Machine Minds: Behavioral Assessment of LLM Therapists

Submitted in partial fulfillment of the requirements
for the degree of

B.E. Computer Engineering

By

Jheel Shah	**66**	**212101**
Shashank Kamble	**68**	**212045**
Anushree Nahak	**70**	**212067**
Harsh Kalgutkar	**72**	**212044**

Guide
Ms. K. Priya Karunakaran
Assistant Professor

Department of Computer Engineering
St. Francis Institute of Technology
(Engineering College)
An Autonomous Institute, Affiliated to University of Mumbai

2024-2025

CERTIFICATE

This is to certify that the project entitled **"Machine Minds: Behavioral Assessment of LLM Therapists"** is a bonafide work of **"Jheel Shah (66), Shashank Kamble (68), Anushree Nahak (70) and Harsh Kalgutkar (72)"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering.

(Ms. K. Priya Karunakaran)

Guide

(Dr. Kavita Sonawane)

Head Of Department

(Dr. Sincy George)

Principal

Project Approval Report for B.E.

This project report entitled "***Machine Minds: Behavioral Assessment of LLM Therapists***" by ***Jheel Shah (66), Shashank Kamble (68), Anushree Nahak (70) and Harsh Kalgutkar (72)*** is approved for the degree of ***B.E. in Computer Engineering***.

Examiners

1. 30/4/25 [K. Priya K]

2. 30/4/25 Dr. Suvarna P.

Date: 30/4/25

Place: Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Jheel Shah 66

Shashank Kamble 68

ANahak

Anushree Nahak 70

Harsh Kalgutkar 72

Date: 30/4/25

Abstract

The growing interest in leveraging Artificial Intelligence (AI) for mental health care has led to the exploration of Large Language Models (LLMs) such as GPT-2, LLAMA, and other conversational AI models as potential therapeutic tools. This study, titled "Machine Minds: Behavioral Assessment of LLM Therapists," evaluates the effectiveness and limitations of these advanced models in providing therapeutic support for mental health care. Despite the promise LLMs hold in enhancing accessibility to mental health services, their reliability in dealing with emotionally complex, sensitive, and crisis situations remains largely untested.

Through a structured evaluation framework, we assess key therapeutic parameters, including empathy, emotional sensitivity, ethical response handling, and consistency of LLMs in simulated therapeutic interactions. By analyzing interactions between AI models and simulated patients, this study highlights the critical performance gaps of LLMs, particularly in the areas of contextual understanding, handling emotional depth, and addressing ethical dilemmas.

The research aims to provide a comprehensive analysis of the potential and current limitations of LLMs in therapeutic settings. The findings will inform future advancements and improvements to ensure that these models can safely and effectively complement human mental health professionals, particularly in improving accessibility, scalability, and efficiency in mental health care. However, it is crucial that the deployment of such AI tools be approached with caution, ensuring responsible usage while prioritizing patient safety, privacy, and emotional well-being.

Keywords: *Large Language Models (LLMs), GPT-2, LLAMA, therapeutic AI, mental health care, emotional sensitivity, empathy, crisis management, ethical responses, contextual understanding, AI bias, accessibility, mental health accessibility, therapy simulations, emotional well-being, AI ethics, human-AI collaboration.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Description

The rise of Large Language Models (LLMs) has generated significant interest in their potential applications across various fields, including mental health care. These models, capable of processing and generating human-like text, offer the promise of increasing accessibility to mental health support, providing scalable solutions, and assisting with routine therapeutic tasks. As mental health services continue to be in high demand, these AI-driven technologies present a unique opportunity to supplement human therapists and expand care options to underserved populations.

This project, titled "Machine Minds: Behavioral Assessment of LLM Therapists," seeks to critically evaluate the effectiveness and limitations of these advanced language models in the domain of behavioral therapy. The central aim of this study is to compare the performance of LLMs in therapeutic contexts by simulating therapy-like interactions with patients. Through these simulations, the study assesses how well these models can understand and respond to emotional states, engage in emotionally sensitive conversations, and build trust with users over multiple sessions.

Key therapeutic factors, including the ability to provide appropriate responses based on contextual understanding, maintaining emotional engagement, and demonstrating empathy, will be examined. The study also focuses on evaluating how LLMs manage complex emotional dynamics, such as addressing emotional distress, crisis situations, and handling sensitive topics like trauma and mental health challenges. Additionally, the models' ability to adapt to user needs, maintain consistency in their responses, and avoid biases that may affect the therapeutic relationship will be explored.

The findings of this research will provide valuable insights into the potential and limitations of LLMs in replicating key aspects of behavioral therapy. It will highlight areas where these models may excel, as well as the performance gaps that need to be addressed for these technologies to be viable as supplementary tools in the mental health field. Ultimately, the project aims to contribute to the ongoing discussion of how LLMs can responsibly support human mental health professionals, ensuring their safe and ethical integration into real-world therapeutic settings.

1.2 Background Study: Terminologies and Definitions of New Terms

Large Language Models (LLMs)

Large Language Models (LLMs) are advanced machine learning models designed to understand, generate, and manipulate human language. These models are typically trained on vast amounts of text data from various sources to develop a general understanding of language patterns. LLMs, such as GPT-2 [11], LLAMA [8], StableLM [10], Falcon [9] and Gemini 2.0 [12], leverage architectures like the Transformer to perform tasks such as text generation, translation, summarization, and question answering. In the context of therapy, LLMs are utilized to simulate conversation and assist in emotional support by generating text-based responses that mimic human interaction.

Behavioral Therapy

Behavioral Therapy refers to a type of psychological treatment that focuses on modifying negative behaviors and promoting healthier, more adaptive ones. It is based on principles of behaviorism, which suggests that all behaviors are learned through interaction with the environment. Behavioral therapy can involve techniques such as cognitive-behavioral therapy (CBT), dialectical behavior therapy (DBT), and exposure therapy. These therapies aim to improve an individual's emotional regulation, reduce distress, and build coping mechanisms. The role of empathy, trust-building, and emotional sensitivity is central to its effectiveness, making it an ideal framework for comparing LLMs in therapeutic settings.

Emotional Sensitivity and Empathy

Emotional Sensitivity refers to the ability to recognize, understand, and appropriately respond to emotions, both in oneself and in others. It is a critical component of effective communication, particularly in therapeutic settings, where recognizing emotional cues can guide appropriate interventions. Empathy, on the other hand, is the ability to share and understand another person's feelings. In a therapeutic context, empathy helps the therapist establish a supportive environment where the patient feels understood, validated, and less isolated. LLMs need to demonstrate a degree of emotional sensitivity and empathy to create a trusting relationship with users.

Contextual Understanding

Contextual Understanding refers to the model's ability to interpret the meaning behind a given statement, considering not just the immediate words used, but also the broader context in which they are communicated. This involves understanding cultural nuances, tone, and prior interactions. In therapeutic conversations, contextual understanding is essential for maintaining an appropriate response that aligns with the user's emotional state. LLMs with limited contextual understanding may misinterpret emotions, leading to inappropriate responses or breakdowns in communication.

Human-AI Collaboration in Therapy

Human-AI Collaboration in Therapy involves integrating AI systems like LLMs into therapeutic settings alongside human therapists. While LLMs can assist in automating certain tasks or provide supplemental support, the goal is to ensure that the AI complements rather than replaces human therapists. Human therapists bring essential qualities, such as empathy, intuitive judgment, and ethical responsibility, that are currently beyond the capability of AI models. Thus, collaboration focuses on augmenting therapeutic care, especially in areas where human therapists are in short supply, and improving the efficiency and scalability of mental health services.

1.3 Fundamental study points of the selected topic and the domain

This research examines the role of Large Language Models (LLMs) like LLAMA, GPT-2, Gemini, Falcon and StableLM in providing therapeutic support, focusing on their potential to simulate behavioral therapy. AI, particularly LLMs, promises to enhance mental health care by improving accessibility and scalability of services. However, their effectiveness in managing emotional complexity and offering meaningful support needs thorough evaluation. The study explores whether LLMs can replicate key therapeutic elements such as empathy, active listening, and emotional regulation, and whether they can handle crisis situations. Ethical concerns, including privacy, bias, and emotional reliance, are critical, as is the psychological impact of relying on AI for therapy. Additionally, the research assesses the ability of LLMs to maintain contextual understanding in long-term interactions, an essential aspect of therapeutic engagement. Lastly, the study explores the potential for human-AI collaboration, where LLMs can support human therapists in delivering mental health care, rather than replacing them entirely.

1.4 Identification of challenges in the selected topic

- **Data-Related Challenges:** A major challenge in evaluating LLMs for therapeutic support is the limited availability of high-quality therapy datasets due to privacy concerns and ethical restrictions.[1] Existing datasets, such as HOPE and the High-Low Quality Therapy dataset, lack diversity, which can introduce biases in model training. Additionally, data cleaning and standardization are complex, as therapy conversations vary in structure, requiring extensive preprocessing. Strict privacy laws like HIPAA and GDPR further limit access to real-world therapy data, making it difficult to develop models that provide contextually appropriate and high-quality responses.

- **Model-Related Challenges:** LLMs often fail to replicate high-quality therapy behaviors, frequently overusing problem-solving instead of emotional validation and reflection. They also exhibit inconsistency in responses, where slight prompt variations lead to unpredictable changes in tone and content. Furthermore, LLMs lack memory, preventing them from maintaining long-term context across sessions, making interactions feel impersonal. These limitations reduce trust and effectiveness, preventing LLMs from delivering meaningful therapeutic support.

- **Evaluation-Related Challenges:** Assessing LLM-generated responses using the 13 psychotherapy techniques is difficult because measuring empathy, compassion, and warmth is highly subjective. While behavioral frequency analysis tracks specific techniques, it does not capture response quality or appropriateness. Additionally, comparing LLM outputs with real therapist responses requires expert human evaluation, which is time-consuming and resource-intensive. LLMs also tend to provide solutions prematurely, skipping essential steps like emotional validation, making reliable evaluation an ongoing challenge.

- **Ethical and Safety Concerns:** LLMs lack clinical judgment and may generate harmful or inappropriate responses, especially in high-risk situations like suicidal ideation or trauma-related distress. They can also hallucinate misinformation, potentially misleading vulnerable users. Moreover, the absence of regulatory oversight for AI-based therapy assistants raises concerns about accountability and legal compliance, as LLMs cannot be licensed therapists. Ensuring safety, ethical use, and responsible AI deployment is crucial before these models can be considered for real-world mental health support.

- **Fine-Tuning and Optimization Challenges:** Fine-tuning LLMs for therapy applications is computationally expensive and resource-intensive, especially for large models like LLaMA and Falcon. Small datasets increase the risk of overfitting, reducing model generalization. Additionally, validating fine-tuned models requires human evaluation, which is time-consuming and costly. Even after fine-tuning, biases, inconsistencies, and ethical risks persist, requiring continuous monitoring. Balancing cost, scalability, and real-world effectiveness remains a key challenge in optimizing LLMs for mental health support.

1.5 Problem Statement and Proposed Solution

The increasing use of Large Language Models (LLMs) such as GPT-2, LLAMA, StableLM, Falcon and Gemini 2.0 in various fields has prompted exploration into their potential to provide therapeutic support. While these models are capable of generating human-like text, their ability to replicate the emotionally sensitive and nuanced responses critical in therapy remains uncertain. In mental health care, where empathy, trust, and ethical considerations are paramount, LLMs may struggle to deliver the required level of emotional sensitivity and understanding. This study aims to assess the effectiveness of LLMs in providing therapeutic support, specifically investigating whether they can replace human therapists in therapeutic settings. The research will examine the strengths and limitations of these models in delivering empathetic, ethical, and contextually appropriate responses. Additionally, the study will compare different LLMs to determine which model performs best in terms of emotional sensitivity, reliability, and adherence to therapeutic principles. By addressing these aspects, the project seeks to provide insights into the feasibility of using LLMs as replacements for human therapists and identify the most suitable LLM for mental health care applications.

1.6 Scope of the system

This project evaluates the potential of Large Language Models (LLMs) like GPT-2, LLAMA, StableLM, Falcon and Gemini 2.0 for mental health support. It assesses their ability to provide empathetic, emotionally appropriate, and ethically sound responses in therapeutic contexts. The research compares these models on key metrics such as emotional sensitivity, understanding of distress signals, and adherence to therapeutic principles like empathy, active listening, and trust-building. Additionally, the project explores the possibility of fine-tuning LLMs to better align with mental health best practices and investigates their potential to supplement or complement human therapists. Ultimately, the goal is to identify the strengths and limitations of various LLMs in mental health care and evaluate their role in enhancing accessibility and providing supportive interventions.

