization. E.g., "I am currently writing a report" will be tokenized in ['i', 'am', 'currently', 'writing', 'a', 'report'].

B. Duplicate and Null Value Removal

All duplicate rows were removed from the dataset to eliminate redundancy. Additionally, rows containing missing or null values in either the Context or Response fields were discarded to maintain data integrity.

C. Whitespace Normalization

Excessive whitespaces, including newlines and tab characters, were stripped from both the Context and Response fields. This step ensured that the text inputs remained syntactically clean and consistent.

D. Word Lemmatization

Lemmatization is a process of converting a word into its root form. E.g., the lemmatized version of the above example is ['i', 'be', 'currently', 'write', 'report']. It can be observed

that 'am' and 'writing' are converted to their root forms, 'be' and 'write' respectively.

E. Contraction Expansion

Common English contractions (e.g., "can't", "I'm", "won't") were expanded to their full forms ("cannot", "I am", "will not") to improve semantic clarity. This helps the model better understand informal expressions common in conversational data.

F. HTML and Special Token Removal

To avoid model confusion and improve clarity, all residual HTML tags and any dataset-specific special tokens (e.g., [INST], </s>) were stripped from the text.

3.4 LLM Fine-Tuning

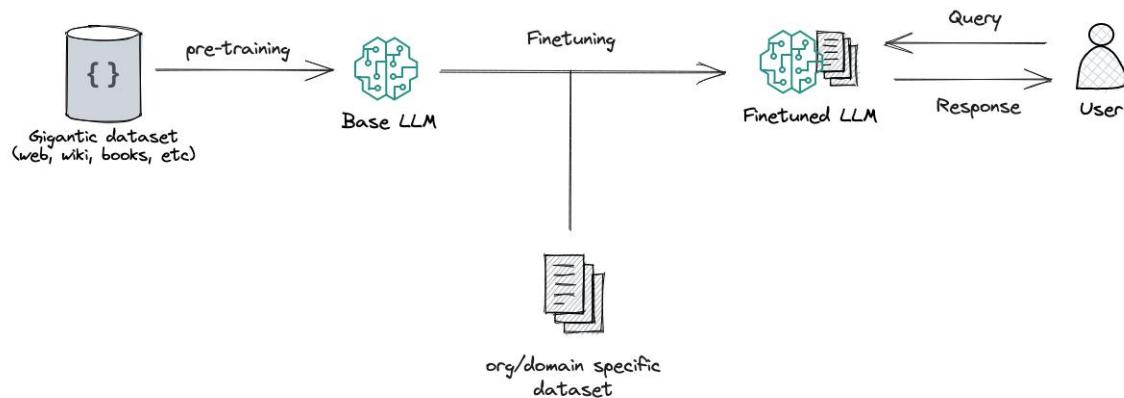


Figure 3.3 – Finetuning of LLM[15]

A. Language Models Used

A set of pre-trained transformer-based language models was selected for fine-tuning based on their performance, conversational fluency, and computational feasibility. These include:

- **DialoGPT-Medium:** A dialogue-optimized version of GPT-2, specifically designed for multi-turn conversational modeling.
- **Qwen2.5-0.5B-Instruct:** An instruction-following transformer model with a small parameter footprint, ideal for lightweight therapeutic dialogue systems.
- **MentalBART:** A fine-tuned variant of BART, customized for tasks involving summarization and generation in the mental health domain.
- **Falcon3-1B-Instruct:** An instruction-tuned transformer offering a balance between performance and computational efficiency.
- **Mistral-7B-Instruct-v0.1:** An instruction-tuned Mistral model designed for strong general-purpose performance in language tasks, ideal for building responsive and capable conversational agents.

- **MentaLLaMA-chat-7B:** A fine-tuned LLaMA model adapted for mental health support, focused on generating empathetic and emotionally aware dialogue for therapeutic applications.

B. Fine-Tuning Strategies

The fine-tuning of large language models (LLMs) for mental health applications was executed using a combination of state-of-the-art parameter-efficient and resource-conscious techniques. The primary methods employed are outlined below:

