Utilizing Counselor-Client Dialogues to Develop a Memory-Efficient Mental Health Question-Answering System with Large Language Models

MSc Research Project
Masters of Science in Artificial Intelligence

Diya Srivastava
Student ID: x23177608

School of Computing
National College of Ireland

Supervisor: Arundev Vamadevan

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Diya Srivastava
Student ID:	x23177608
Programme:	Masters of Science in Artificial Intelligence
Year:	2024-2025
Module:	MSc Research Project
Supervisor:	Arundev Vamadevan
Submission Due Date:	12/12/2024
Project Title:	Utilizing Counselor-Client Dialogues to Develop a Memory-Efficient Mental Health Question-Answering System with Large Language Models
Word Count:	XXX
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Diya Srivastava
Date:	24th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Utilizing Counselor-Client Dialogues to Develop a Memory-Efficient Mental Health Question-Answering System with Large Language Models

Diya Srivastava
x23177608

Abstract

Healthy living for an individual most certainly encompasses Mental well-being enabling humans to endure emotions. A healthy state of mind is premiere to a healthy life, yet nearly millions of individuals globally suffer from mental disorders such as anxiety, depression, and PTSD, with the COVID-19 pandemic further exacerbating this crisis. Traditional therapy is often challenging due to scarcity of psychologists, expensive sessions or apprehensions associated with people belonging to different demographics. Hence, in response this research explores the advancements in Natural Language Generation(NLG) domain of Artificial Intelligence(AI) to conduct Virtual Therapy. The study proposes the use of sequence-to-sequence large Language Models built on decoder-transformer architecture, leveraging Parameter efficient Fine-tuning technique like LoRa and Memory Efficient Quantization strategy to develop a mental health domain specific question-answering system in a resource constraint environment. The study experiments with Flan T-5-Base, Tiny Llama-1.1B, Llama-2 7B, Gemmma-2 2B and GPT-Neo 2.7B to prospect their performance after being fine-tuned on 'MentalChat16k' dataset of question-answer pair from a therapist, client conversation. The study evaluates model generated outputs qualitatively, while Conducting quantitative analysis on diverse LLMs by computing BLEU, BERT and ROUGE Score. Concluding with Gemma-2 achieving 0.5 ROUGE-1 score outperforming other models, while Llama-2 prevails in delivering more empathetic and coherent responses.
Keywords: Cognitive Behavioral Therapy, PEFT, Quantization, Mental Health Support.

1 Introduction

Mental Health is a state of Well-being enabling an individual to experience diverse emotions, cope with stress, thibk clearly and make informative decisions, be productive and effectively contribute to their community. The state of Mental well-being is a complex continuum and an inevitably crucial part of a healthy life style. According to the WHO, around 970 million people world-wide were living with mental disorders World Health Organization (2023) in 2019 encompassing anxiety, depression, PSTD i.e. Post Traumatic Depression or Eating disorders. Among these, 300 million were diagnosed with anxiety disorder, 58 million of them being children and adolescents. The following year, Covid-19 pandemic caused a 25% increase, reaching to 32% of the total population globally suffering from mental health conditions, as reported by AXA AXA (2024). the surge in mental health issues during and after pandemic, often referred to as 'Long Covid' was scrutinized and published in The BMC Psychiatry in February 2024. Additionally, research published in the Lancet in 2021 Seighali et al. (2024) enunciated an extra 54 million cases of major depressive disorder and 76 million cases of anxiety disorders world-wide in 2020 due to the pandemic.

Hence, these meta analysis and statistics serve the pressing necessity of a comprehensive intervention to address psychological disorders. Putting an emphasis on accessible, cost effective and sustainable solution. This need is particularly critical in densely populated nations like India, China, USA, where a Mental Health or Emotional Support System is a necessity.

Recent advancements in the domain of Natural language Processing (NLP) of Artificial intelligence (AI) have led to the evolution of Language Models that unlock a wealth of opportunities through Text analysis, Intent Recognition, Context Management and Text Generation. Models like OpenAI's Chat-GPT, Google's BERT have embarked remarkable standards of tasks in Natural Language Understanding (NLU) and Natural Language Generation (NLG). The development of pre-trained models like these has profoundly influenced fields of Finance (Shah et al.,2023) Hiraoka et al. (2022) and Medicine (Madani et al., 2023) Madani et al. (2023). Amid these developments, the integration of AI and mental health has emerged as a promising frontier. Notable breakthrough researches in the Mental Health, Psychology domain, aimed to develop AI-based Mental/Emotional Support systems leveraging Large Language Models (LLM) in their respective study, Lai et al. (n.d.), (Hongbin, 2023) Na (n.d.).

Large Language Models (LLMs) represent Deep Learning application in a Transformer Network. These Neural Networks learn context and meaning, by analyzing relationship in sequential input data, enabling then to perform tasks like Recognition, summarizing, Translation, Prediction and Text Generation. This research delves into the coalition of LLMs and Mental health Therapy to develop a Virtual Mental Health Support system. One of 'Golden Therapy' intervention David et al. (2022), Cognitive Behavioral Therapy or CBT on a large horizon is a talking Therapy that helps and individual to refine thoughts and hence give them a better perspective, simply through talking and dividing problems into smaller parts is commonly used to treat anxiety, depression.

Extensive research has focused on developing mental health support systems driven by LLMs. Recent advancements have seen the creation of fine-tuned task specific models have on substantial therapy dataset. For instance, Psy-LLM by Lai et al. (n.d.), CBT-LLM by (Hongbin,2023) Na (n.d.) have leveraged LLMs to develop a Mental Health question answering systems in Chinese language. However, full Instruction-tuning of large language models is computationally intensive and impractical in a resource constrained environment. Moreover, pioneering studies like Na (n.d.) introduced a Chinese CBT-LLM are limited by language constrains, and do not suffice the requirement for an english virtual therapy.

Motivated by these researches and need for Mental Health support in today's high-pressure society, where mental health issues are wide spread, relying solely on limited number of psychologists is challenging. Additionally, exploring the potential of Transformer based language models in resource constrained environment via Parameter Efficient Fine-Tuning (PEFT) techniques like LoRa (Low-Rank Adaptation) and Memory efficient 'Quantization' strategy play a very essential role in leveraging Language Models

to develop AI-driven Mental Health support system. Moreover, influenced by traditional cultures, and apprehensions associated people belonging to different demographics, individuals often restrain to adjure for help. Thus, this research proposes a virtual Mental-Health Therapist, that eliminates apprehension and unease associated with human interaction.
Research Question: How can Large Language Models such as Flan-T5, Tiny-Llama, Llama-2, Gemma-2 and GPT-Neox, Fine-Tuned on therapy questionnaire dataset 'MentalChat16k' be utilized to develop a Virtual Therapy Question Answering System and how effective is the system in delivering stigma-free mental health support as evaluated by metrics like BLEU, ROUGE, and BERTScore?