Chapter 2

Review of Literature

2.1 Survey of Existing systems

In paper [1], Chiu et al. (2024) introduce the BOLT framework, a behavioral evaluation method designed to assess large language models (LLMs) like GPT-4 and LLaMA within therapeutic contexts. Unlike traditional metrics that focus on fluency or relevance, BOLT evaluates responses based on adherence to techniques from Motivational Interviewing (MI) and Cognitive Behavioral Therapy (CBT). The study reveals that while LLMs can generate coherent and human-like responses, they often overlook critical therapeutic elements such as empathy, emotional validation, and reflective listening. For example, GPT-4 frequently prioritizes problem-solving over emotional support, potentially disrupting the therapeutic process. These findings emphasize that conversational quality alone is not enough for effective mental health support. The paper calls for a shift toward evaluation frameworks that incorporate therapeutic principles and highlights the need for collaboration between AI developers and mental health professionals to improve the emotional and relational quality of LLM responses.

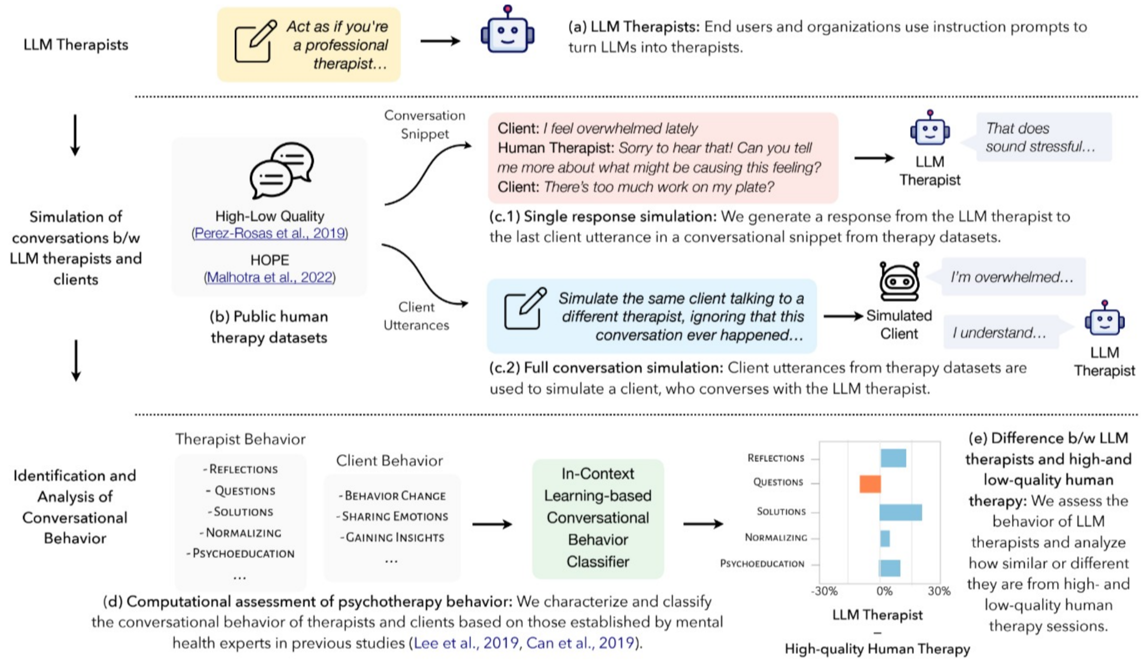

Figure 2.1: Visualization of the BOLT framework, which assesses LLM responses based on MI and CBT techniques. Adapted from Chiu et al. (2024) *[1]*.

As illustrated in Figure 2.1, the BOLT framework maps LLM-generated therapeutic responses to clinically relevant behavioral codes. This structure enables a deeper understanding of how well these models adhere to key principles in human-centered therapy.

In paper [2], Demszky et al. (2023) examine the role of large language models (LLMs) in psychological research and clinical practice, focusing on their potential to model cognitive development and assist in interventions. Although LLMs generate human-like text, the authors note that these models do not possess true understanding; instead, they rely on pattern recognition from large datasets. This fundamental limitation is particularly important in psychology, where comprehension of complex emotional and cognitive nuances is crucial. To address this, Demszky et al. propose two methods for improving LLMs in psychological contexts: fine-tuning and prompt-tuning. Fine-tuning involves retraining models on specialized psychological datasets to help the model better align with psychological principles. Prompt-tuning, on the other hand, involves crafting specific prompts to guide the model's responses without altering its core structure. Both approaches aim to enhance the model's ability to understand and respond to emotional cues in therapeutic settings. The authors emphasize the need for integrating keystone psychological datasets, which would enable LLMs to more effectively handle complex emotional responses. This highlights the importance of continued adaptation and development of LLMs to increase their utility in psychological interventions. While LLMs hold promise for enhancing clinical practice, their effectiveness in psychology will depend on further advancements in model training and integration.

In paper [3], Chung et al. (2023) explore several significant challenges in applying large language models (LLMs) to mental health counseling, with a particular focus on the issue of model hallucination. Hallucination occurs when LLMs generate information that, while plausible, is factually incorrect. This phenomenon raises concerns about the accuracy and reliability of LLMs in clinical contexts, where incorrect information can be harmful and counterproductive. The authors also stress the challenges associated with the interpretability of LLMs, as their decision-making processes are often opaque, making it difficult to understand how responses are generated. Furthermore, the ethical use of electronic health records (EHRs) is highlighted, pointing out the critical need for privacy and security measures when handling sensitive client data in AI-driven therapeutic systems. The lack of empathy and emotional intelligence in LLMs is another key concern, as these models struggle to replicate the deep emotional understanding required for effective therapeutic engagement. To mitigate these issues, the authors suggest several approaches, including fine-tuning models on specific therapeutic datasets, integrating real-time updates to improve accuracy, and employing Reinforcement Learning from Human Feedback (RLHF) to reduce biases and promote fairness. They also recommend the use of explainable AI (XAI) techniques to improve transparency and ensure that the models' actions can be better understood and trusted, ultimately allowing LLMs to be used more ethically and responsibly in clinical settings.

In paper [4], De Choudhury et al. (2023) delve into the dual-edged nature of large language models (LLMs) like GPT in the context of digital mental health care. The authors highlight the transformative potential of LLMs, particularly their ability to democratize mental health services by providing accessible, real-time, and culturally sensitive interventions, especially in areas where professional mental health support is scarce. This could significantly improve access to mental health care for underserved populations. However, the paper also underscores several key limitations that pose risks to the therapeutic process. Although LLMs can simulate human-like conversations, they lack true emotional comprehension, which may undermine the therapeutic alliance. Without this understanding, LLMs might misinterpret emotional distress or fail to respond with the necessary empathy, potentially leading to poor therapeutic outcomes. Moreover, the study raises concerns about the ethical implications of using these models, particularly around misinformation, bias, and privacy. Marginalized groups may face increased vulnerability, as the automated nature of LLMs could exacerbate biases or lead to incorrect interventions. Additionally, privacy risks regarding the handling of sensitive mental health data are a significant concern, especially given the potential for unauthorized access or misuse. In response to these challenges, the authors advocate for an ”AI-in-the-loop” model, where LLMs are used to support, rather than replace, human therapists. This hybrid approach would allow LLMs to provide preliminary support, such as psychoeducation or initial assessments, while ensuring that human therapists retain control over critical decision-making and emotional engagement, thus safeguarding the quality of care.

2.2 Limitations of Existing Systems or Research Gaps

Current systems for evaluating LLMs as therapists face key limitations, underscoring the need for a structured framework. LLMs lack consistent therapeutic context, often failing to apply core techniques reliably and missing an evaluation framework for such behaviors. Their emotional sensitivity is inconsistent, leading to robotic or even harmful responses, while the absence of benchmarks against high-quality therapy practices makes it difficult to gauge their therapeutic reliability. LLMs also tend to over-prioritize problem-solving, which can hinder emotional validation essential for a strong therapeutic alliance. Their "black box" nature complicates ensuring responses align with therapeutic values, raising trust and accountability issues. Additionally, biases in training data can result in insensitive responses, creating ethical concerns, and they are ill-prepared for crisis intervention, lacking standardized safety protocols. These limitations highlight the need for a robust framework that includes therapeutic benchmarks, interpretability, bias mitigation, and safety standards in LLM-assisted therapy.

2.3 Additional Insights from Recent Research

Recent studies further enrich the understanding of LLMs in therapeutic contexts.

Brown et al. (2020) [14] introduced GPT-3, demonstrating that large-scale language models are capable of few-shot learning. This highlights the potential for minimal intervention-based fine-tuning of LLMs in mental health support, though it also raises concerns regarding control and predictability of outputs in sensitive scenarios.

Sharma et al. (2020) [15] proposed a computational method to measure empathy expressed in text-based mental health support conversations. Their work emphasizes the critical role of empathetic language, reinforcing the need for LLMs to be evaluated not only on fluency but on emotional resonance as well.

Building on this, Sharma et al. (2024) [16] explored how human-LLM interaction could facilitate self-guided mental health interventions, particularly in cognitive restructuring tasks. This study supports the idea that LLMs can serve as scaffolds for therapeutic exercises when properly guided, but still lack autonomous therapeutic intelligence.

Finally, Shah et al. (2024) [17] presented the Machine Minds framework for behavioral evaluation of LLMs in therapy, focusing on emotional sensitivity, reliability, and therapeutic techniques. This research, forming the foundation of the current project, identified critical gaps in LLM-generated therapeutic conversations and emphasized the importance of human-aligned evaluation methods.

2.4 Motivation

The motivation for our research stems from the growing recognition of the challenges associated with deploying Large Language Models (LLMs) in mental health care. While LLMs like GPT-2 and LLAMA are capable of generating human-like responses, their lack of true emotional understanding raises significant concerns. These models may struggle to provide the nuanced, sensitive, and validating support required in therapeutic contexts, especially when dealing with individuals who are vulnerable or experiencing distress. In high-stakes scenarios, where emotional well-being is at risk, the reliability and safety of these models are paramount. However, LLMs often misinterpret distress signals, which can result in inappropriate or harmful responses that undermine trust in the therapeutic process. Furthermore, LLMs lack the ability to genuinely form therapeutic alliances, a crucial aspect of effective mental health support. This shortcoming compromises the depth of the client-therapist relationship, which is integral to the success of traditional therapy. Our research aims to address these critical gaps by developing a comprehensive assessment framework that evaluates key therapeutic parameters such as emotional sensitivity, reliability, contextual understanding, and transparency. By focusing on these factors, our goal is to create a framework that enhances the trustworthiness, accountability, and ethical alignment of LLMs with therapeutic values, ultimately improving their potential to safely and effectively complement human therapists in mental health care.

Chapter 3

Proposed System

3.1 Detailed Explanation of Proposed System

The proposed system, consists of three main modules: Preprocessing Models, Therapeutic Response Generation Module, and Output Module. The process begins with Text Input, which undergoes Preprocessing involving Tokenization (breaking text into smaller components), Context Management (ensuring conversational continuity), and Analysis (extracting key information from the input). The processed text is then passed to the Therapeutic Response Generation Module, where a suitable Model Selection is performed. The chosen model is then Fine-Tuned using specialized mental health datasets to enhance response accuracy and relevance. After fine-tuning, the model generates a Therapeutic Response. The final stage, the Output Module, involves Generating Responses, Feedback Collection (evaluating response quality and effectiveness), and Summary Generation, where a concise summary of the conversation is produced. This structured approach ensures that the generated responses are contextually appropriate, empathetic, and effective for mental health support.