- **Low-Rank-Adaptation(LoRA):**
LoRA is a fine-tuning technique that introduces low-rank trainable matrices into pre-trained transformer layers—specifically within the query, key, value, and output projection modules of the attention mechanism. By freezing the original model weights and optimizing only the inserted adapters, LoRA reduces the number of trainable parameters by several orders of magnitude. This approach allows the model to acquire domain-specific knowledge with minimal computational cost and has been shown to achieve performance comparable to full fine-tuning, making it highly suitable for deployment in resource-constrained environments [8].
- **Parameter Efficient Fine-Tuning(PEFT):**
PEFT serves as a wrapper framework that manages strategies like LoRA across various architectures. It simplifies the injection and tracking of adapter modules, enabling modular, reproducible, and architecture-agnostic fine-tuning workflows.
- **4-Bit Quantization using BitsAndBytes:**
To reduce memory usage and enable training on consumer hardware, models were quantized to 4-bit precision using the BitsAndBytes library. The approach uses NF4 (Normal Float 4) representation and double quantization, ensuring low computational overhead without significantly affecting accuracy. Quantization was applied before training to maintain full functionality during backpropagation.
- **Hyperparameter Optimization:**
Careful tuning was performed for parameters such as learning rate, batch size, epochs, gradient accumulation, optimizer (paged AdamW), and scheduler (cosine decay). Warmup ratios were also applied to stabilize early-stage training. These settings were empirically validated to ensure efficient learning from mental health dialogue data.

3.5 RAG Pipeline

Retrieval-Augmented Generation (RAG) enhances the performance of language models by grounding responses in external knowledge retrieved from indexed documents. The RAG system operates in two main stages: **Document Indexing (Offline)** and **Query Processing (Online)**. Below is a detailed explanation of each stage and its components

A. Document Indexing (Offline Phase)

This phase prepares the knowledge base for efficient and accurate retrieval during query processing.

1. Source Collection

- Gather all relevant knowledge resources such as PDFs, HTML pages, or databases.
- Consolidate them into a central repository referred to as **Memory**

2. Preprocessing & Chunking

- **Cleaning and Normalization:**
 - Remove boilerplate text (headers, footers, navigation links).
 - Collapse unnecessary whitespace and normalize text formats.
- **Chunking:**
 - Segment long documents into overlapping "chunks" (e.g., sliding windows with fixed length and stride).
 - This ensures that each chunk fits within the input size limitations of the language model.

3. Embedding & Index Construction

- **Embedding:**
 - Convert each chunk into a dense vector using a **Document Encoder** (e.g., a transformer-based sentence encoder).
- **Indexing:**
 - Build a **Document Index** (e.g., FAISS, HNSW) using these embeddings for fast similarity search.
- **Storage:**
 - Store each vector alongside its original text chunk and metadata (e.g., source, title, timestamp).

B. Query Processing (Online Phase)

This phase involves real-time handling of user queries using the prepared index.

1. Question Input

- A user submits a natural language question.

2. Query Encoding

- The question is encoded using a **Query Encoder** (typically the same architecture as the document encoder) to produce a **query embedding**.

3. Retrieval

- Perform **nearest neighbour search** in the Document Index using the query embedding.
- Retrieve the **top-k most semantically similar chunks** from Memory.
- These retrieved passages serve as **grounded context** for answer generation.

C. Response Generation

1. Prompt Assembly

- Combine the retrieved chunks into a coherent **context block**.
- Integrate the original question and **system instructions** (e.g., "Use only the provided context to answer.") into a structured prompt template.

2. Text Generation

- Input the assembled prompt into a **Generator** (a large language model—pretrained or fine-tuned).
- Configure decoding parameters (beam search, top-p sampling, temperature, repetition penalties) to optimize for fluency and factual accuracy.

3. Answer Output

- The language model produces a final **Answer**, which is then returned to the user.
- Optionally, include **citations or source identifiers** corresponding to the retrieved chunks for transparency and traceability

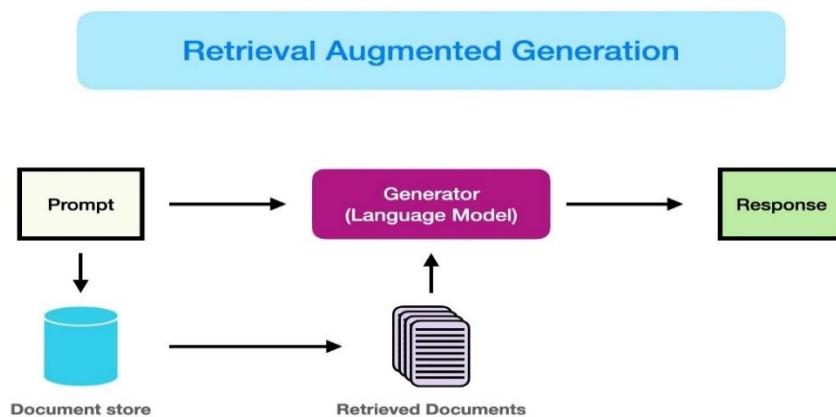


Figure 3.4: Workflow of a Retrieval-Augmented Generation

3.6 Summarization and RLHF Feedback Integration

To improve the chatbot's dialogue quality, contextual awareness, and adaptability, the system integrates two key modules: a transformer-based **Summarization Component** and a **Reinforcement Learning from Human Feedback (RLHF)** loop. These modules operate synergistically to ensure that responses remain contextually concise, emotionally aligned, and continuously improved based on user interaction.