2 Related Work

There have been several research in the field of Mental Health and its amalgamation with language models, demonstrating textual, audio and other health parameters utilized to develop Emotion Classification and Evaluation System, Mental Health Chat-bots and even for explicit tasks of mental health therapy like Cognitive Behavioral Therapy (CBT) which as stated by David et al. (2022) is an infamous gold-standard psychotherapy leveraged to cure a number of mental disorders like anxiety, depression, bipolar, etc.

2.1 AI-driven Psychological Therapy: Leveraging Large Language Models for Metal Health Systems

In the realm of AI-driven LLMs for Mental Health applications, a substantial body of researches explore the potential of LLMs. One such inspiring work by Hongbin (2023) Na (n.d.) explores the capabilities of different language models to develop a CBT-based question-answering system. The author designed CBT specific prompts instilled in a CBT-based Q/A dataset tailored for Chinese Mental Health dialogues utilizing PsyQA dataset Sun et al. (2021), crafted by exploiting GPT-3.5 Turbo 16k model. Post dataset engineering, Hongbin probed the efficacy of LLMs including LlaMA-Chinese-7b, Alpaca-Chinese-7b, Qwen-7b, Biachuan-7b employed with Parameter Efficient Fine-Tuning technique (PEFT) technique i.e. LoRa for cost effective and minimal parameter tuning on the refined dataset, to evaluate using automatic evaluation metrics like BLEU, BERTSCORE, METEOR, CHRF, BLEURT accompanied with Human Evaluation metrics like Relevance, CBT Structure Measure and Helpfulness. Concluding with Baichuan-7b marginally outperforming Alpaca-7b in all aspects, particularly in adhering to CBT frameworks and providing helpful responses. While, the study limits its application to only Chinese and relying entirely on generative models to encapsulate the comprehensive process of CBT-responses in single dialogues may inadvertently create a sense of pressure for the users and entail societal bias. Similar to Hongbin's research David et al. (2022) in a recent study i.e. CBT-Llama, explores the latent ability of LLMs like Llama-3 (trained on 8b parameters) fine-tuned on mental health conversational dataset extensively formulated with the help of Claude-Haiku. The study employed Claude for synthetic dataset generation leveraging emotional and demographic seeds encompassing attributes: 'Sex', 'Age', 'Occupation', 'Relationship Status', 'Negative Emotions', wherein each of them are paired with a negative emotion (i.e. disgust, fear, hopelessness, guilt, envy, hate, etc.) to generate realistic scenarios. The study also assess Quantized Low-Rank Adaptation (Q-LoRa) technique to fine-tune the transformer model while optimizing computational cost and thereby generating competitive results. Although, the doesn't talk about evaluation of the trained model as per automatic evaluation metrics. Additionally only training one model i.e. Llama-3, the study lacks to compare and evaluate its performance with any other Language Model in the LLM realm. Simultaneously, a completely synthetic dataset formulation posts a threat for bias, unrealistic assumptions and degraded performance on realistic data/conversations leading to suboptimal results. Another study, bridging the gap for Real Feedback from Humans, Lai et al. (n.d.) leveraged the caliber of LLMs to generate Psy-LLM, a language model built upon large scale pre-trained corpus models, specifically PanGu by Huawei and WenZhong model based on GPT-architecture by Idea Research Institute. for the dataset, the researchers combined the PsyQA dataset Sun et al. (2021) and a substantial number of Chinese psychological articles. As a part of their comprehensive human evaluation, the study explores real feedback from people by deploying the LLMs on a website and later using the rating to re-fine-tune the model. While, Xu et al. (n.d.) Xu et al. in their pioneer experiment of utilizing a melange of different LLMs leveraged Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, and GPT-4 for tasks encompassing: Zero-shot and few-shot prompting, along with Instruction-tuning on different LLMs on considerable mental health datasets like (Dreddit, DepSeverity, SDCNL, etc.) to perform predictive tasks based on scenarios to apprehend user's condition. The key takeaways from this study highlights the preeminence of smaller language models like Flan T-5 and Alpaca surpassing GPT-3.5, GPT-4 after being trained on diverse dataset enabling to achieve more generalized results. In their research of adapting LLMs towards Emotional Support Systems, Liu et al. (2021) proposed an Emotional Support Conversation (ESC) framework that's based on Helping Skills Theory accompanied with rich annotations in a 'Help-Seeker' and 'Supporter' mode amid ESC dataset construction. This study stands out for capitalizing on LLMs to provide support through social interactions rather than professional counseling grounded on Hill's helping theory stages i.e. (Exploration, Comforting and Action). For implementation, the researchers employed two state-of-the-art pre-trained models i.e. 'Blender-Bot', an open-domain language model trained with conversational skills optimized for empathetic responding. 'Dialo-GPT' is the other LLM used that is a GPT-2 based model trained on large scale corpora. These models were trained in three variants: 'Vanilla' , 'Joint' , 'Oracle' with different responses, annotation and reference strategies. While the study evaluated on automatic evaluation metrics, it emphasized on Human Interactive Evaluation Metrics like: 'Fluency', 'Identification', 'Comforting', 'Suggestion'. In conclusion, the blender-bot 'Joint' variation model outperformed all competitors on all metrics. The research focuses on modest level emotional support explicitly divulging from professional peer support. While the system performs to alleviate negative emotions it doesn't resolve the questions regarding the extents to which dialog system provides emotional support. Additionally, the research completely relies on the Hill's strategy while the real application may need more refinement and the evaluation metrics leveraged here may not fully capture the nuanced and subjective nature of effective emotional support, potentially leading to gaps in assessing real-world performance.