3.1.1 Block Diagram of Proposed System / Workflow

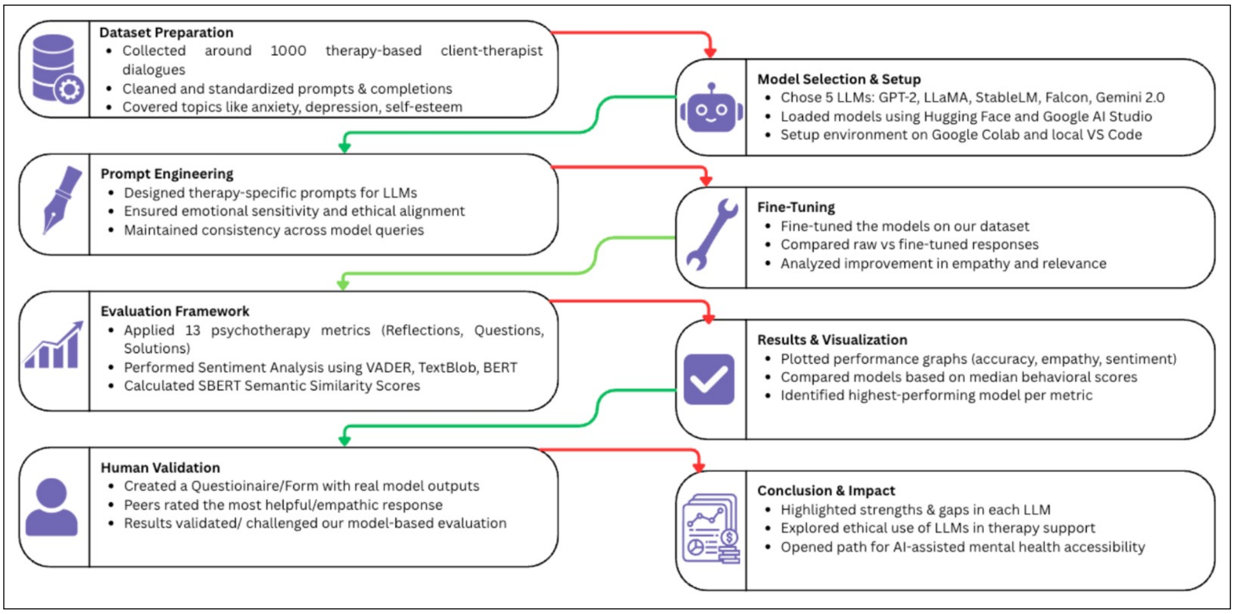

Figure 3.1: Block Diagram of the Proposed System

Figure 3.1 presents the workflow of the proposed system, illustrating the core phases involved in analyzing LLM-based therapeutic responses.

3.1.2 Working Principle (Algorithm)

The following algorithm outlines the execution flow of the proposed system, which evaluates and generates therapeutic responses using multiple LLMs. The algorithm integrates input handling, model processing, and human-in-the-loop evaluation to ensure response relevance and empathy.

Step 1: User Input Acquisition

1. Accept a user query or therapy-related prompt as input.
2. Preprocess the input to ensure tokenization and proper formatting.

Step 2: Prompt Routing and Standardization

1. Map the input query to a standardized therapy-specific prompt.

2. Ensure prompt consistency and ethical alignment before sending it to any model.

Step 3: Parallel Model Invocation

1. Forward the prompt to each of the fine-tuned LLMs (GPT-2, LLaMA, Falcon, StableLM, Gemini 2.0).
2. Collect raw responses generated by each model.

Step 4: Automated Response Evaluation

1. Evaluate each model's response using:
   - 13 psychotherapy metrics (e.g., empathy, active listening, solution orientation).
   - Sentiment analysis using VADER, TextBlob, and BERT.
   - SBERT semantic similarity scores.
2. Assign performance scores to each model response.

Step 5: Human-in-the-Loop Validation

1. Display anonymized model responses in a questionnaire or form.
2. Collect human ratings for empathy, helpfulness, and relevance.
3. Compare human scores with automated metrics to validate effectiveness.

Step 6: Final Output and Analysis

1. Identify the top-performing model based on both human and automated evaluations.
2. Generate a summary of the interaction for record-keeping or therapeutic continuity.
3. Store feedback to improve future model responses.

3.1.3 Phase/Module-Wise Explanation

Phase 1: Dataset Preparation & Prompt Engineering

According to figure 3.1, this phase focuses on gathering and preparing the therapy-based dataset and crafting effective prompts to guide LLMs in generating emotionally intelligent responses.

- **Dataset Preparation:** Collected around 1000 client-therapist dialogues focused on topics such as anxiety, depression, and self-esteem. The dialogues were cleaned and standardized to ensure quality input for model training.
- **Prompt Engineering:** Designed therapy-specific prompts for large language models (LLMs), ensuring emotional sensitivity and ethical alignment. Prompts were made consistent across all model queries to enable fair evaluation.

Phase 2: Model Setup & Fine-Tuning

In this phase, the appropriate models are selected, set up, and fine-tuned using the curated dataset.

- **Model Selection & Setup:** Five LLMs were used—GPT-2, LLaMA, StableLM, Falcon, and Gemini 2.0. Models were accessed via Hugging Face and Google AI Studio, with environments set up on Google Colab and local VS Code.
- **Fine-Tuning:** The models were fine-tuned on the client-therapist dataset. This helped improve empathy, emotional relevance, and therapeutic alignment. Comparisons between raw and fine-tuned responses were also performed.

Phase 3: Evaluation Framework

This phase involves a comprehensive evaluation of model performance using psychotherapy metrics and sentiment analysis tools.

- **Psychotherapy Metrics:** Applied 13 metrics such as Reflections, Questions, and Solutions to evaluate therapeutic alignment.
- **Sentiment Analysis:** Used VADER, TextBlob, and BERT to assess the emotional tone of responses.
- **Semantic Similarity:** Measured using SBERT to evaluate relevance and coherence of the model outputs.

Phase 4: Results Visualization

Visualized and analyzed the results of model responses across different performance metrics.

- **Performance Graphs:** Plotted graphs based on accuracy, empathy, and sentiment scores.
- **Model Comparison:** Compared models based on median scores and highlighted top-performing models per metric.

Phase 5: Human Validation

This module validates model-generated outputs through human feedback.

- **Questionnaire-Based Validation:** Created a form containing real model outputs, which were rated by peers on helpfulness and empathy.
- **Cross-Verification:** Human feedback validated or challenged the model-based evaluations, enhancing the credibility of the results.

Phase 6: Conclusion & Impact

Summarized insights from the project, discussed ethical implications, and highlighted future directions.

- **Insights:** Identified strengths and limitations of each LLM in therapeutic settings.
- **Ethics & Accessibility:** Explored the ethical use of AI in therapy and the potential to improve mental health accessibility.

3.2 System Analysis

3.2.1 Functional Requirements

- **Simulated Therapy Conversations:** Simulated Therapy Conversations: Design a framework for conducting simulated therapy sessions with each LLM, generating responses to standardized prompts based on common therapeutic scenarios (e.g., anxiety, depression, interpersonal issues). This setup will allow consistent testing across different LLMs to assess their therapeutic potential.
- **Behavioral Metrics Evaluation:** Define and implement behavioral assessment metrics to evaluate each LLM's responses, with a focus on empathy, emotional appropriateness, ethical sensitivity, and conversational coherence in simulated therapeutic interactions.
- **Comparative Analysis:** Develop a comparative system to collect and analyze responses from multiple LLMs (e.g., GPT-3, GPT-4, LLAMA, Bloom, Falcon) across various scenarios, identifying strengths, weaknesses, and variations in therapeutic effectiveness.
- **Data Collection and Storage:** Implement a secure storage system to archive session transcripts and associated metrics, enabling further analysis and potential model fine-tuning, while adhering to data management best practices.
- **User Feedback Loop:** Integrate feedback from mental health professionals who will review the LLM responses. This input will guide improvements by providing insights on emotional understanding, ethical appropriateness, and potential risks if these models were deployed in real-world therapeutic settings.

3.2.2 Non-Functional Requirements

- **Security and Privacy:**Ensure robust data protection and compliance with privacy standards, particularly regarding sensitive information generated during simulated therapeutic conversations.
- **Reliability and Consistency:** Design for high reliability and consistency in LLM performance, ensuring stable and coherent responses across similar scenarios and avoiding erratic or harmful advice.
- **Usability:** Develop an intuitive, user-friendly interface for mental health professionals, featuring clear, interpretable metrics and tools to assess LLM performance efficiently.
- **Ethical Safeguards:** Establish ethical safeguards by embedding clear guidelines and disclaimers on AI limitations in therapeutic contexts, making it clear that LLMs are supportive tools rather than replacements for human therapists.

3.2.3 Software and Hardware Requirements

Software Requirements

- **Programming Languages:**
  - Python 3.x (primary language for LLM interactions, data processing, and analysis).
- **Libraries and Frameworks:**
  - **Transformers (Hugging Face)**: For loading and interfacing with different LLMs.
  - **PyTorch or TensorFlow**: For model management and fine-tuning.
  - **NLTK or SpaCy**: For natural language processing tasks, sentiment analysis, and text pre-processing.

  - **Flask or Django**: If building a web-based interface for interacting with the system.
  - **NumPy and Pandas**: For data manipulation and analysis.
  - **Matplotlib or Seaborn**: For visualizing data and model performance metrics.
  - **SQLite, PostgreSQL, or MongoDB**: For secure storage of conversation logs, feedback, and metrics.
- **Development Environment:**
  - Jupyter Notebook or Google Colab (for prototyping and model testing).
  - IDEs like PyCharm or VSCode (for development and integration).
- **APIs:**
  - Hugging Face API or OpenAI API (if using API-based access to LLMs).
  - Secure authentication protocols (OAuth2, API keys) for safe model usage.
- **Operating System:**
  - Compatible with Linux, macOS, or Windows 10/11.

Hardware Requirements

- **CPU:**
  - Intel i7/AMD Ryzen 7 or above (minimum for testing and running LLMs locally).
  - High-performance multi-core CPU recommended for faster computations.
- **GPU:**
  - NVIDIA RTX 30-series (3060 or higher) or equivalent AMD GPU, with at least 8 GB VRAM for local model fine-tuning.

  - Preferably an NVIDIA GPU (A100, V100) for enhanced support with PyTorch and TensorFlow if large model fine-tuning is required.

- **RAM:**
  - Minimum 16 GB (for basic testing and smaller model handling).
  - Recommended 32 GB or higher for efficient handling of larger datasets and model processing.

- **Storage:**
  - SSD with a minimum of 512 GB storage (for faster data access and model loading).
  - Additional external or cloud storage (if dealing with large datasets or multiple LLMs).

- **Internet Connection:**
  - Reliable, high-speed internet (for accessing cloud-based LLMs, APIs, and collaborative work environments).

- **Additional Hardware (Optional):**
  - High-resolution display (for ease in analyzing metrics and UI design).
  - External hard drive or NAS for secure backup of data and project files.