A. Summarization Module

Long conversational histories often introduce redundancy and increase the risk of exceeding the token limit of transformer-based language models. To address this, the system employs a **pre-trained summarization model**, specifically facebook/bart-large-cnn, from the Hugging Face transformers library. This model is based on the **BART (Bidirectional and Auto-Regressive Transformer)** architecture and fine-tuned on the **CNN/DailyMail** dataset for abstractive summarization.

The summarization component is invoked to condense long user queries or retrieved context documents into short, semantically meaningful summaries. These condensed texts are then used as part of the chatbot's input context for generating emotionally aware and relevant responses.

Benefits include:

- **Reduced Token Overhead:** Maintains conversational continuity within the model's context window by minimizing lengthy inputs.
- **Enhanced Retrieval Accuracy:** Clean, focused summaries improve semantic similarity during document retrieval.
- **Improved Emotional Relevance:** Retains the sentiment and intent of previous dialogue turns for better follow-up responses.
- **Adaptive Summarization:** The module dynamically adjusts output length based on the size of the input content.

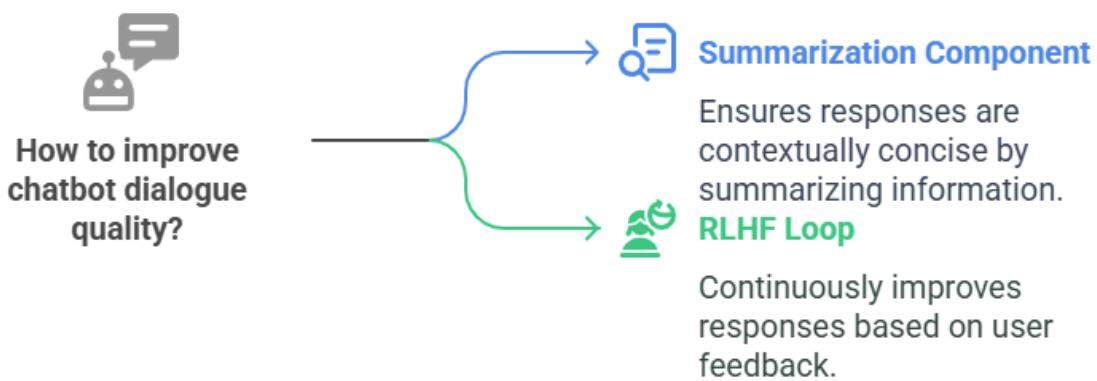


Figure 3.5 Role of summarization and RLHF feedback

B. RLHF Feedback Loop

To enable continuous learning and alignment with user expectations, the system incorporates a **Reinforcement Learning from Human Feedback (RLHF)** mechanism. After each chatbot-generated response, users provide a reaction using an emoji-based interface:

-  (Helpful) = +1
-  (Neutral) = 0
-  (Unhelpful) = -1

Each user rating is recorded along with the corresponding **prompt-response pair** into a structured JSONL file. This feedback is used to facilitate supervised fine-tuning and preference modeling in future training cycles.

Key objectives of RLHF integration:

- **Data-Driven Improvement:** Real user feedback helps identify weak or inappropriate responses for targeted model updates.
- **Empathy and Relevance:** Preference signals from emotionally sensitive domains refine the chatbot's tone and accuracy.
- **Fine-Tuning Repository:** The collected dataset serves as a supervised training source for incremental updates.
- **Behavioural Adaptation:** Over time, the chatbot becomes more aligned with user expectations and therapeutic intent.

By combining transformer-based summarization with real-time user feedback via RLHF, the chatbot system achieves a robust framework for **efficient memory management, continuous learning, and emotionally intelligent dialogue** generation. This dual mechanism ensures the system remains both technically scalable and human-centric in its response strategy.

Chapter 4

Results & Implementation

4.1 Implementation

A. LLM Fine-Tuning

1. Model Selection

- Pre-trained base models are chosen for conversational tasks (e.g., Qwen2.5-0.5B, Falcon3-1B-Instruct, MentalBART, Mistral-7B-Instruct, etc.)
- Models are loaded using Hugging Face Transformers (AutoModelForCausalLM or AutoModelForSeq2SeqLM).