2.2 Dataset Curation: Towards Mental Health Conversation Annotation and Analysis

Across multiple studies including the ones by Na (n.d.) and David et al. (2022), there is a recurring emphasis on dataset formation and annotation of conversations with detecting emotions accompanied with cognitive distortion analysis. In a similar study focusing on exploiting Deep learning for language understanding of mental health concepts Rojas-Barahona et al. (n.d.) present their ontology demonstrated from Cognitive Behavioral Therapy (CBT) by leveraging deep learning models to label mental health associated corpus exhibiting psychological phenomenon. This study works with the exceptional corpus curated from 500k anonymous posts from Koko platform i.e.(based on peer-to-peer therapy) introduced by Morris et al. (2015), employing two therapists for annotation. The study utilized the superiority of Convolutional Neural Network (CNN) inspired by Kim (2014) operating over pre-trained 'GloVe' embeddings, over Recurrent Neural Network (RNN) since the posts are generally too long for an RNN model to maintain memory over words. CNN-GloVe models performed the best especially in capturing the sentiments of words with both 100d and 300d embeddings, indicating stability across dimensions and effeciency in capturing psychological concepts, whereas Sentence-level embeddings (Skip-thought) required higher dimensionality but were useful for capturing nuanced sentence-level meanings. The takeaway from this study is the use of an exceptional dataset integrated with psychological expertise rather than relying on pre-trained transformer models like David et al. (2022) and other studies, making the research more realistic and scalable. In another work by Sharma et al. (n.d.), the researchers investigated the performance of language models to Re-frame Negative thoughts as therapy. Similar to the findings of Rojas-Barahona et al. (n.d.), this study experimented labeled corpus curation and development of automatic evaluation metrics with 600 situations accompanied by psychologist's expertise in Cognitive Re-framing as per 7 linguistic attributes: whether the thought is rational, positive, empathic, actionable, etc. Thought Record Dataset Burger et al. (2021) and Mental health America(MHA) serve as dataset for this work. The research leverages GPT-3 to train a retrieval-enhanced in-context learning model. For development of evaluation metrics, the authors have demonstrated metrics like 'Rationality', 'Positivity', 'Empathy', etc. For positive re-framing, the researchers used (1) DialoGPT GPT-2 Zhang et al. (2020), (1) BART based Positive re-framing model Ziems et al. (2022), (3) Flan T-5 and (4) GPT-3. For extended furtherance, the work was deployed to examine "People's preference of linguistic attributes of re-framed thoughts" and "Relation of those attributes with desired cognitive re-framing outcomes". Finally, the model was deployed on the MHA website with a random field study of 2067 participants to conclude that people struggling from negative thoughts highly emphasis on 'Empathy' and 'Specificity' but don't prefer re-frames that are highly positive. The respective study motivates the use of LLMs and development of automatic evaluation metrics in mental health domain, while the dataset is not publicly available and did not address the impact of socio-cultural factors on cognitive reframing. Moreover, the study focused primarily on creating effective in-the-moment interventions and did not explore short-term or long-term clinical outcomes. A study Malhotra et al. (n.d.) with the objective of developing a comprehensive dataset aggregated from open-source therapist-patient conversation with the objective of 'Dialogue Act Classification' in counseling conversations. The researchers compiled 12.9k utterances of conversations engineered from publicly available YouTube counseling sessions to generate state-of-the-art HOPE dataset, simultaneously performed empirical and quantitative analysis on the same SPARTA. The study therefore, proposes 12 dialogue-act annotations for the development of a Classification System that combines speaker-dynamics and local context through a time-aware attention mechanism. The study motivates the utilization of datasets like these for further tasks in mental health domain prominently Question-Answering system or Chatbots, but this study couldn't foster such benefits from the HOPE dataset since for QnA system the corpus accounted for only 6.4 k utterances that couldn't suffice LLMs training. The limitation of the study is highlighted in the information loss due to automatic transfer of utterance from the speech modality to the text and incorporated bias due to data formulated from online counseling sessions in US.

2.3 Standalone Researches: Mental Health Condition Analysis and Evaluation

Aiming on "Predictive Analysis on Mental Health on Leveraging LLM Embeddings and Machine Learning Models for Social Media Analysis" Radwan et al. (2024) in their work proposed a methodology to use LLMs with emphasis on GPT-3 embeddings and Machine Learning algorithms to classify 'Social Media' posts as indicative of stress disorder. The study leveraged models like GPT-3, BERT, Metapath2Vec, mBERT for embeddings, vector representations accompanied with ML algorithms including support vector machines, random forests, XGBoost, KNN, and neural networks trained on a substantial dataset with 10000 labeled Reddit posts wherein SVM and XGBoost prevailed in accuracies for classification. The study inspires the amalgamation of LLMs and ML algorithms for dataset curation and classification tasks. A diverse research study "Integrating BERT with CNN and Bi-LSTM" for explainable depression detection by Xin and Zakaria (2024). The study endeavors to explore integrated performance of transformer-based models and hybrid architectures to enhance depression detection from a corpus of social media posts. The researchers proposed three BERT-based approaches—fine-tuned BERT, BERT-BiLSTM, and BERT-CNN—and evaluated them against MentalBERT, a state-of-the-art mental health-focused model, using datasets from Reddit, Twitter, and mental health corpora. While achieving a superior accuracy of 98.2% with BERT-CNN. The amalgamation of BERT and CNN, LSTM not only proposes improved accuracy, but also ensures explain-ability in the 'black box' nature of deep learning models. This study fosters advance explain-able AI tools for early intervention in mental health domain by integrating cutting-edge NLP models with transparent interpretability techniques. Another imperative approach introducing PaperQA Lála et al. (n.d.), employed Retrieval-Augmented Generation (RAG) on science QnA System proved to outperform existing LLMs at the task of answering scientific questions. The study promotes the use of RAGs systems to circumvent 'Hallucination' in training LLMs. With the combinations of Claude-2, GPT-3.5, and GPT-4 providing strong results on LitQA (benchmark of 50 questions). The research highlights the ability of a RAGs system to dynamically retrieve full-text papers and and iterate over them for answers without hallucinating. MindGuard Ji et al. (2024) marks a pioneering work introducing a mobile mental health first aid equipped with a LLM containing extensive mental health domain knowledge. The research is a standalone and one of its kind embarked by its objective to capture data from mobile sensors with the superior conversational and analytical capabilities of the LLM to facilitate accessible and stigma-free mental health support, encompassing mental health diagnosis, continuous monitoring, and personalized intervention. The study employed

GPT-4o for question-answering tasks demonstrated from the personalized data curated via mobile sensors and evaluated against GPT-3.5o, Claude-3.5 and Llama-3, Mixtral, InternLM2, Qwen-2 on publicly available data where their model prevailed. Delving into "Deployment of LLMs with RAGs" Prabhune and Berndt (2024) in their study highlighted the key challenges and opportunities associated with implementing RAGs systems with deploying LLMs. Although RAGs are the emerging solutions to the 'Hallucination' challenge of LLMs, the findings cover the complexities and challenges with it's deployment. Moreover, the study discusses the AI governance framework and essential practices to be followed for successful deployment of RAGs, with introducing governance models for unique AI system demands for structured prompt engineering while acknowledging the ethical considerations, and ongoing system evaluation. A considerably new ontology proposed in the work "Virtual and Augmented Simulations in Mental Health" by Carlson (2023) explored the integration of new technologies i.e. Virtual Reality(VR) and Augmented Reality(AR) and LLMs in mental Health systems. The research discusses the application, implications and promising results of VR/AR, particularly Virtual Reality Exposure Therapy (VRET) in treating anxiety disorders and phobia. The study also highlights the integration of LLMs in AR/VR for real-time feedback and emotion detection, the potency of intelligent agents to be semantically reliable and adapt to diverse therapeutic scenarios. For a mental-health domain-specific NLP task, this coalition of VR/AR and LLM technologies provide scalable approach to deliver virtual mental health therapy unlike traditional methods.