3.3 Use Case Modelling

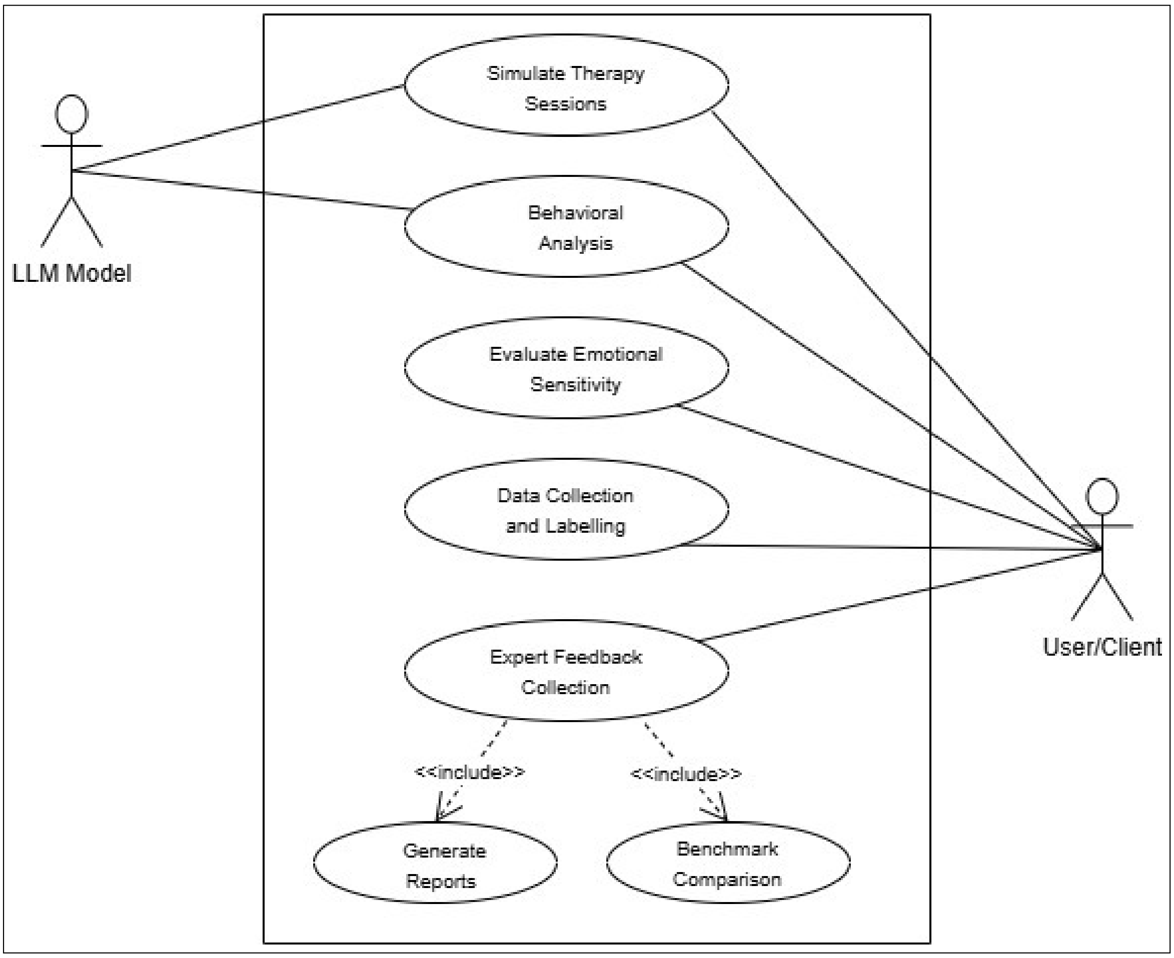

Figure 3.2: Use Case diagram

Figure 3.2 presents the use case diagram for the Machine Minds system, illustrating the interactions between users (mental health professionals or individuals seeking support) and large language models (LLMs). The diagram outlines key processes such as initiating a therapy session, receiving feedback, and providing continuous evaluation for improvement. The system starts with the user engaging in a session, where the LLM generates responses tailored to their needs. A feedback mechanism ensures the system adapts to the user's emotional tone and therapeutic needs over time. Continuous evaluation, both during and after sessions, helps track response quality and effectiveness, ensuring ongoing improvement. This iterative process aims to enhance the system's ability to provide personalized and effective therapeutic support, bridging the gap between

AI and human expertise.

Table 3.1: Simulate Therapy Sessions Use Case

Use Case	Simulate Therapy Sessions
Primary Actor	LLM Model
Goal in Context	User Interacts with the system for therapy simulation, LLM Model engages in simulated therapy sessions.
Preconditions	The LLM model is trained and capable of mimicking therapeutic conversations.
Basic Flow	The User/Client initiates a simulated therapy session. The LLM Model generates responses based on therapeutic context.
Alternative Flows	If the model fails to respond appropriately, it logs the issue for refinement.
Postcondition	The session is completed, and responses are saved for analysis.

Table 3.2: Behavioral Analysis Use Case

Use Case	Behavioral Analysis
Primary Actor	System
Goal in Context	The system analyzes the LLM Model's behavior during simulated sessions. Results are compiled for evaluation.

Preconditions	Analysis results are stored.
Scenario	System detects notes and onsets from the spectrogram that it creates and then aggregates and interprets data for a complete transcription.
Alternative Flows	If data is insufficient, prompts the user to continue collecting interaction data.
Postcondition	Analysis results are stored.

3.4 Proposed System: Analysis, Modelling and Design

3.4.1 Activity Diagram

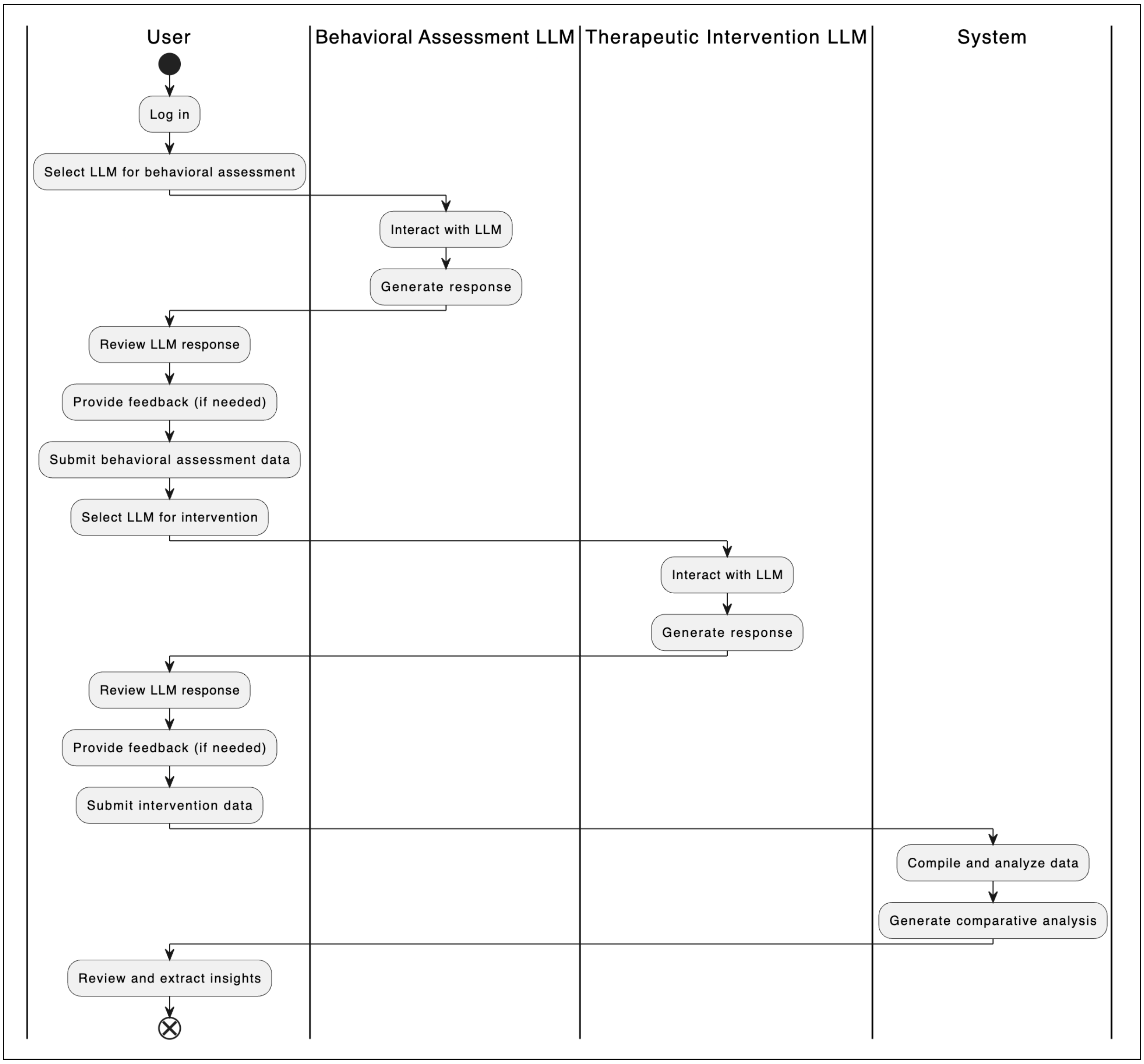

Figure 3.3: Activity Diagram

Figure 3.3 outlines the workflow for assessing Large Language Models (LLMs) in therapeutic contexts. The process starts with user login and selection of an LLM for behavioral assessment. The user interacts with the LLM, reviews its responses, and submits feedback and behavioral data. Finally, the system compiles the assessment and intervention data to generate a comparative analysis, enabling insight extraction for model evaluation.