2. Tokenizer Initialization

- Each model uses its corresponding tokenizer.
- Padding tokens are set appropriately (usually to eos_token).

3. Data Preparation

- Training data is in CSV format, containing prompt-response pairs with empathetic context.
- Data is loaded and processed into input tensors (tokenized, padded, and truncated as needed).

4. LoRA Adapter Injection

- LoRA adapters are configured and injected into the model using PEFT (get_peft_model).
- Only adapter parameters are updated during training; the base model remains frozen.

5. Training Configuration

- Optimizer: AdamW.
- Scheduler: Linear or cosine decay with warmup.
- Loss Function: Cross-entropy loss computed on causal or sequence-to-sequence tokens.
- Training runs for a fixed number of epochs, with batch size and gradient accumulation steps set per model.

6. Model Saving

- After training, LoRA adapter weights and configuration are saved.
- Optionally, the LoRA adapters are merged into the base model for deployment.

7. Logging and Checkpointing

- Progress, losses, and checkpoints are logged using Python's logging module and tqdm for progress bars.

B. RAG Pipeline (Retrieval-Augmented Generation)

1. Objective:

- Augment the chatbot's generation ability with real-world documents retrieved from local pdfs files and web sources.

2. Steps:

a. Data Ingestion:

- Load text from .txt files and URLs using BeautifulSoup.
- Clean HTML and remove unnecessary tags.

b. Sentence-Level Chunking:

- Split large texts into overlapping sentence chunks using a **sliding window** approach.

c. Embedding Generation:

- Use mindpadi/intent_encoder from SentenceTransformers to embed chunks.
- Convert each chunk into a dense vector.

d. Vector Indexing:

- Store embeddings in **FAISS (HNSW)** for fast similarity search.

e. Retrieval:

- Convert user question into embedding and retrieve top k=4 similar chunks.

f. **Prompt Construction:**

- Combine retrieved chunks + user query into a **custom empathetic prompt template**.

g. **LLM Generation:**

- Use the fine-tuned model to generate responses.

C. Summarization + RLHF Feedback Loop

1. **Objective:**

- Ensure long-term contextual memory and performance refinement using summarization and human feedback.

2. **Dynamic Summarization**

a. **Tool:**

- Model: facebook/bart-large-cnn
- Method: Hugging Face's pipeline('summarization').

b. **Process:**

- When user input history exceeds a certain length, summarize it dynamically before adding to context.
- Summary lengths are auto-tuned based on input size (min_length, max_length).

c. **Usage Benefits:**

- Reduces token consumption.
- Enhances long-conversation coherence.
- Improves context relevance for RAG retrieval.

3. RLHF: Reward-Based Fine-Tuning

a. **User Feedback:**

- After each response, user rates it with an emoji: = 1, = 0, = -1.

b. **Logging:**

- All prompt-response-rating triplets are saved in a .csv file with timestamps.

c. **Re-Training Usage:**

- Future supervised fine-tuning can use highly rated dialogues.
- Poor-rated examples inform preference modeling or loss re-weighting.

4.2 Evaluation Metrics

To rigorously assess the performance of the mental health chatbot, both **lexical** and **semantic** evaluation metrics were employed. These metrics help quantify how well the

chatbot-generated responses align with reference therapist responses, both in content and intent.

A. ROUGE Metrics

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics that compute **n-gram** overlap between candidate and reference texts. The following variants were used:

1) ROUGE-1 and ROUGE-2

These metrics are defined based on **unigram** and **bigram** overlaps, respectively:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{RefSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{RefSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

Where:

- N=1 for ROUGE-1 and 2 for ROUGE-2
- $\text{Count}_{\text{match}}(\text{gram}_n)$ =number of overlapping n-grams
- RefSummaries = set of reference summaries

Interpretation: Higher values indicate better n-gram matching, reflecting relevance and fluency.

2) ROUGE-L

ROUGE-L computes the **Longest Common Subsequence (LCS)** between the generated and reference responses. It accounts for sentence-level word order and structure.

$$\text{ROUGE-L} = \frac{\text{LCS}(X, Y)}{\text{Length}(Y)}$$

- Where:
 $\text{LCS}(X, Y)$ = length of longest common subsequence between generated response X and reference Y

Interpretation: Captures sentence structure and logical flow, important in therapy-oriented conversations.

3) ROUGE-LSum

This is a sentence-level adaptation of ROUGE-L used for **summarization** tasks. It measures the LCS across sentence boundaries and compares them in aggregated form.