2.4 Conclusion

Recent advancements in language models have contributed significantly to develop a mental health support system exhibited in diverse domain-specific tasks including emotion detection and classification, chat-bots and therapy tools like CBT-based response generation. Researches have also explored LLM functionality in resource-constrained environments, thereby implementing state-of-the-art techniques like LoRa and Quantization for parameter efficient fine-tuning(PEFT) of LLMs, like Baichun-7b, Llama-3 and GPT-3 in generating empathetic and coherent responses to user's questions. But the Chinese Mental Health assistant limits the scalability of such research due to language barrier. The integration of multi-modal approaches, encompassing VR/AR for real-time feedback, and hybrid architectures such as BERT-CNN for depression detection, highlights the potential for enhanced explainability and therapeutic applications. Not to forget, emerging AI frameworks like RAGs address hallucination issues whereas utilizing mobile sensor's data with LLMs add to personalization. Despite these innovations, ongoing research emphasizes the need for a more robust english dataset, ethical AI governance and real-world validation to ensure scalable, inclusive, and clinically effective mental health solutions.

3 Methodology

3.1 Research Methodology Workflow Overview

This study adopts a Secondary Mixed Empirical Deductive Research Methodology to develop and evaluate a cost efficient Mental-Health question-answering system using Large Language Models. This research involves Secondary approach, that leverages pre-existing open-source dataset constituting conversations between a therapist and a client Jia Xu

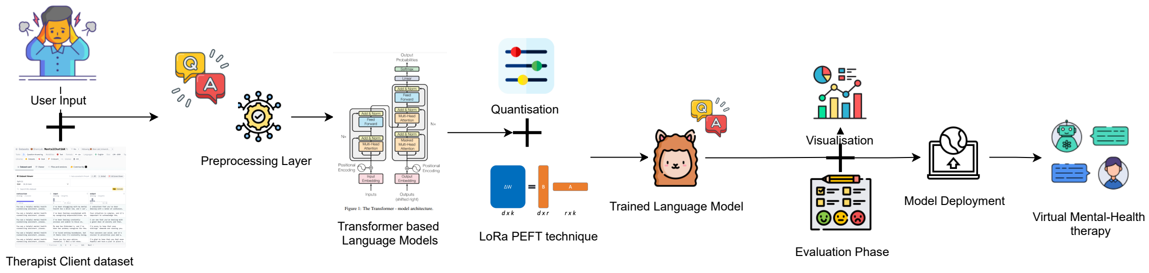

Figure 1: Research Architecture

(2024) available on HuggingFace. The work combines Empirical methods inspired from the pioneering research based on Instruction-Tuning of LLMs for mental health therapy like Na (n.d.), with Mixed-method strategy encompassing Qualitative evaluation demonstrated by the coherence and relevance of LLM's generative responses on therapeutic conversations evaluated by Quantitative analysis on the automatic evaluation metrics i.e. BLEU, ROUGE, BERTScore.

The deductive aspect of the study is highlighted in the Hypothesis driven delineation: Fine-tuning large language models with memory-efficient Quantization and parameter-efficient LoRa techniques without compromising the quality of generated responses. The coalescence of these methodologies foster a comprehensive framework for evaluating the potency and feasibility of LLMs in Mental Health domain.

3.2 Data Collection

At the onset of this research, HOPE dataset curated by Malhotra et al. (n.d.) from publicly available counseling sessions elucidated with dialogue-act annotations using LLMs, was the primary choice that constitute 12.9k utterances. While it provided a good starting point for the research, certain limitations including incoherent question-answer pairs. While the HOPE corpus was meaningful, it followed a continuum of conversations and not discrete advice/answers of questions. This structure made it difficult to isolate the answers for explicit questions that would suffice the requirement of standalone advice by therapist.

To address this issue, an attempt was made to augment the existing dataset into segments of question-answers. However, it faced the challenges of Incoherent QnA pair, where many consecutive head-to-head lacked clear boundaries, were fragmented and presented vague exchanges like greetings or patient's background. Additionally, incompatible dataset size was another challenge. After augmentation, the dataset reduced to 6.4k utterances of question-answer pairs, which wasn't sufficient to Instruction-tune LLMs. The limited data posted risks for overfitting and and restricted the model's ability to generalize across diverse mental health scenarios.

To overcome the challenges of HOPE dataset, this study adopted the 'MentalChat16k' dataset Jia Xu (2024) as primary corpus sourced from Huggingface repository (ShenLab/MentalChat16K). This publicly available dataset constitute of 16,084 rows with size of 46.4 MB counselor-client conversations, with columns named: 'Instruction', 'Input', 'Output'. The dataset is an amalgamation of 'Synthetic data' i.e. 9,775 synthetic conversations between counselor and a client and 'Interview data' i.e. 6,338 question-answer pairs from 378 interviews.

The 'Synthetic data' covers 33 mental health topics such as Relationships, Anxiety, De-

pression, Intimacy, and Family Conflict. This dataset is curated leveraging the OpenAI GPT-3.5 turbo model and a customized adaptation of the Airoboros self-generation framework. The Airoboros framework helped to create a discrete instruction based topic specified prompt mechanism to create patient's queries, which were then fed to GPT-3.5 turbo model to generate respective responses, aligning with the complexity and diversity of the authentic human therapist-client exchange.
The 'Interview data' on the other hand consists of 6338 question-answer pairs from 378 interview transcripts that are curated from ongoing clinical trials transcribed by human experts based on the audio recordings of behavior intervention sessions between behavior health coaches and caregivers of individuals in palliative or hospice care. These transcripts were paraphrased and summarized with filtering conversations less than 40 words (each question and answer) by employing the local Mistral-7B-Instruct-v0.2 model. Finally, the researchers contributing to this dataset instruction-tuned Mistral LLM to summarize the whole conversations into single page to deliver concise advice by the behavioral health coach.
The 'ShenLab/MentalChat16K' dataset is was designed to train and evaluate language models for Mental Health virtual assistance or chat bot or dialogue systems. Therefore, its extensive structure and explicit prompting instructions eliminate the need for pre-processing, with key factors adding to this:

- Well Defined Structure: The corpus is well designed with clear prompting style accompanied with 'input' by the client and coherent 'output' as responses by the therapist. This formatting style is adaptable to train and evaluate language models for the task of question-answering system without further pre-processing.
- Consistency with LLM Responses: The dataset employs a uniform conversational format with discrete separation prompt and response. Additionally, the contextually rich dialogues make it easier to train the model with individual exchanges.
- Quality Assurance by Dataset Creator: As a publicly available dataset on platforms like Huggingface, the dataset has gone through rigorous validation process by its creator (ShenLab). Thus, this ensures minimal noise, redundancy or inconsistency in the data.
- Compatibility with training pipeline: Dataset available on platforms like HuggingFace, are designed to be compatible with ML libraries i.e. (Hugging Face, Dataset, Transformers), which facilitate direct tokenization, integrating into training pipelines, and saving trained models in huggingface repository.