3.5 Data Flow Diagram (DFD)

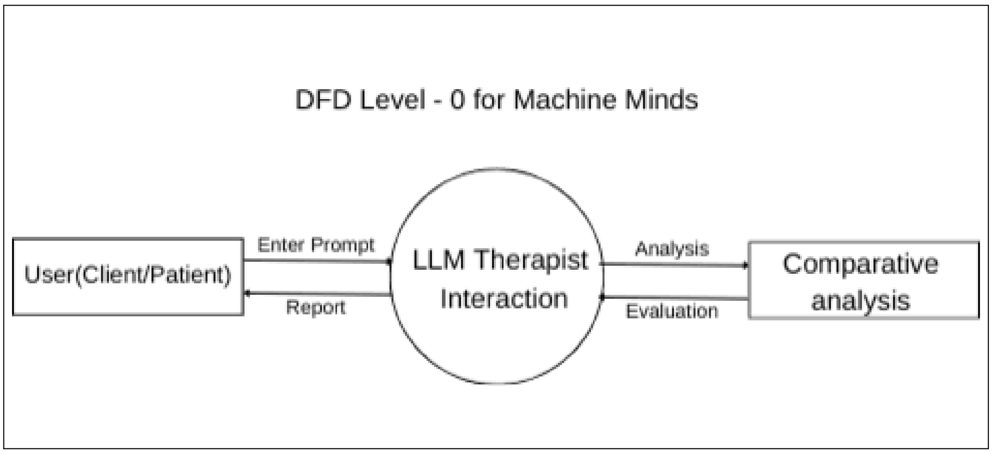

Figure 3.4: Data Flow Diagram Level-0

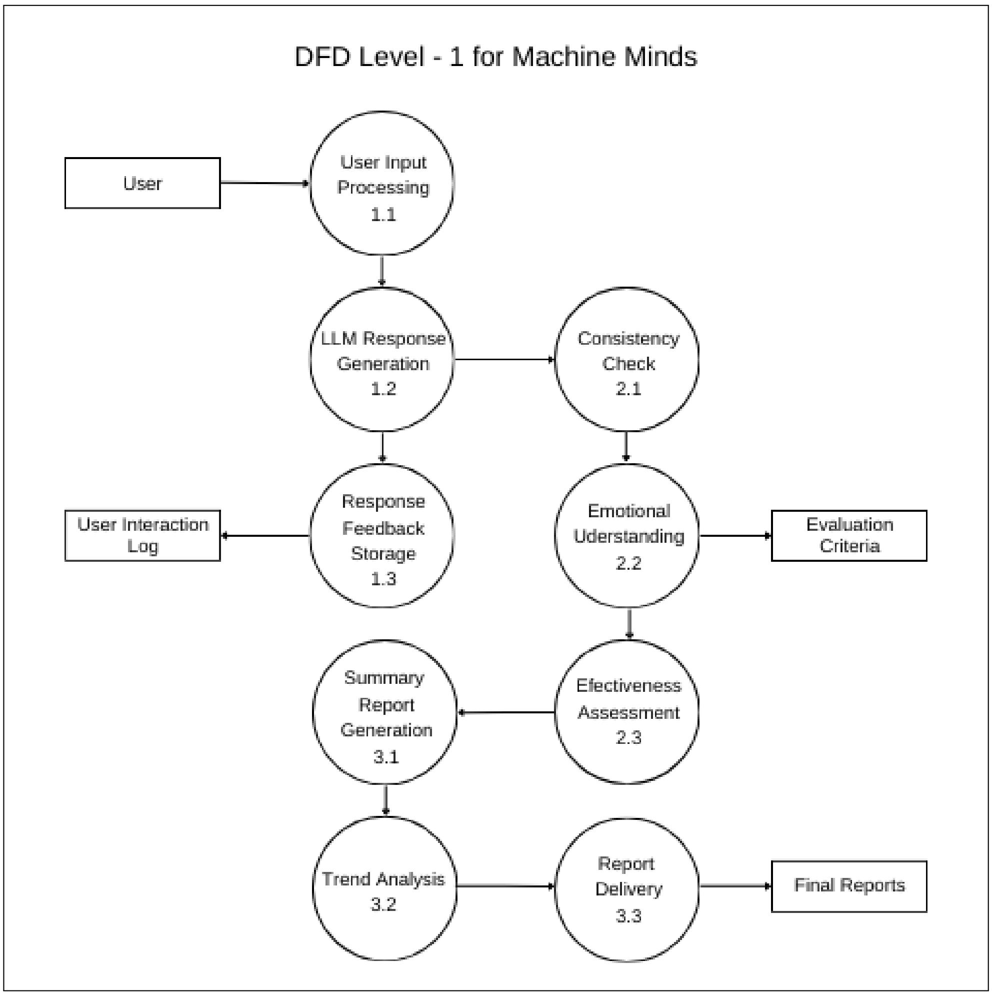

Figure 3.5: Data Flow Diagram Level-1

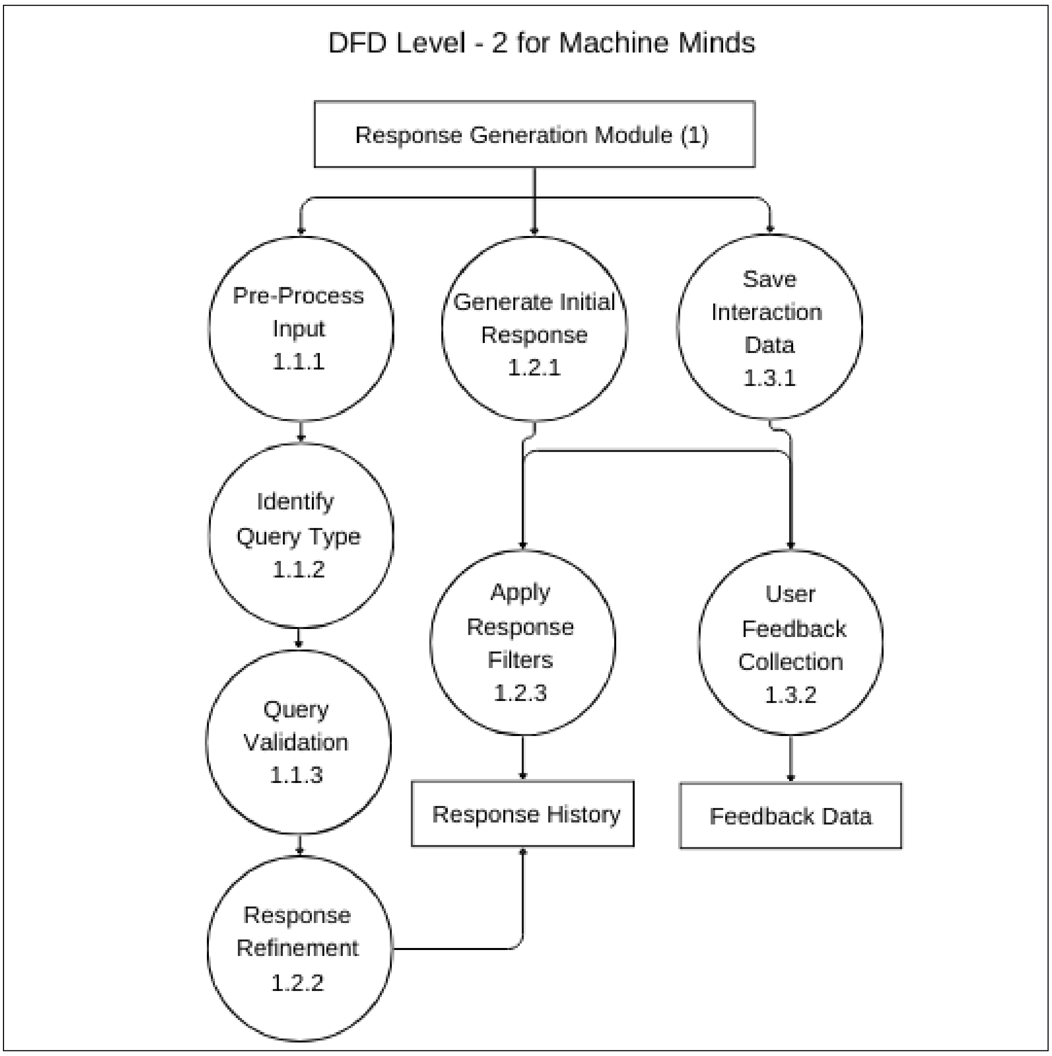

Figure 3.6: Data Flow Diagram Level-2

Figure 3.4, figure 3.5 and figure 3.6 illustrate the workflow of assessing Large Language Models (LLMs) for therapeutic support. Level 0 presents the basic user interaction, where input prompts lead to LLM responses and subsequent metrics evaluation. Level 1 further details the process, including prompt generation, LLM response analysis, and storage of results. Level 2 breaks down the metrics evaluation into submodules, covering empathy, ethicality, coherence, and comparative analysis, giving a comprehensive view of how each LLM's therapeutic effectiveness is assessed.

3.6 Architectural View

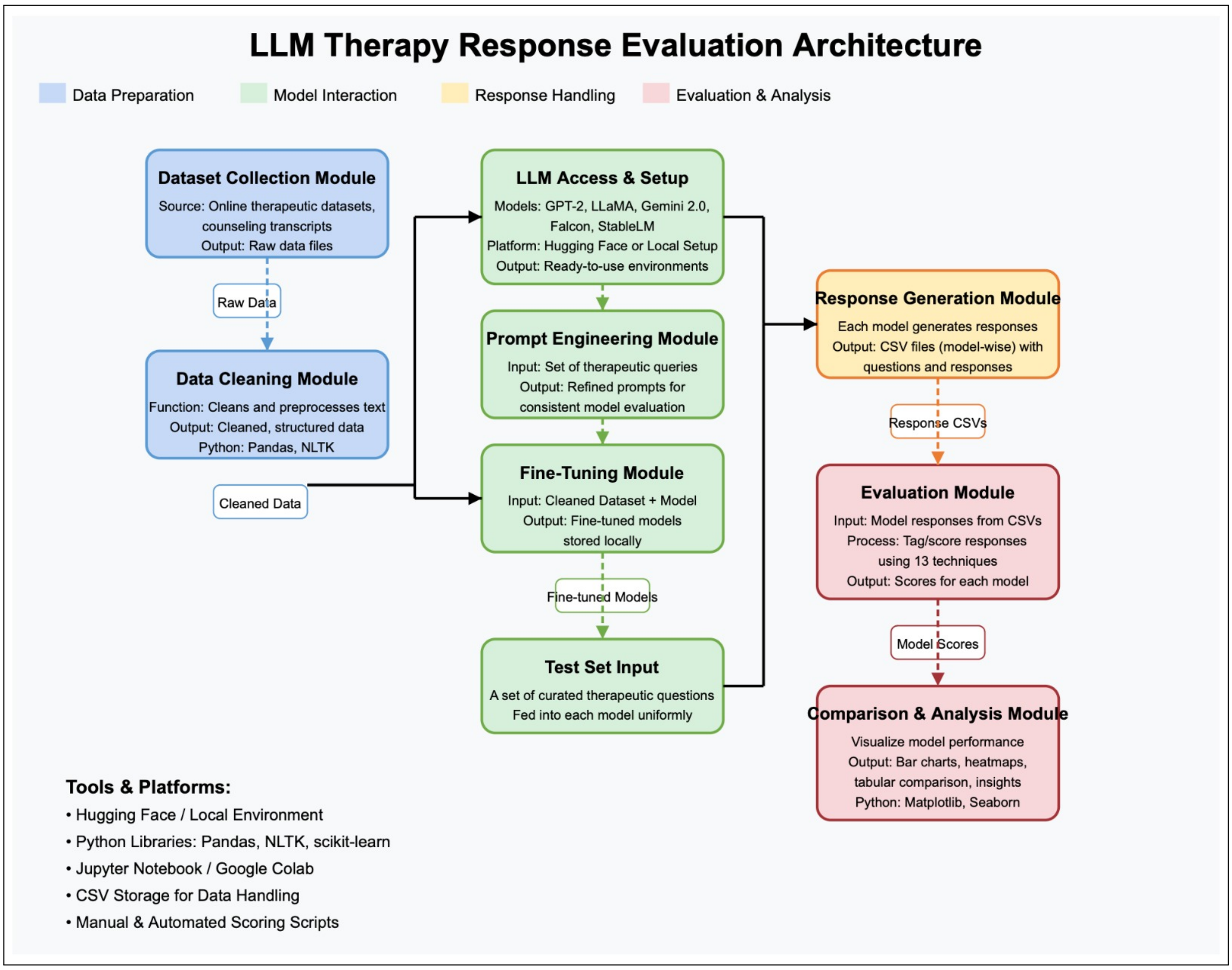

Figure 3.7: System Architecture

Figure 3.7 illustrates the LLM Therapy Response Evaluation Architecture. The workflow is divided into four main stages: Data Preparation, Model Interaction, Response Handling, and Evaluation Analysis. It begins with collecting and cleaning therapeutic conversation datasets. The cleaned data is used for prompt engineering and fine-tuning various LLMs such as GPT-2, LLaMA, Gemini, Falcon, and StableLM. A curated test set is input uniformly to all models, which then generate responses stored in CSV format. These responses are evaluated using 13 psychotherapy techniques. The final step involves comparison and analysis of model scores using visualization tools like bar charts and heatmaps, supported by Python libraries such as Matplotlib and Seaborn.

3.7 Algorithms / Methodology

3.7.1 Simulated Therapy Conversations:

To evaluate LLMs in a therapeutic context, we simulate therapy-like sessions by generating prompts based on common mental health issues, such as anxiety, depression, and interpersonal conflicts. Each model's responses are collected across standardized scenarios to ensure consistency in data for evaluation.

3.7.2 Behavioral Metrics Evaluation:

We apply a set of behavioral metrics to assess LLM responses. These metrics include:

- Empathy Detection: Evaluates the degree to which responses show understanding and validation of users' emotions.
- Ethical and Sensitive Response Scoring: Measures how appropriately the model addresses sensitive topics and adheres to ethical guidelines.
- Conversational Coherence: Assesses logical flow, response relevance, and linguistic clarity.
- Consistency: Ensures the model's responses align across similar scenarios, indicating reliability.

3.7.3 Sentiment and Emotion Analysis:

Sentiment analysis algorithms help quantify the emotional tone of responses, checking if the LLM conveys appropriate levels of empathy and sensitivity. Emotion classifiers are employed to determine the emotional alignment with the user's inputs.

3.7.4 Comparative Analysis Framework:

A comparative analysis algorithm aggregates the metric scores for each LLM and evaluates them across multiple therapeutic scenarios, identifying strengths, weaknesses, and variations in each model's performance.

3.7.5 Feedback Integration with Reinforcement Learning:

We incorporate insights from mental health professionals reviewing the LLM responses to guide iterative improvements. Techniques like Reinforcement Learning from Human Feedback (RLHF) allow LLMs to adapt based on professional input, refining their performance in future interactions.

3.7.6 Data Collection and Storage Mechanisms:

Each session's data, including prompts, responses, and metric evaluations, is securely stored to support continuous analysis and potential model fine-tuning.

Chapter 4

Implementation Plan and Experimental Setup

4.1 Experimental Setup

4.1.1 Dataset and Input Description

The **Patient-Therapist Dataset** [13] is designed for analyzing mental health interactions between patients and therapists, particularly focusing on the behavior of large language models (LLMs) in therapeutic settings.

Dataset Information:

- **Name of the dataset:** Patient-Therapist.csv
- **Sources:** The dataset has been compiled from various sources, including Kaggle, Reddit, Twitter, and other existing datasets to ensure a comprehensive analysis. This dataset has been sourced from Kaggle, contributed by Zuhair Hasan Shaik.
- **Fields Used:** The dataset includes the fields `prompt` and `completion`, which represent conversations between the client and the AI therapist, respectively.
- **Size and Number of Records:** The dataset contains 640 records, with a file size of approximately 900 KB.

About the Dataset:

- The dataset covers a wide range of mental health topics, including anxiety, depression, relationships, self-esteem, coping mechanisms, and more.
- The questions and answers in the dataset aim to help individuals gain insights into various aspects of therapy and mental health.

4.1.2 Performance Evaluation Parameters

To assess the model's responses, we use 13 psychotherapy techniques, categorized as follows:

1. **Reflections**
   Reflections are a core component of high-quality therapy, used to show empathy and deepen the client's understanding of their thoughts and emotions.

   - **Reflections on Needs** – The therapist identifies and verbalizes the client's underlying needs.