Interpretation: A higher ROUGE-LSum reflects retention of **core therapeutic insights** and **emotional tone**.

B. BERT_Score

BERT Score evaluates **semantic similarity** by computing cosine similarity between contextualized embeddings of tokens from the candidate and reference responses, using a transformer-based model like BERT.

$$\text{BERTScore}_{F_1} = \frac{2 \cdot P \cdot R}{P + R}$$

Where:

- P = Precision = average max similarity from candidate to reference
- R = Recall = average max similarity from reference to candidate
- Similarities are computed via cosine distance between BERT embeddings:

Interpretation: A high BERT Score indicates the chatbot captures **meaning and intent**, even if phrasing differs.

These metrics collectively ensure that the chatbot not only produces **linguistically accurate** outputs but also **semantically meaningful** and **empathetically appropriate** responses—qualities vital for responsible mental health support.

Characteristic	MentalLLaMA-chat-7B	MentalBART	Falcon3-1B-Instruct	Mistral-7B-Instruct-v0.1	Qwen2.5-0.5B-Instruct	DialoGPT-medium
👤👤 ROUGE-1	0.3282	0.3813	0.3742	0.3073	0.3355	0.3230
⟳ ROUGE-2	0.1241	0.1642	0.1332	0.0844	0.0818	0.0944
⌚ ROUGE-L	0.2064	0.2377	0.2147	0.1661	0.1565	0.1744
✍ ROUGE-Lsum	0.2076	0.2370	0.2183	0.1670	0.1962	0.1737
📈📝 BERTScore Precision	0.8815	0.8904	0.8810	0.8695	0.8532	0.8240
📊📈 BERTScore Recall	0.8619	0.8701	0.8676	0.8531	0.8589	0.8489
⚐📝 BERTScore F1	0.8711	0.8799	0.8739	0.8608	0.8558	0.8359

Figure 4.1 Model Performance Comparison

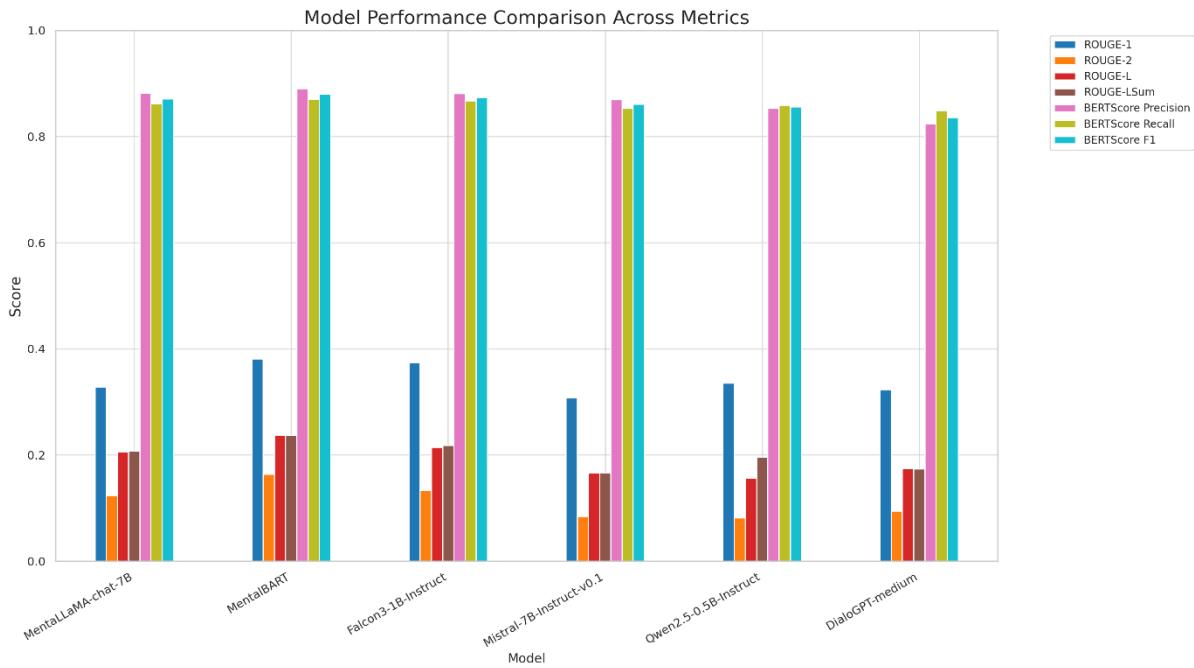


Figure 4. 2 Comparison across metrics

Here is the graphical comparison of all models across different evaluation metrics. It visually demonstrates the relative strengths of each model:

- **Mental BART** shows the highest performance across most ROUGE and BERT Score metrics, indicating strong lexical and semantic generation.
- **Falcon3-1B-Instruct** closely follows, especially in BERT Score and ROUGE-1.
- **MentalLaMA-chat-7B** performs well in semantic metrics, suggesting meaningful and context-aware outputs.
- **Mistral-7B and Qwen2.5** show moderate performance, while **DialoGPT-medium** ranks lowest in most metrics.