3.3 Tools and Resources Used

3.3.1 Hardware

This study was conducted on 2020 MacBook M1 with 8GB of memory, running MacOS Sonoma version 14.3.1, and equiped with 245.11GB of internal storage. Further, the model training and evaluation was conducted in Google Colaboratory Pro.

3.3.2 Software and Libraries

This research is a result of several software and libraries working in synergy. Starting with the programming language used 'Python' version: 3.10.12. As stated above,

the study was conducted in Google Colaboratory Pro leveraging GPU: T4 High RAM which works on 1.67 Computing Units per hour, equipped with 12.7GB system RAM, 15.0GB GPU RAM and 112.6GB Disk memory and A100 which performs on 10 Computing Units per hour, and avails 83.5GB system RAM, 40GB GPU RAM and 112.6GB Disk memory. Covering the libraries, the study leveraged 'HuggingFace Transformer' for model handling and fine-tuning LLMs, 'Datasets Library' for loading and processing 'MentalChat16k', 'accelerate' from HuggingFace to optimize resource allocation (CPU, GPU), 'Peft' (Parameter-Efficient fine-tuning) for employing LoRa for memory efficient fine-tuning, 'bitsandbytes' for 4bits and 8bits Quantization of language models, 'PyTorch' for backend training LLMs, 'Tensorboard' for logging and monitoring model performance, 'trl' to provide SFTTrainer class for supervised fine-tuning of LLMs.
For evaluation of fine-tuned models, the research installed libraries including 'evaluate' for standardized evaluation metrics for NLP tasks, 'bertscore' to calculate bertscore i.e. semantic similarity between two pieces of text, 'rouge_score' to evaluate the overlap between the generated and reference text.

3.3.3 Model Architecture/ Design Specifications

In this research on Virtual Mental Health domain-specific question-answering task, I chose large language models (LLMs), leveraging their extensive design for Natural Language Understanding and Generation tasks. To better tailor their potential, this study employs Parameter-Efficient Fine-Tuning (PEFT) technique LoRa (Low-Rank Adaptation) Hu et al. (2021) accompanied with Memory Efficient Quantization technique.
The fundamental architecture of the employed LLMs is based on the Decoder-Transformer, characterized by auto-regressive framework which works on the principle of sequential word prediction based on the preceding text. The mathematical representation of Sequential Prediction that follows the Chain Rule of Probability is:

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

Large language Models are subsequent to many Natural language Processing (NLP) tasks, encompassing understanding and generation of coherent and contextually relevant text. Language Models are formulated with Transformer Block which uses Attention Mechanism to process input sequence and generate contextually coherent text.
The process initiates with Tokenization, where textual data is broken down into smaller units called tokens of words, punctuations or signs. These tokens are passed into the Embedding layer to assign high-dimensional vector corresponding to each token, with the objective to establish a relationship between tokens in Numerical format. To maintain contextual consistency and relevance, these embeddings are passed through Positional Encoding. Since transformers lack the inherent knowledge about the word order, positional encoding injects information about position by adding pre-defined embedding vectors to embeddings. This demonstrates the order, meaning and relevance of a word in the sentence.
This process is followed by the prediction of the next possible word based on the preceding text by calculating probabilities over vocabulary, following an auto-regressive framework that generates text sequentially, one at a time. This process is performed by a really large neural network, comprising multiple layers of Transformer Blocks, which allows the model to capture complex patterns and nuances in language.

To fully exploit the potential of LLMs for complex domain-specific tasks of Virtual mental-health assistant, this study incorporates Instruction-tuning by leveraging LoRa fine-tuning strategy. LoRa (Low-Rank Adaptation) is a Parameter-Efficient fine-tuning technique that enables task-specific adaptation of LLMs by modifying only a small fraction of their parameters. Models are trained on enormous number of parameters (eg. 250M, 50B, 300B), updating all parameters of the model isn't computationally feasible and often impractical for most use cases. Rather than updating all parameters of the model during fine-tuning, LoRa injects 2 rank decomposition matrices to specific layers, such as the attention layers (q_proj, v_proj), which usually contain most parameters and are critical for learning task-specific patterns, while keeping the other weights of the model frozen. Specifically, for a for a pre-trained weight matrix W, LoRa augments its parameters as by calculating:

$$\Delta W = A \times B$$

Where, A and B are low-rank matrices, while the original parameters of W remain frozen. This approach is computationally efficient and minimizes memory usage making it feasible to Instruction-tune LLMs in limited resources and time. By targeting only the most impactful layers, LoRa balances efficiency with performance, ensuring the model generates responses tailored to the Virtual Mental Health domain.

Quantization is another technique that compliments LoRa in the pursuit of fine-tuning LLMs with constrained resources. This process reduces the precision of model weights and activations is a conversion from higher-memory format to lower-memory format on the fundamentals of Calibration. For instance, converting 32-bit floating-point values to lower-precision formats, such as 8-bit or 4-bit integers, significantly reduces memory usage and speeds up computations. This study uses Post-Training Quantization (PTQ), which applies quantization after the model is pre-trained, and Quantization-Aware Training (QAT), which incorporates quantization during fine-tuning for better accuracy. By leveraging libraries like bitsandbytes, the quantization process ensures that the computational efficiency of the fine-tuned LLMs does not compromise their performance in generating coherent, contextually appropriate responses.

In addition to training, the language model calculates loss using the Cross-Entropy, a common loss function for language models. The SFTTrainer from trl library is built on HuggingFace's Trainer and thus uses the default Cross-Entropy loss function to configure loss in the transformer model. For language modeling tasks, cross-entropy loss measures the difference between the predicted probability distribution over the vocabulary and the true target token. Formulated on the mathematical representation of cross-entropy loss:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i \mid x_1, x_2, \ldots, x_{i-1})$$

4 Implementation

4.1 Data Preparation

The dataset selected for this study is 'MentalChat16k' loaded from HuggingFace repository and splits the original training set into 80% training and 20% test set. Further, the training set is split into 90% training and 10% validation sets combined into 'Dataset dictionary' for easy access. The dataset is then formatted into a structure compatible with LLMs (like Llama). This re-formation combines the 'Instruction' and 'Input' as a coalition prompt with '[INST]' annotation for clarity and better understanding, to serve the model stripped of leading/trailing spaces. Finally, each reformatted example is returned to a dictionary format with a single annotation 'text' encompassing the altered Input.