     Client-Therapist Dialogue

     Client: "I feel like no one understands me."
     Therapist: "It sounds like you need to feel heard and validated by those around you."

   - **Reflections on Emotions** – The therapist acknowledges and labels the client's emotional state.

     Client-Therapist Dialogue

     Client: "I just feel so lost and unmotivated."
     Therapist: "You seem to be experiencing a deep sense of uncertainty and frustration."

   - **Reflections on Values** – The therapist reinforces the client's core beliefs and values.

Client-Therapist Dialogue

Client: "I don't want to be a burden to my family."

Therapist: "It seems like being independent and not relying too much on others is very important to you."

- **Reflections on Consequences** – The therapist highlights potential outcomes.

Client-Therapist Dialogue

Client: "I always avoid difficult conversations, even when I know I should speak up."

Therapist: "Avoiding those conversations might give you temporary relief, but could lead to long-term misunderstandings."

- **Reflections on Conflicts** – The therapist identifies internal struggles.

Client-Therapist Dialogue

Client: "I want to succeed, but I also feel scared to take risks."

Therapist: "You seem to be torn between your ambition and your fear of failure."

- **Reflections on Strengths** – The therapist acknowledges the client's strengths.

Client-Therapist Dialogue

Client: "Even though things are tough, I keep pushing forward."

Therapist: "It sounds like you have a lot of perseverance and inner strength to keep going despite challenges."

2. **Questions**

- **Questions on Experiences**

Client-Therapist Dialogue

”Can you tell me more about what happened when you felt that way?”

- **Questions on Perspectives**

Client-Therapist Dialogue

”How do you think someone else might interpret this situation?”

- **Questions on Emotions**

Client-Therapist Dialogue

”What emotions come up for you when you think about this experience?”

3. **Solutions**

- **Problem-Solving**

Client-Therapist Dialogue

”One way to manage your anxiety could be practicing deep breathing before stressful situations.”

- **Planning**

Client-Therapist Dialogue

”What small steps could you take this week to work toward feeling more confident in social situations?”

4. **Normalizing**

Client-Therapist Dialogue

Client: ”I feel like I should be able to handle this on my own, but I can't.”
Therapist: ”It's completely normal to feel that way—many people struggle with asking for help.”

5. **Psychoeducation**

Client-Therapist Dialogue

Client: ”I don’t understand why I keep overthinking everything.”
Therapist: ”Overthinking can be linked to anxiety, where the brain tries to anticipate problems. One way to manage it is by practicing mindfulness.”

4.2 Model Training and Implementation

4.2.1 Fine-tuning an LLM – GPT-2 Medium

Fine-tuning GPT-2 Medium

```
from transformers import AutoTokenizer, AutoModelForCausalLM, Trainer
    , TrainingArguments
from datasets import Dataset
import pandas as pd
import torch

# Load tokenizer and model
tokenizer = AutoTokenizer.from_pretrained("gpt2-medium")
model = AutoModelForCausalLM.from_pretrained("gpt2-medium")
tokenizer.pad_token = tokenizer.eos_token

# Load and preprocess dataset
data = pd.read_csv("cleaned_dataset.csv")
dataset = Dataset.from_pandas(data)
dataset = dataset.train_test_split(test_size=0.1)

# Tokenize datasets
def preprocess_function(examples):
    inputs = [f"Client: {p}\nTherapist: {r}" for p, r in zip(examples
        ["client"], examples["therapist"])]
    tokenized = tokenizer(inputs, truncation=True, padding="
        max_length", max_length=128)
    tokenized["labels"] = tokenized["input_ids"].copy()
    return tokenized

dataset = dataset.map(preprocess_function, batched=True,
    remove_columns=["client", "therapist"])

# Define training arguments
```

```
training_args = TrainingArguments(
    output_dir="./gpt2-medium-therapeutic",
    evaluation_strategy="epoch",
    learning_rate=5e-5,
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    num_train_epochs=5,
    save_strategy="epoch",
    save_total_limit=2,
    logging_dir="./logs",
    fp16=True,
)

# Train model
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=dataset["train"],
    eval_dataset=dataset["test"],
    tokenizer=tokenizer,
)
trainer.train()
```

4.2.2 Generating Responses

Generating Responses

```
def generate_response(input_text):
    prompt = f"Client: {input_text}\nTherapist:"
    inputs = tokenizer(prompt, return_tensors="pt", padding=True, truncation=True, max_length=128).to("cuda")
    outputs = model.generate(inputs["input_ids"], max_new_tokens=120, temperature=1, top_p=0.9, repetition_penalty=1.5)
    response = tokenizer.decode(outputs[0], skip_special_tokens=True)
    response = response.split("Therapist:")[1].strip() if "Therapist:" in response else response.strip()
    return response

# Test Inputs
test_inputs = [
    "I'm feeling so overwhelmed at work. What should I do?",
    "I can't stop thinking about my breakup. It's consuming me.",
    "I feel like  I m  not good enough for my family.",
    " I ve  lost someone I deeply care about, and I  c a n t  move on."
]
```

```


for input_text in test_inputs:
    print(f"Input: {input_text}")
    print(f"Response: {generate_response(input_text)}\n")
```

4.3 Model Output and Analysis

Table 4.1: Client Input and Model Response

Client Input	**Model Response**
I’m feeling so overwhelmed at work. What should I do?	Taking breaks and prioritizing tasks can help manage stress. You’re not alone in feeling this way.
I can’t stop thinking about my breakup. It’s consuming me.	Breakups can be painful, and it’s okay to take time to heal. Talking to close ones may help.
I feel like I’m not good enough for my family.	You are valuable just as you are. Self-worth is not defined by others’ opinions.
I’ve lost someone I deeply care about, and I can’t move on.	Grieving takes time. Surrounding yourself with support and cherishing good memories can help.

The table 4.1 contains examples of emotionally vulnerable client statements alongside thoughtful, supportive responses. The inputs reflect common mental health challenges such as workplace stress, heartbreak, low self-esteem, and grief. Each response offers empathy, reassurance, and gentle guidance, aiming to reflect the kind of understanding and encouragement often found in therapeutic conversations. The purpose of this table is to show how emotionally attuned and compassionate language can help individuals feel heard and supported in moments of distress.

Chapter 5

Results and Discussions

5.1 Presentation and Validation of the Results

5.1.1 Overview of the Evaluation Framework

This study systematically evaluates the therapeutic capabilities of five large language models (LLMs): LLaMA, GPT-2, Gemini, Falcon, and StableLM. Each model was tested on its ability to simulate therapist-like responses across 13 defined psychotherapy techniques. The responses were evaluated using a combination of semantic similarity (via Sentence-BERT) and sentiment polarity (via VADER).

The dual evaluation framework ensured that the assessment captured not only the factual alignment of the responses but also their emotional appropriateness—both being essential in mental health conversations.

5.1.2 Step-by-Step Evaluation Process

Step 1: Dataset Preparation and Model Inference

A carefully curated dataset comprising 600 therapy-like prompts was constructed. These prompts covered diverse emotional scenarios, ranging from anxiety and depression to motivational issues and relationship conflicts. Each model generated responses to the same set of prompts, ensuring a uniform comparison basis.

Step 2: Application of Evaluation Metrics

Two complementary metrics were calculated for each model's responses:

- **Semantic Similarity Score:** Using MiniLM-L6-v2, the similarity between the generated response and the ideal therapist response was computed.
- **Sentiment Polarity Score:** Using VADER, the positivity or negativity of the emotional tone was scored.

A weighted composite score was calculated as:

$$CompositeScore = (0.6 \times SemanticSimilarity) + (0.4 \times SentimentPolarity)$$

Step 3: Psychotherapy Technique-Specific Classification

Each response was then mapped to one of the 13 psychotherapy techniques: Reflection, Validation, Clarification, Confrontation, Immediacy, Interpretation, Open Questions, Planning, Psychoeducation, Relationship Focus, Resource Activation, Risk Assessment, and Self-disclosure.

Technique-specific behavioral patterns were recorded, such as the ability of models to accurately validate client emotions or formulate effective plans.

Step 4: Aggregation, Visualization, and Statistical Summarization

The results were visualized using bar charts for individual techniques and a combined heatmap for overall comparison. Statistical summaries, including mean scores and standard deviations, were prepared to aid in comprehensive analysis.

5.1.3 Graphical and Tabular Representation of Results

Technique-wise Comparative Charts:

The following figure 5.1 illustrates "13 Psychotherapy techniques" performed on all 5 models:

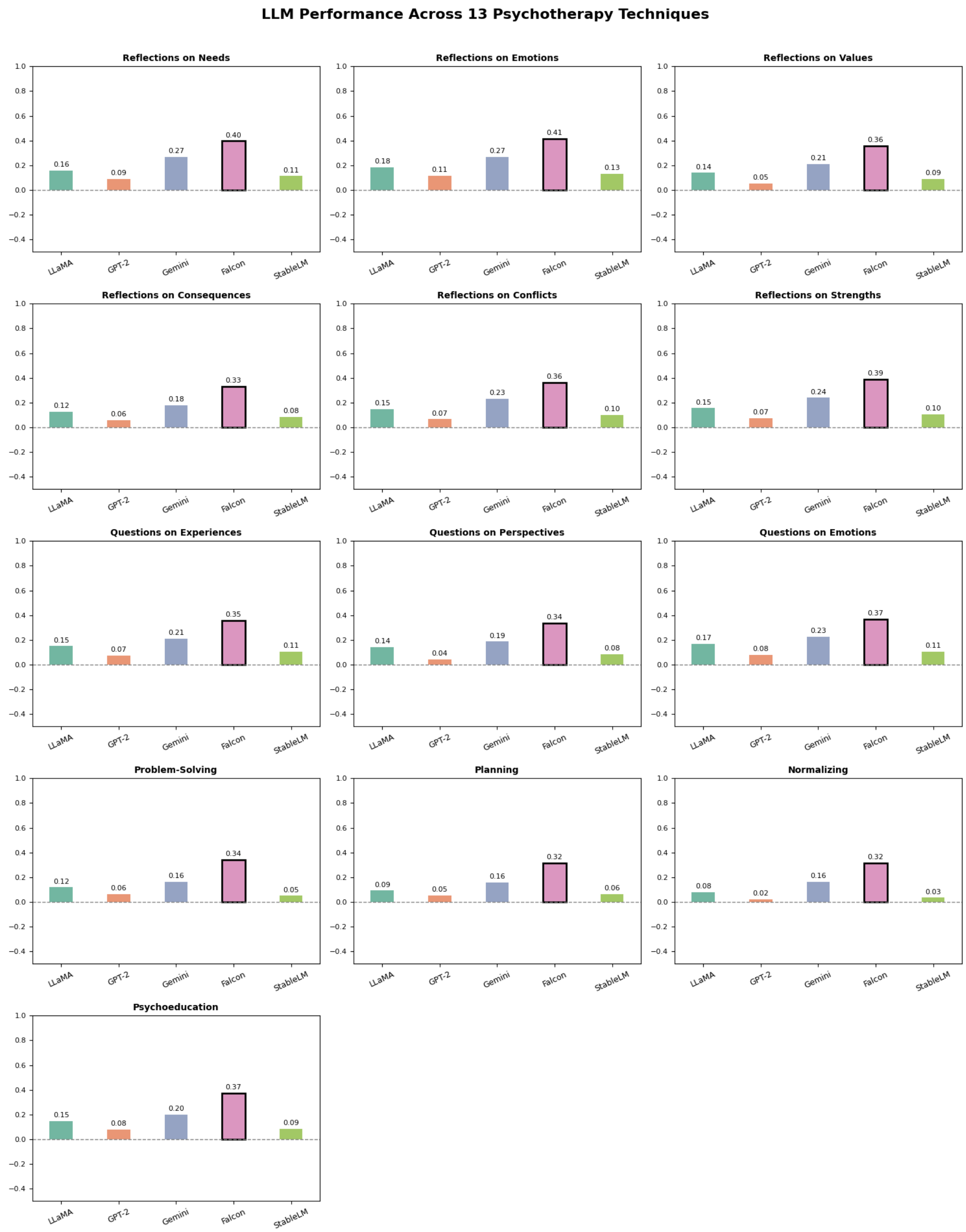

Figure 5.1: Performance of models on "13 Psychotherapy techniques"

Combined Performance Heatmap:

The heatmap shown in figure 5.2 provides an at-a-glance view of the models' comparative strengths and weaknesses across different psychotherapy skills:

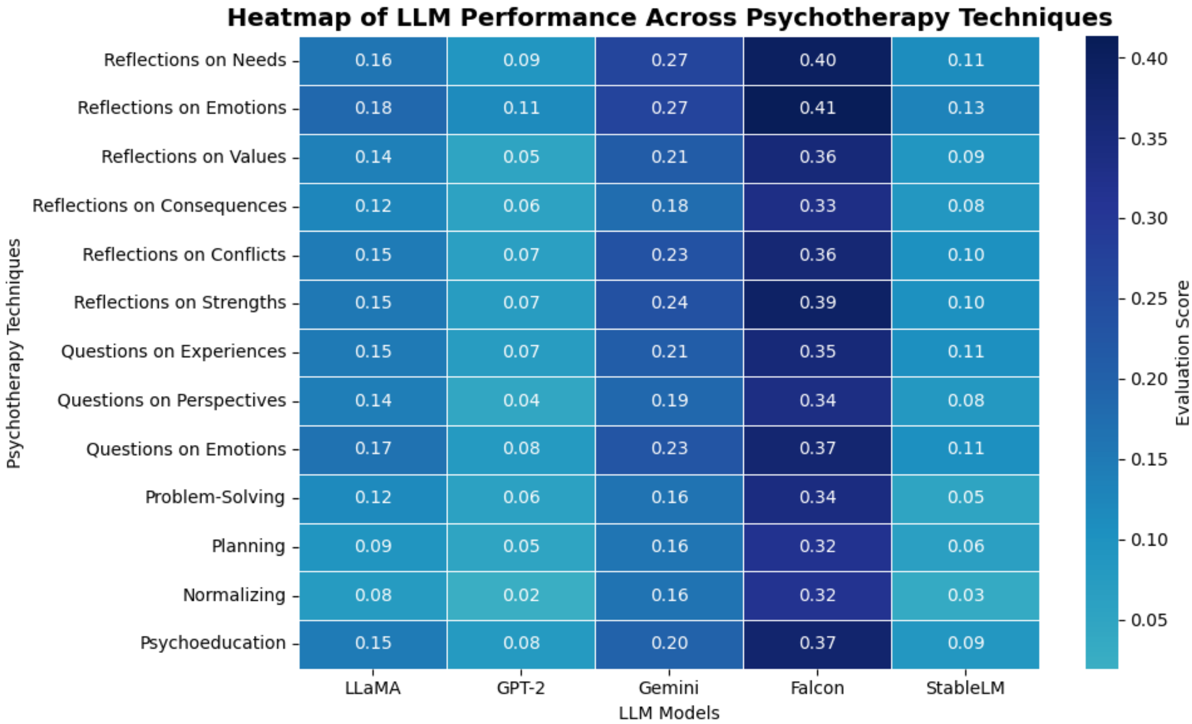

Figure 5.2: Comparative performance of LLMs across all psychotherapy techniques

Tabular Summary of Average Scores:

The overall average scores across all techniques are summarized below:

Table 5.1: Composite Average Scores of LLM Models

Model	Average Composite Score
Falcon	0.36
Gemini	0.21
LLaMA	0.14
StableLM	0.09
GPT-2	0.07

Table 5.1 provides a consolidated view of the average composite scores obtained by each LLM model across all psychotherapy techniques. The Falcon model achieved the highest average score, indicating comparatively stronger

alignment with therapeutic conversation standards. Gemini and LLaMA followed with moderate performance, while StableLM and GPT-2 scored the lowest, reflecting weaker adherence to the desired therapeutic response qualities. This summary highlights the variability in how effectively each model engages in psychologically supportive dialogue.

SBERT-Based Model Similarity Visualization

To better understand how semantically aligned each model's responses were to ideal therapeutic replies, Sentence-BERT (SBERT) was used to generate embeddings for all model outputs. SBERT is a variation of BERT optimized for capturing semantic similarity between sentences. By encoding both the model-generated and reference therapeutic responses, cosine similarity scores were computed to evaluate how closely each model approximated ideal conversational behavior.

To visualize behavioral similarities between models, Principal Component Analysis (PCA) was applied to the SBERT embeddings. PCA reduces the high-dimensional sentence embeddings into two dimensions, preserving the most significant variance:

- **PCA1 (x-axis)** represents the primary dimension of variation between model behaviors.
- **PCA2 (y-axis)** represents the second largest orthogonal variance component.

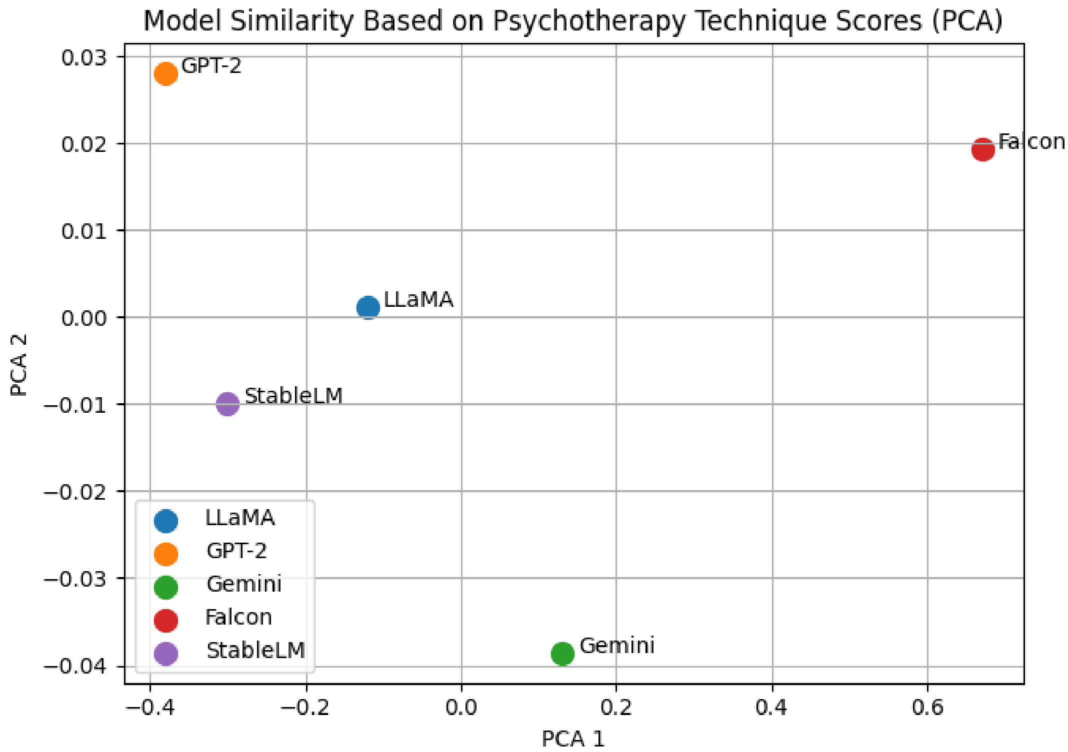

Figure 5.3: Average SBERT Similarity Scores Across All Models and Techniques

The plot shown in figure 5.3, gives the approximate spatial positioning of each model based on PCA-transformed SBERT scores:

- **Falcon** $(0.65, 0.02)$ lies furthest right along PCA1, suggesting the strongest semantic alignment with therapeutic reference responses.
- **Gemini** $(0.15, -0.04)$ also displays good alignment, especially in emotionally reflective responses.
- **LLaMA** $(-0.1, 0.0)$ remains near the center, implying moderate alignment with therapeutic standards.
- **StableLM** $(-0.3, -0.01)$ and **GPT-2** $(-0.04, 0.03)$ are clustered further left, indicating less therapeutic relevance and more generalized, non-specific language use.

This visualization complements the results of table 5.1, reinforcing that Falcon and Gemini offer more semantically accurate and therapeutically appropriate outputs compared to GPT-2, LLaMA and StableLM.

5.1.4 Observations from Behavioral Analysis

- **Falcon** consistently outperformed other models across all 13 psychotherapy techniques as mentioned in table 5.1, it has the highest average score and particularly excelling in reflective and emotionally attuned responses. Its strong scores in techniques like Reflections on Needs, Emotions, and Strengths highlight its alignment with therapeutic conversation norms.

- **Gemini** also showed strong performance, ranking second overall as shown in table 5.1. It demonstrated consistent empathetic engagement, especially in reflective and emotion-oriented techniques, although it was slightly less nuanced than Falcon in delivering depth.

- **LLaMA** showed moderate performance with a noticeable gap from the top two models. While it exhibited occasional alignment in reflective and psychoeducational techniques, it often lacked sustained depth across more complex therapeutic categories.

- **StableLM** and **GPT-2** consistently scored the lowest, suggesting a limited capacity for handling emotionally nuanced or client-centered dialogue. Their responses were generally surface-level, with minimal reflection or empathy.

- The results demonstrate that models like Falcon and Gemini, which either possess larger parameter sizes or benefit from fine-tuning and advanced instruction-following capabilities, are significantly better aligned with therapeutic communication compared to smaller, less specialized models like GPT-2 and StableLM.

These findings support the hypothesis that models equipped with more advanced architectures or fine-tuned on instruction-following data produce higher-quality, empathetic, and therapeutically-aligned conversations. The performance gap suggests that future therapeutic AI development should prioritize both model scale and domain-specific fine-tuning.

5.2 Validation with Test Cases

To complement automated evaluation, we conducted a human-centric validation using peer feedback collected via a Google Form. The form presented 10 diverse prompts, each with 5 model-generated responses from Falcon, LLaMA, GPT-2, Gemini, and StableLM. Over 30 respondents rated each response on a scale from 1 (Poor) to 5 (Excellent), evaluating the quality, empathy, and relevance of the replies. Figure 5.4 shows a graphical representation of the ratings given to different models by the responders.

Peer Evaluation Methodology:

- **Number of Respondents:** 30+
- **Rating Scale:** 1 to 5 (Low to High)
- **Prompts Evaluated:** 10
- **LLMs Evaluated:** Falcon, Gemini, LLaMA, StableLM, GPT-2

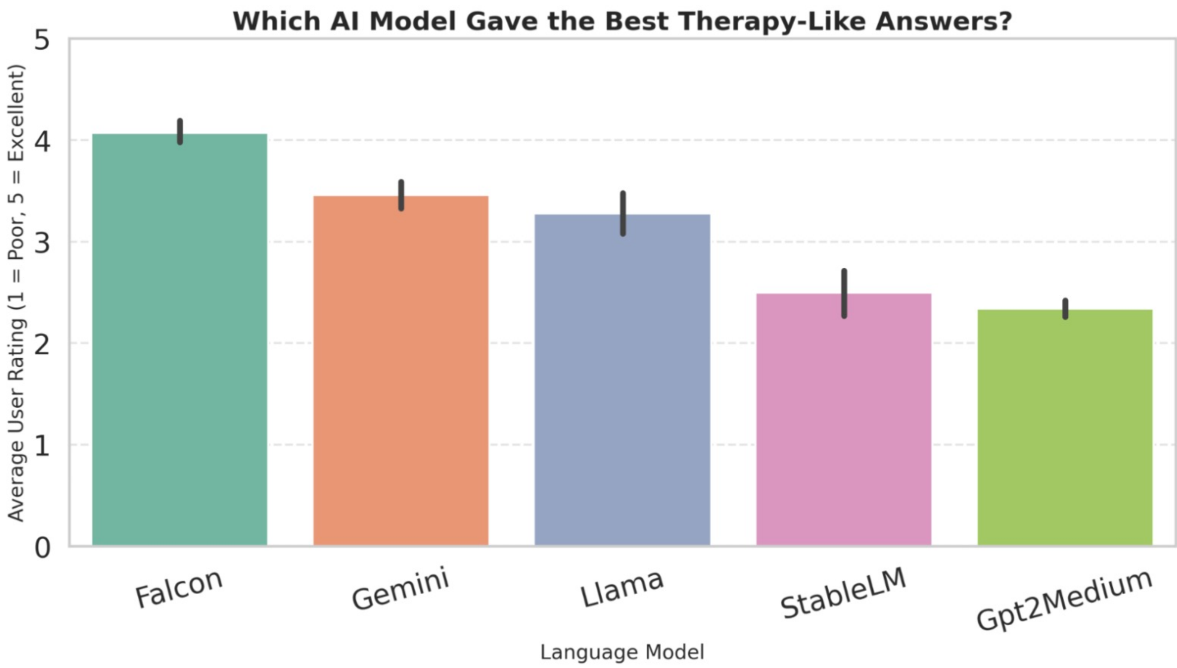

Figure 5.4: Average User Ratings per Model from Google Form Peer Feedback

Key Insights from User Feedback:

- **1. Top Performing Models:**
  Gemini and **Falcon** consistently received the highest average ratings across prompts. Their responses were found to be contextually accurate, emotionally aligned, and more human-like. They also showed low variance, indicating reliable behavior across diverse situations.
- **2. Mid-Tier Performers:**
  LLaMA and **StableLM** received moderate ratings. Their performance was prompt-dependent — excelling in structured queries but struggling with open-ended or emotionally sensitive prompts. Higher variance suggests inconsistency.
- **3. Underperforming Model:**
  GPT-2 had the lowest average score with the highest variance. Users described its responses as generic, occasionally irrelevant, and emotionally flat — indicating limitations due to lack of fine-tuning and outdated architecture.
- **4. Distribution Patterns:**
  Gemini and Falcon responses clustered in the 4–5 range with minimal outliers, showing general approval. In contrast, GPT-2 responses had a wider distribution, including multiple low ratings, and a large interquartile range, showing unpredictability.
- **5. Prompt Sensitivity:**
  Certain prompts (e.g., involving ethical dilemmas or deep grief) yielded lower ratings across all models. This highlights a shared limitation in addressing nuanced emotional complexity — an area requiring further model refinement.