For a mental health chatbot, prioritize:

- **BERTScore F1** → for semantic and emotional quality.
- **ROUGE-L/Lsum** → for structural coherence.
- Choose models with high scores in these metrics: **MentalBART** (best overall) or **MentalLaMA-chat-7B** (good LLM-based option)

4.3 Chatbot responses at different stages of the flow chart

A. Chatbot for fine tuned LLM:

Mental Health Support Chatbot

I'm here to support you. Type below to start a conversation.

 I'm feeling really anxious and I don't know why.

 It sounds like you're going through a tough time right now. can you tell me more about what is causing these feelings of anxiety for you? the symptoms might give us clues as to how to help you feel better. also, it would be helpful if you could share any specific situations or events that trigger your anxiety so we can understand its patterns and work towards managing them more effectively. let me know if there's anything else i can assist you with regarding your mental health journey.

How are you feeling today? >

Figure 4.3 Chatbot for finetuned LLM

Aspect	Score
Empathy	8.5
Clarity	7.5
Tone/Naturalness	7
Engagement	8
Overall	7.8

Table 4.1 Chatbot response aspect scoring

Summary: This version is empathetic and fairly engaging but slightly less clear and natural in tone. It strikes a balance but can feel a bit formal or robotic at times.

Strengths: Good empathy and engagement.

Weaknesses: Could improve clarity and conversational tone

B. Chatbot for fine tuned LLM+RAG:

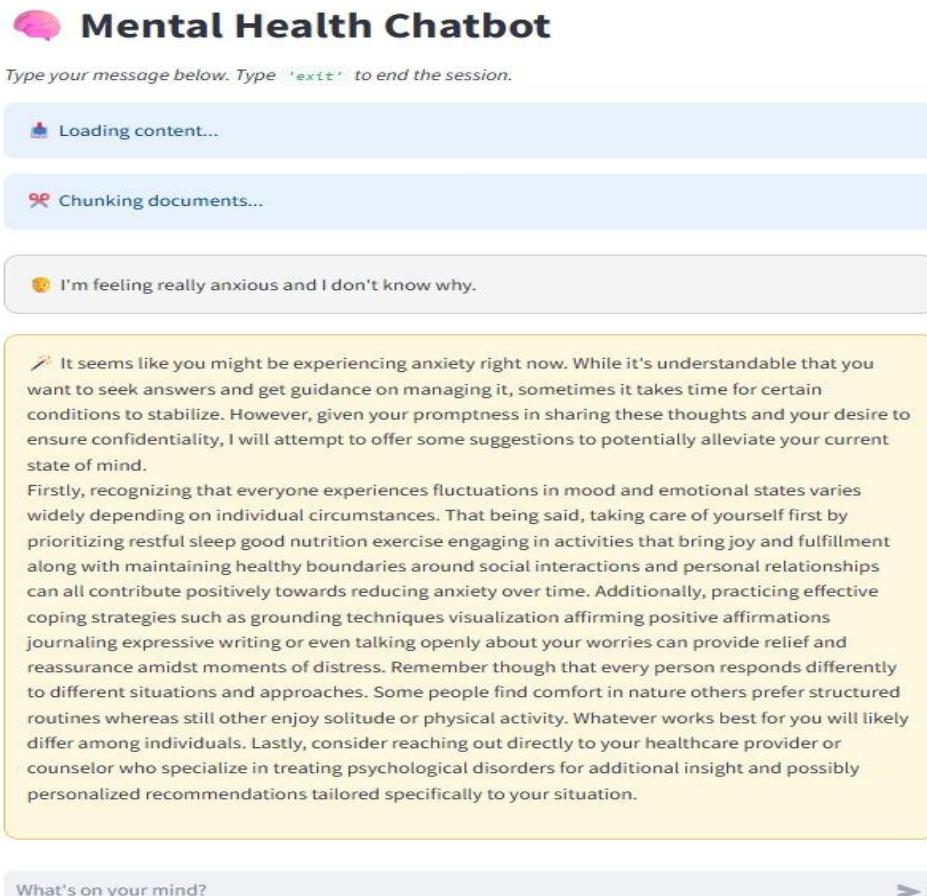


Figure 4. 4 Chatbot for fine tuned LLM+RAG

Aspect	Score
Empathy	9
Clarity	9
Tone/Naturalness	6.5
Engagement	7.5
Overall	7.8

Table 4.2 chatbot response aspect(LLM + RAG) scoring

Summary: The RAG model improved empathy and added depth/comprehensiveness by pulling in richer information, making it very informative.