4.2 Baselines

4.2.1 Flan-T5 Base 250M:

Flan-T5 Hugging Face (2023) by Google is the Instruction Fine-tuned version of T5 or Text-to-Text Transfer Transformer language model, trained on C4 dataset (Colossal Clean Crawled Corpus) approx. 750GB of cleaned data after processing. It is an Encoder-Decoder architecture, sequence-to-sequence based model designed for domain-specific fine-tuning tasks like question-answering system. The encoder part of transformer processes the input sequences, while the decoder generates output. It's comparatively smaller size makes it suitable for resource-constrained environment and is a great start to initiate working with language models.

4.2.2 Tiny-llama 1.1B:

Tiny-llama 1.1B Zhang et al. (2024) is a reduced version of llama, trained on approx. 1 Trillion tokens for about 3 epochs.Tiny-llama is optimized for efficiency while retaining the potential, usually utilized for tasks that require generative abilities with a smaller memory footprint. This model like Llama models is a Transformer-Decoder based auto-regressive model. The model incorporates self-attention mechanism and uses optimizations such as FlashAttention and Lit-GPT to capture dependencies in sequences, and achieve high-quality text generation.

4.2.3 Llama-2 7B:

Llama-2 7B Touvron et al. (2023) is a decoder only transformer model from the Meta AI's language model family trained on approx. 2 Trillion tokens of online publicly available data, and by default supports a context length of 4096. The Llama-2 7B is a large-scale state-of-the-art language model renown for its extensive performance in complex tasks, incorporated with 7 billion trainable parameters. The model is encapsulated with Normalization- equipped with RMSNorm(normalizing inputs based on their root mean square values) , self-attention layer and rotatory positional-embeddings (RoPE) ensuring the model can handle long-range dependencies in sequences and SwiGLU activation function enabling expressiveness. Additionally, it's Reinforcement Layer with Human Feedback (RLHF) that aligns model's performance with human preferences.

4.2.4 Gemma-2 2B:

Gemma-2 2B Team (2024) is a mid-sized Transformer based text-to-text language model that is a part of high-performance Gemma family available in three sizes: 2B, 7B, 27B, trained through distillation from larger models. Gemma-2 2B is a 28 stacked layer, decoder-only transformer model with 2 billion trainable parameters and trained on 256128 vocabulary size and context length of 8192 tokens, configuring it can process approx. 6144 words at a time. With a feed-forward layer after the attention-mechanism layer, the standard ReLu non-linearity is replaced by the GeGLU activation function, a variation of GLU (Gate Linear Unit), dividing the activation into two parts: a sigmoidal part and a linear projection, resulting in a non-linear activation block.

4.2.5 GPT-Neo 2.7B:

GPT-Neo 2.7B Black et al. (2021) is an open-source transformer based model belonging to the GPT family, replicating the Eleuther AI's GPT-3 architecture. This model was trained for 420 billion tokens over 400,000 steps, as a masked auto-regressive language model, making it an extensive fit for next text generation/prediction. GPT-Neo was trained on the Pile 800GB, a dataset curated by EleutherAI comprising diverse text sources, enabling understanding and generation for a wide range of language patterns.

4.3 Experimental Setup

The study initially experiments with fine-tuning the 5 language models on 1 epoch and evaluates their performance. Based on the results, three best performing LLMs i.e. Gemma-2, GPT-Neo and Llama-2 are further retrained on 10 epochs to asses the consistency of their performance. Leveraging the Google Colaboratory Pro for implementation of training and evaluating language models. The GPU employed for Q-LoRa Fine-tuning in **1 epoch experiment** is T4 High RAM. While, in **10 epoch experiment** A100 GPU is utilized.
The base model is loaded from HuggingFace repository and corresponding tokenizer was used to automatically tokenize the textual data. Ensuring proper alignment of input sequence with the pre-trained model. Special tokens, such as 'eos' and 'pad', were used to handle sequence termination and padding, ensuring consistency during fine-tuning.
The 'LoRa Rank'(low-rank parameters for layer) is set to be 64 'LoRa Alpha'(alpha parameter or scaling factor) is set to 16, 'LoRa Droput'(dropout probability for lora layers) is set to be 0.1 to prevent overfitting. These parameters demonstrate the model layers adjusted for minimal training and efficient performance of PEFT technique.
Adjustments for Quantization parameters include 'use_4bit' activating 4 bit precision to reduce memory usage, 'bnb_4bit_compute_dtype' (compute datatype for 4bit model) is set to be float16 for efficient computation and 'bnb_4bit_quant_type' (specified quantization type) is set to be nf4 optimized for Transformers and 'use_nested_quant'(Optional double quantization) is set to be false to deactivate double quantization. These settings enable resource-constrained and memory efficient Instruction-Tuning of LLMs.
Post Q-LoRa parameter setting, comes the Training Argument section that controls the training process. The output directory is mentioned to save model checkpoints and logs, 'save_steps' is set to 0 to avoid intermediate checkpointing, Number of 'epochs' for training is set to be 20. 'per_device_train_batch' and 'per_device_eval_batch' size is set to 4 for GPU batch size for each training and evaluation. For Gradient Optimization: 'gradient_accumulation_steps' is set to 1 which demonstrates the accumulation of number of updated steps to save the 'gradient_checkpoint' set to True, 'max_grad_norm'(gradient clipping to prevent exploding gradients during training) is 0.3. Setting for Learning Parameters: 'learning_rate' is set to be 2×10^{-4} and 'learning_rate_scheduler' as cosine for 'Adam Optimizer' i.e. 32bit paged adam, 'weight_decay'(apply to all layers except bias and Norm layer) set to 0.001, 'warmup_ratio' set to over 3% of training steps for linear warmup. Setting 'max_steps' to -1 and 'logging_steps'(log every X steps post updates) to 25 and 'group_by_length' setup True to group sequences into batches with

same lengths, thereby minimizing padding and improving efficiency. Finally, both precisions 'fp16' and 'bf16' are set to be False, but 'bf16' is set True if operated on A100 GPU. 'Max_seq_length' is set to None for the model to adapt according to examples and 'packing' is disabled to configure each example as independently, loading the entire model on GPU 0.

5 Evaluation

LLMs	Training Loss	Final Loss	Train Runtime (sec)
1 Epoch- T4 GPU: (15GB)			
GPT-Neo 2.7B	1.336982	1.245100	17217.08
Gemma-2 2B	0.881864	0.982700	**16813.02**
Llama-2 7B	**0.777635**	**0.765700**	26722.05
10 Epochs- A100 GPU: (40GB)			
GPT-Neo 2.7B	0.923790	0.724100	18027.73
Gemma-2 2B	**0.360185**	**0.038000**	**17615.65**
Llama-2 7B	1.246209	0.329300	38162.32

Table 1: Comparison of Training and Final Losses for LLMs across 1 and 10 Epochs.