This peer-based validation complements metric-driven analysis by offering a real-world perspective on model effectiveness. High-rated models such as Falcon and Gemini demonstrate stronger alignment with human expectations, suggesting their viability in therapeutic-like settings when guided by proper oversight.

5.3 Future Work

Building on the findings of this study, several future directions are identified to enhance the impact, scope, and real-world applicability of LLMs in mental health support. These include technical expansion, ethical reinforcement, domain-specific collaboration, and global adaptability.

1. Broader Model Inclusion

Future iterations of this project will evaluate newer and more advanced Large Language Models such as **Claude**, **GPT-4**, and **Mistral**. These models offer state-of-the-art improvements in contextual understanding, emotional intelligence, and response coherence. By comparing their therapeutic behaviors with existing models, we aim to assess whether newer architectures can better replicate the sensitivity and empathy required in therapeutic interactions.

2. Ethical Safeguards and Crisis Detection

To ensure responsible deployment, ethical safeguards must be embedded into future versions of the system. This includes developing mechanisms to **detect and respond to high-risk prompts** involving self-harm, abuse, or suicidal ideation. Future work will integrate real-time **crisis deflection mechanisms**, such as emergency resources, disclaimers, or escalation to human support channels. These features are crucial for ensuring that AI systems are safe, supportive, and not harmful in emotionally sensitive situations.

3. Collaboration with Mental Health Professionals

To validate and refine the therapeutic accuracy of model responses, the next phase of research will involve close collaboration with licensed psychologists, therapists, and clinical counselors. Their expertise will be instrumental in:

- Designing domain-relevant evaluation rubrics
- Conducting manual audits of model output
- Guiding safe and ethical interaction strategies

Such partnerships will ground the project in real-world therapy practices, ensuring that AI behavior aligns with professional standards.

4. Cultural and Linguistic Diversity

Therapeutic communication varies greatly across languages and cultures. Future research will include **multilingual and multicultural datasets** to test LLM performance in diverse contexts. The aim is to assess and improve **cross-cultural empathy, inclusivity**, and relevance of responses for users from different cultural backgrounds. This work is critical for building equitable and globally applicable AI systems in mental health.

5. Prototype Chatbot Development

A real-time therapeutic chatbot will be developed using top-performing models (e.g., Falcon, Gemini). The chatbot will feature:

- Contextual memory for continuity in sessions
- Safety protocols for handling emotional distress
- Referral pathways to mental health professionals

The goal is to build a supportive AI system that can assist users with low-intensity mental health concerns, while **clearly maintaining the distinction between AI assistance and licensed therapeutic care**.

These proposed extensions aim to make LLM-based mental health tools more robust, ethical, inclusive, and clinically meaningful — bridging the gap between AI capabilities and human-centered therapeutic care.

Chapter 6

Conclusion

In this project, titled **“Machine Minds: Behavioral Assessment of LLM Therapists,”** an extensive evaluation of various Large Language Models (LLMs) was conducted to assess their effectiveness in providing therapeutic support. The models were analyzed based on their performance across thirteen core psychotherapy techniques, including reflections, questioning, planning, normalizing, psychoeducation, and problem-solving skills.

The evaluation revealed significant variability among the models. Some models exhibited strong performance in specific techniques while underperforming in others. Upon detailed analysis of the response evaluation graphs, the following observations were made:

- **Falcon:** Emerged as the most **highly recommended model** for mental health counseling. It consistently achieved the **highest positive scores** across a broad range of psychotherapy techniques. Falcon demonstrated a well-balanced approach, showing strong capabilities in reflection, emotional support, psychoeducation, and problem-solving, while maintaining very low negative evaluations. Its ability to offer human-like, empathetic, and contextually appropriate therapeutic responses makes it the most suitable model for counseling applications.
- **Gemini:** Identified as the **second most recommended model**. It performed exceptionally well in techniques such as reflections on needs and emotions, psychoeducation, and planning. Although it showed slightly

lesser consistency in open-ended questioning compared to Falcon, Gemini's overall alignment with human therapeutic practices remains strong, making it a reliable secondary option.

In contrast, models like **LLaMA**, **StableLM**, and **GPT-2** demonstrated weaker therapeutic behavior. LLaMA often remained neutral without sufficiently engaging with client emotions, StableLM leaned excessively towards directive responses without promoting exploration, and GPT-2 struggled with consistency and depth across various techniques.

Thus, based on the behavioral assessment conducted, **Falcon is recommended as the primary model for therapeutic support applications**, followed by **Gemini as a secondary choice**. These findings highlight the importance of carefully selecting and evaluating LLMs for sensitive domains like mental health, ensuring that the deployed models align closely with established therapeutic communication standards.

Appendix-I

A. Sample Prompts and Model Responses

Client Prompt	Model Response (Falcon-1B)
”I feel like anxiety is taking over my life.”	”It sounds like you’re feeling overwhelmed. Anxiety can be intense, but there are ways to manage it. Would you like to explore some coping strategies?”
”I don’t think I can ever be happy again.”	”Feeling hopeless can be painful. But know that emotions do change, and with the right support, many people find joy again.”

B. Evaluation Metrics Used

1. SBERT Semantic Similarity

Semantic similarity was calculated using Sentence-BERT embeddings. The formula for cosine similarity is:

$$Similarity = \frac{A \cdot B}{\|A\|\|B\|}$$

where A and B are the embedding vectors for the ground-truth and model-generated responses.

2. VADER Sentiment Analysis

The VADER sentiment analyzer provides a compound score that ranges from -1 (most negative) to +1 (most positive). It was used to quantify the emotional tone of each response.

- Positive Score
- Negative Score
- Compound Score (Overall sentiment)

C. Model Configuration Summary

Model	Version	Fine-Tuned	Quantized	Parameter Size
LLaMA	3.2-1B	Yes	No	1.1B
GPT-2 Medium	OpenAI	Yes	No	345M
StableLM	3B	No	Yes (4-bit)	3B
Falcon	1B	Yes	Yes (8-bit)	1B
Gemini	2.0 (API)	No	No	N/A

D. Technical Tools and Libraries

- **Language:** Python 3.10
- **Libraries Used:** transformers, datasets, scikit-learn, sentence-transformers, vaderSentiment, matplotlib, seaborn
- **Training Platforms:**
  - MacBook Air M2 (8GB RAM) – local training
  - Google Colab (Free Tier) – cloud-based execution

E. Example Code Snippets

SBERT Similarity Code

```
from sentence_transformers import SentenceTransformer, util

model = SentenceTransformer('all-MiniLM-L6-v2')
score = util.cos_sim(model.encode(gt_response), model.encode(
    model_response))
```

VADER Sentiment Analysis

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()
sentiment = analyzer.polarity_scores(response)['compound']
```

F. Human Evaluation Tool

Human validation was conducted using a Google Form, where participants rated model responses based on:

- Relevance
- Empathy
- Therapeutic helpfulness

References

[1] Chiu, Y. Y., Sharma, A., Lin, I. W., and Althoff, T. (2024). 'A Computational Framework for Behavioral Assessment of LLM Therapists'. *arXiv (Cornell University)* [Preprint]. doi:10.48550/arxiv.2401.00820.

[2] Demszky, D. et al. (2023). 'Using large language models in psychology'. *Nature Reviews Psychology*. Available at: Semantic Scholar (Accessed: 19 October 2024).

[3] Chung, N. C., Dyer, G., and Brocki, L. (2023). 'Challenges of Large Language Models for Mental Health Counseling'. *arXiv.org* [Preprint]. Available at: `https://arxiv.org`.

[4] De Choudhury, M., Pendse, S. R., and Kumar, N. (2023). 'Benefits and Harms of Large Language Models in Digital Mental Health'. *arXiv.org* [Preprint]. Available at: `https://arxiv.org`.

[5] Singh, U. and Aarabhi, P. (2023). "Can AI have a personality?", *Proc. IEEE CAI*, vol. 2, pp. 205–206. doi:10.1109/cai54212.2023.00097.

[6] World Health Organization. (2022). *Mental Health Atlas 2020*. Available at: `https://www.who.int/publications/i/item/9789240036703`.

[7] National Institute of Mental Health and Neurosciences (NIMHANS). (2016). *National Mental Health Survey of India, 2015–16*. Bengaluru: NIMHANS Publication.

[8] Meta. (2024). "meta-llama/Llama-3.2-1B," *Hugging Face*. Available at: `https://huggingface.co/meta-llama/Llama-3.2-1B`.

[9] TII UAE. (2025). "tiiuae/Falcon3-7B-Instruct," *Hugging Face*. Available at: https://huggingface.co/tiiuae/Falcon3-7B-Instruct.

[10] Stability AI. (2023). "stabilityai/stablelm-3b-4e1t," *Hugging Face*. Available at: https://huggingface.co/stabilityai/stablelm-3b-4e1t.

[11] OpenAI Community. (2023). "openai-community/gpt2-medium," *Hugging Face*. Available at: https://huggingface.co/openai-community/gpt2-medium.

[12] Google AI. (2024). "Google AI Studio." Available at: https://aistudio.google.com.

[13] Shaik, Z. H. (2024). "Patient-Therapist Dataset," *Kaggle*. Available at: https://www.kaggle.com/datasets/zuhairhasanshaik/datacsv.

[14] Brown, T. *et al.* (2020). "Language Models are Few-Shot Learners." *NeurIPS 2020*. Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[15] Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020). "A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support." *arXiv.org* [Preprint]. Available at: https://arxiv.org/abs/2009.08441.

[16] Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., and Althoff, T. (2024). "Facilitating Self-Guided Mental Health Interventions Through Human-Language Model Interaction: A Case Study of Cognitive Restructuring." *Proceedings of CHI 2024*. doi: https://doi.org/10.1145/3613904.3642761.

[17] Shah, J., Kamble, S., Nahak, A., Kalgutkar, H., and Karunakaran, K. (2024). "Machine Minds: Behavioral Assessment of LLM Therapists."

Journal of Technology, vol. 12, issue 12, Paper 63. Available at: https://journaloftechnology.org/volume-12-issue-12-2024/.

Acknowledgements

As students of Group A-1, BE CMPN A, we are grateful for this educational opportunity and for our teamwork in creating this project. Ms. K. Priya Karunakaran guided our project. She provided invaluable advice and feedback for our project, for which we are grateful. Her expertise and recommendations aided in the project's effective development and workflow planning. We also thank our college for letting us collaborate with them and for giving us the greatest resources and assistance we needed for the project. We also want to express our gratitude to the head of the computer department, Dr. Kavita Sonawane, as well as to all of the lab assistants and other faculty members who helped us along the way as we developed this project.

Sincerely,

Jheel Shah

Shashank Kamble

Anushree Nahak

Harsh Kalgutkar