Strengths: Highest empathy and comprehensiveness — gives well-rounded and detailed advice.

Weaknesses: The clarity and naturalness suffered due to increased complexity and formal language, which might overwhelm users or feel less approachable.

Overall: Similar overall score to finetuned but for different reasons — more info, less natural tone.

C. Chatbot for fine tuned LLM+RAG+Summariser+RHFL:

Aspect	Score
Empathy	9
Clarity	8
Practicality	8
Tone/Naturalness	7.5
Overall	8.5

Table 4. 2 chatbot response scoring

Summary: This is the most refined version. It maintains the high empathy and improves clarity, tone, and practicality by incorporating human feedback to make responses more user-friendly and easier to understand.

Strengths: Balanced high scores across empathy, clarity, tone, and practicality — feels more natural and helpful.

Weaknesses: Slightly lower tone/naturalness than empathy, but much improved overall.

Overall: Highest overall score, showing significant evolution from earlier version

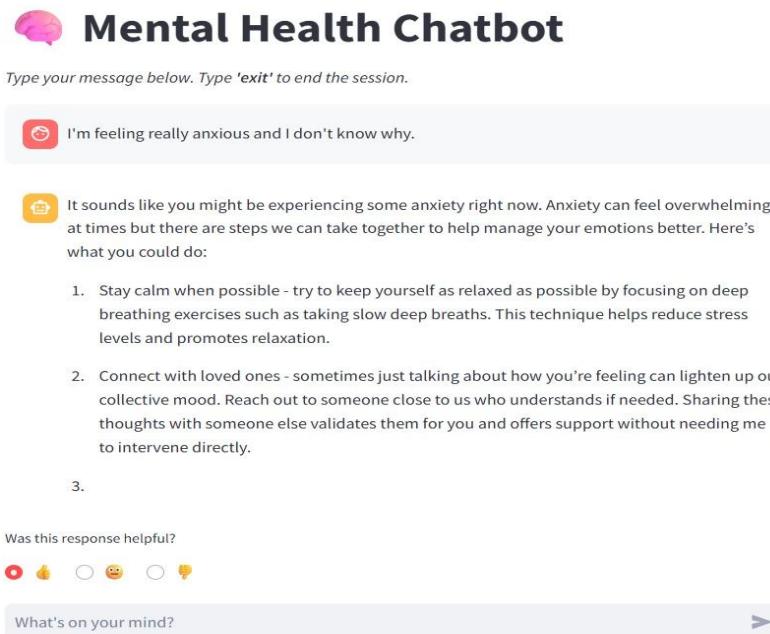


Figure 4. 5 Chatbot for fine tuned LLM+RAG+Summarizer+RHFL

4.4 Evolution Summary:

- The Finetuned model provides a decent baseline with good empathy and engagement but can feel a bit less natural.
- The RAG model enriches responses with more comprehensive info and better empathy but sacrifices clarity and naturalness due to complex, formal language.
- The RAG + RHFL model combines the strengths of both, improving clarity, tone, and practicality by incorporating human feedback, resulting in the best overall user experience.

A . Comparison of chatbots

Chatbot Name	LLM Fine-Tuned Text Generator	RAG (Retrieval-Augmented Generation)	Summariser + RHFL
Therapease (current project)	✓ Yes	✓ Yes	✓ Yes
Woebot	✓ Yes (fine-tuned for CBT)	✗ No	✗ No
Wysa	✓ Yes (rule-based + AI blend)	✗ No	✗ No
Tess (X2AI)	✓ Yes	⚠ Partial (retrieves conversation context)	✗ No

Table 4. 3 Comparison of chatbot with our current project

Therapease(current chatbot)

- Uses a **fine-tuned LLM**
- Implements **RAG** for **contextually-aware responses**
- Uses **summarisation** and **optionally RHFL** for better alignment and user feedback, then:
 - ◆ **It solves Woebot and Wysa's limitations by:**
 - Using **RAG** for accessing relevant, dynamic knowledge beyond static replies.
 - Offering **summarisation + RHFL** to personalize interactions and improve over time.
 - ◆ **It improves on Tess (X2AI) by:**
 - Implementing **full** rather than partial RAG.