To evaluate the experiment of training the models on 1 epoch and 10 epochs, although using different GPU from google colaboratory pro. On illustration of Table 1, the loss has considerably reduced after re-training models for 10 epochs on same dataset lowering the 'Final Loss' from 0.76 to 0.03 for Gemma-2 2B, for GPT-Neo from 1.24 to 0.72 and for llama-2 from 0.76 to 0.32.
The study utilizes automatic evaluation metrics to assess the performance of large language models. The evaluation metrics leveraged for quantitative testing of the LLMs are: BLEU, ROUGE and BERTScore. Test-set is utilized to measure their performance on these metrics (50% for 1 epoch, 20% for 10 epoch model). The study has evaluated on only a small portion of tests because of time and resource constrains, since even this process took 7-20 hrs. These metrics provide insights into fluency, relevance and semantic accuracy of the generated response compared to the reference. Each of the automatic evaluation metrics have individual significance in evaluating the quantitative performance of the different language models.
While evaluating the fine-tuned models, **Tiny llama 1.1B** demonstrated **high tendency of Hallucination**, generating incoherent responses that deviated from the therapy context. On the other hand, **Flan-T5** consistently generated **single sentence responses** just asking the question repeatedly or no response at all, which lacked depth and relevance. Thus, due to these limitations the study did not pursue evaluating these models further as they did not suffice the requirement of an effective Virtual Mental Health Therapist.

5.1 Metrics Used

- **ROUGE Score:**

LLMs (1 Epoch)	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
Llama-2	0.375506	0.111392	0.172646	0.297598
GPT-neo	0.382881	0.114003	0.172279	0.300026
Gemma-2	**0.393325**	**0.124284**	**0.175737**	**0.306454**
LLMs (10 Epochs)	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-LSUM**
Llama-2	0.485221	0.1817300	0.215679	0.395310
GPT-neo	0.471221	0.166945	0.203523	0.383208
Gemma-2	**0.507834**	**0.235510**	**0.265641**	**0.432990**

Table 2: Comparison of ROUGE Scores for LLMs trained on 1 epoch and 10 epochs.

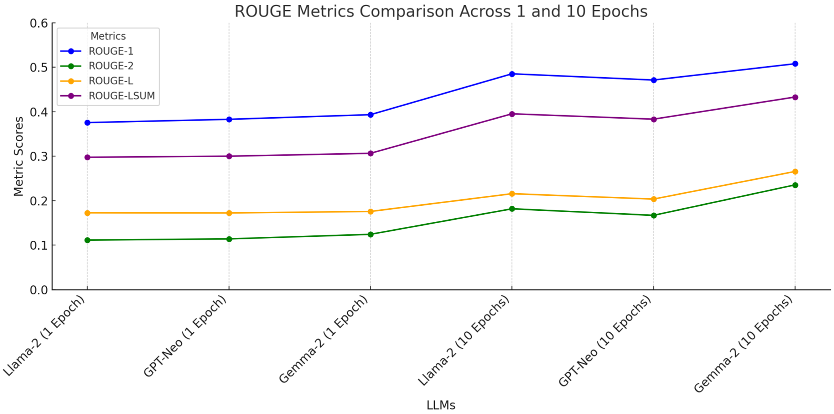

Figure 2: ROUGE Comparison Graph

ROUGE or (Recall-Oriented Understudy for Gisting Evaluation), is a set metrics specially designed for automatic summarization, but can also be used for other NLP tasks. ROUGE-N measures the number of matching N-grams between the model-generated text and reference. 'N-gram' is a widely used concept from text processing in different NLP tasks. It refers to 'N' (i.e. 1,2,3..) consecutive words in a context. ROUGE-1 precision is computed as the ratio of number of uni-grams in Reference that also appear in Model generated text over total number of uni-grams in generated response, while recall is computed against total uni-grams in Reference.

$$\text{ROUGE-N Precision} = \frac{\text{Number of overlapping N-grams}}{\text{Total N-grams in candidate}}$$

$$\text{ROUGE-N Recall} = \frac{\text{Number of overlapping N-grams}}{\text{Total N-grams in reference}}$$

Enunciating Table 2, Gemma-2 achieves the highest ROUGE score across all metrics demonstrating better lexical and structural similarity. Wherein, GPT-Neo yet being a smaller model slightly outperforms Llama-2, with modest improvements in ROUGE-1 and 2 scores. While ROUGE Score is independent of language and is coherent with Human Evaluation being inexpensive to compute, it doesn't account for synonyms or words that have the same meaning, as it measures syntactical matches rather than semantics.

- **BERT Score:**

LLMs (1 Epoch)	**BERT-Precision**	**BERT-Recall**	**BERT-F1**
Llama-2	0.853338	0.847776	0.850484
GPT-Neo	0.856429	0.851492	0.853880
Gemma-2	**0.859081**	**0.854886**	**0.856904**
LLMs (10 Epochs)	**BERT-Precision**	**BERT-Recall**	**BERT-F1**
Llama-2	0.8640052	0.875414	0.869581
GPT-Neo	0.857535	0.873456	0.865255
Gemma-2	**0.863883**	**0.882919**	**0.873012**

Table 3: Comparison of BERT Scores for LLMs trained on 1 epoch and 10 epochs.

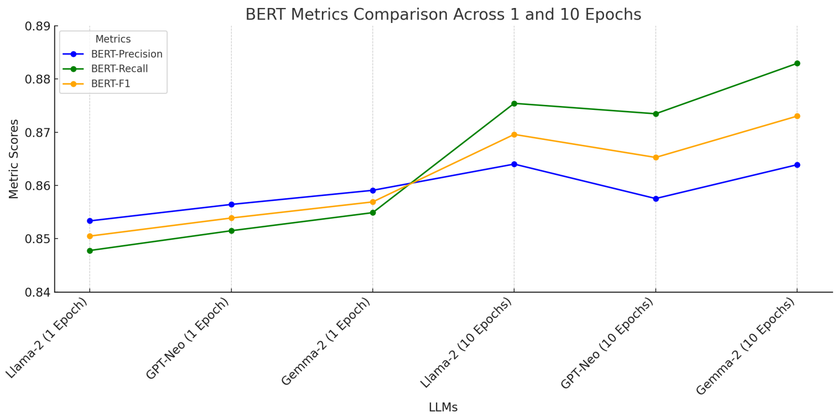

Figure 3: BERT Comparison Graph

BERT is a vital metric that acts as an extensive alternative to traditional evaluation metrics in the NLP domain. It was introduced to measure the quality of generated text in terms of similarity with the reference text. Traditional evaluation metrics based on n-grams do not follow semantic reordering and penalize long-range dependencies. To cover for the disadvantage of n-gram metrics, leading to underestimated performance as semantically correct phrases deviate from the reference. Whereas, in BERTScore the similarity between two sentences is computed as the sum of the cosine similarities between their token embeddings, thereby providing the capability to detect paraphrases.
To demonstrate the quantitative results in Table 3, Gemma-2 achieves the highest BERT score across all metrics, portraying it's superiority in capturing semantic meanings. While, GPT-Neo follows slightly better performance than llama-2, even while being a comparatively smaller model. As a result, Gemma-2 proves to provide the most semantically accurate and contextually relevant responses amid the three language models.