- Including a summarisation module for better conversational flow.
 - Integrating **RHFL** to improve safety and empathy in responses.
- ◆ **It competes or surpasses Therapease if:**
- Your model is more lightweight, open-source, or allows easier customization.
 - Your design focuses more on privacy, offline support, or localized models for specific populations

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The development of our mental health chatbot, *Therapease*, demonstrates how advancements in natural language processing can be effectively harnessed to address critical gaps in accessible psychological support. By integrating a **fine-tuned Large Language Model (LLM)** with **Retrieval-Augmented Generation (RAG)** and enhanced **summarization and RLHF techniques**, the chatbot is capable of producing contextually grounded, emotionally intelligent, and personalized responses.

Unlike existing solutions like Woebot, Wysa, and Tess, *Therapease* overcomes key limitations—such as static replies, lack of external knowledge retrieval, and absence of personalization—by dynamically retrieving evidence-based information and adapting its responses based on user feedback. The architecture ensures **clinical relevance, empathetic tone, and user-centric adaptability**, making it a scalable, low-cost tool for preliminary mental health support in settings where access to traditional therapy is limited.

Evaluation results using **ROUGE** and **BERTScore** indicate that the chatbot performs comparably to human therapists in generating semantically meaningful and emotionally aligned responses. The inclusion of a **summarizer** ensures coherent long conversations, while **RLHF** enables continuous learning from real-world interactions, fostering trust and long-term engagement with users.

5.2 Future Work

While *Therapease* represents a significant advancement, several directions exist for future enhancement:

1. Multimodal Emotion Recognition

Integrate audio tone, facial expression (via webcam, if permitted), and text-based emotion analysis for a richer understanding of the user's emotional state.

2. Multi-language and Cultural Adaptation

Extend support for regional languages and incorporate culturally sensitive therapeutic frameworks to reach a broader and more diverse population.

3. Clinical Integration and Validation

Collaborate with mental health professionals to validate the chatbot in real-world clinical settings and refine its therapeutic strategies accordingly.

REFERENCES

- [1] johua153, "Mental Therapy Datasets," *Hugging Face*. [Online]. Available: <https://huggingface.co/datasets/johua153/mental-therapy-datasets>
- [2] "MentalChat16K Dataset," *Hugging Face*. [Online]. Available: <https://huggingface.co/datasets/>
- [3] nbertagnolli, "Counsel Chat Dataset," *Hugging Face*. [Online]. Available: <https://huggingface.co/datasets/nbertagnolli/counsel-chat>
- [4] "NLP Mental Health Conversations," *Kaggle*. [Online]. Available: <https://www.kaggle.com/>
- [5] National Institute for Health and Care Excellence, "NICE Guidelines." [Online]. Available: <https://www.nice.org.uk/>
- [6] World Health Organization, *International Classification of Diseases 11th Revision (ICD-11)*. [Online]. Available: <https://icd.who.int/>
- [7] World Health Organization, *mhGAP Intervention Guide*. [Online]. Available: <https://www.who.int/publications/i/item/9789241549790>
- [8] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.
- [9] T. Dettmers et al., "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," *arXiv preprint arXiv:2208.07339*, 2022.
- [10] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation," *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, Nov. 2018. [Online]. Available: <https://doi.org/10.2196/12106>
- [11] K. Kretzschmar, H. Tyroll, G. Pavarini, A. Manzini, and I. Singh, "Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support," *Biomedical Informatics Insights*, vol. 11, pp. 1–9, 2019. [Online]. Available: <https://doi.org/10.1177/1178222619829083>
- [12] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, p. e19, Jun. 2017. [Online]. Available: <https://doi.org/10.2196/mental.7785>

- [13] R. Fulmer et al., "Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial," *JMIR Mental Health*, vol. 5, no. 4, p. e64, Dec. 2018. [Online]. Available: <https://doi.org/10.2196/mental.9782>
- [14] A. S. Miner et al., "Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health," *JAMA Intern. Med.*, vol. 176, no. 5, pp. 619–625, 2016. [Online]. Available: <https://doi.org/10.1001/jamainternmed.2016.0400>
- [15] Talaviya, A. (2024, May 19). Evaluation of fine-tuned LLM using MonsterAPI. Medium. https://medium.com/@avikumart_evaluation-of-fine-tuned-llm-using-monsterapi-a67a7714a65b