- **BLEU Score:**

BLEU or Bilingual evaluation understudy is an algorithm usually utilized for in machine translation (evaluating text from one machine language to another). The

LLMs (1 Epoch)	BLEU Score	BLEU Precision	Brevity Penalty	Length Ratio
Llama-2	0.066838	**0.491728**	0.572064	0.641641
GPT-Neo	0.069953	0.478261	0.629262	0.683430
Gemma-2	**0.076635**	0.489027	**0.632543**	**0.685868**
LLMs (10 Epochs)	**BLEU Score**	**BLEU Precision**	**Brevity Penalty**	**Length Ratio**
Llama-2	0.1468767	**0.474781**	1.0	1.115437
GPT-Neo	0.121965	0.430964	1.0	1.256138
Gemma-2	**0.200134**	0.469546	1.0	**1.218975**

Table 4: Comparison of BLEU Scores and related metrics for LLMs trained on 1 epoch and 10 epochs.

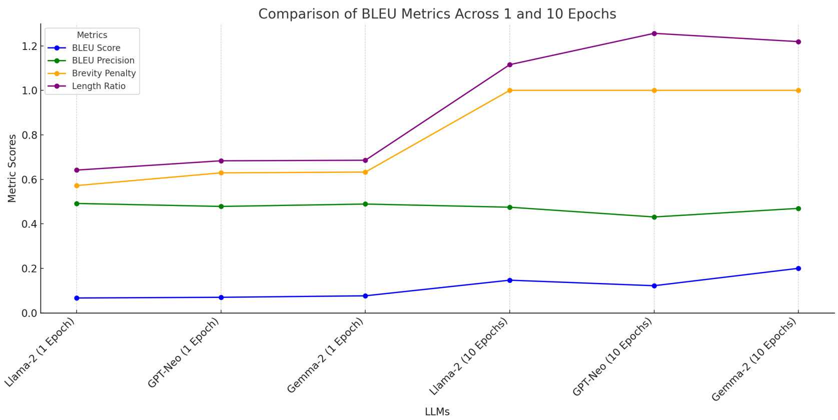

Figure 4: Comparison Graph for BLEU evaluation metrics

evaluation metric assesses the quality of text generated by the model to the human generated text. To assess the quality of text quantitatively, BLEU Score has further subdivisions testing performance including precision, brevity penalty (measure if the translated text is shorter than the reference text) and length ratio (ratio of translation text to the actual text).
To enunciate the results in Table 4, although BLEU score is relevant for translation tasks Gemma-2 achieves the highest BLEU score and a balanced precision and length score while Llama outperforms Gemma in BLEU precision, demonstrating better fluency and adequacy among the evaluated language models.

5.2 Comparative Quantitative Analysis

LLMs (10 epochs)	BERT-Precision	BERT-Recall	ROUGE-1	ROUGE-Lsum
Llama-2	0.864005	0.875414	0.485221	0.395310
GPT-Neo	0.857535	0.873456	0.471221	0.383208
Gemma-2	**0.863883**	**0.882919**	**0.507834**	**0.432990**

Table 5: Performance comparison across BLEU, BERT, and ROUGE metrics for different LLMs for 10 epoch retraining.

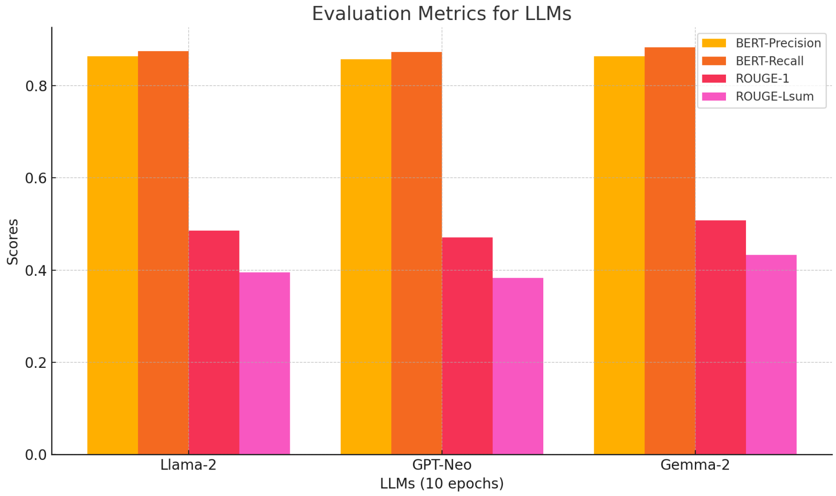

Figure 5: Comparison Graph for all evaluation metrics

On Quantitative analysis from Table 5 of all three LLMs, the BLEU score is collectively low for all trained language models since BLEU Score is an evaluation metric to measure translational similarity between the translated text and reference/ human generated text, yet Gemma-2 outperformed the other LLMs. Since, this is a text-generation task, and the study doesn't benefit from text similarity it evaluates on other metrics like BERT and ROUGE. Wherein BERT Precision score is high altogether, being a smaller model that Llama-2 which is trained on 7 billion parameters, Gemma-2 (2 Billion parameters) outperforms with precision score 0.86. Gemma-2 with 0.88 BERT Recall score outperforms GPT-Neo with 0.873 and Llama-2 with 0.847 recall. ROUGE is a significant score that primarily measures the overlap of word sequence and word pairs of the candidate and reference text which makes more sense in test-generation tasks, ranging from 0-1, Gemma-2 outperformed in ROUGE-1 with 0.50 which is relatively a good ROUGE score. Although, these scores do not fully capture the adaptability of the language model for task of virtual therapy, they don't demonstrate the potential of model to produce empathetic, contextually appropriate, and semantically rich responses.

5.3 Qualitative Analysis

Emphasizing on Qualitative analysis is a more significant evaluation technique for Mental Health domain since the language model generated outputs should reflect empathy, care and emotional while providing relevant advice. All three models were tested on the same prompt to assess their responses.

While Flan-T5-base had incoherent responses and Tiny Llama demonstrated high tendency of Hallucination, other models performed comparatively intensively better. The response generated by Llama-2 model exhibits empathy and support, with phrase like "I understand" and providing different relevant approaches. Its concise structured advice ensured mental, physical, emotional well-being. Gemma-2 response demonstrates a clear structure, coherent advice i.e. deep-breathing exercise and slow exposure to crowded places, providing tips with an encouraging tone. Effectively balancing stress reduction. Response generated by GPT-Neo exhibits generic stress-management techniques with helpful advice, providing situational reassurance and acknowledging user's anxiety, yet the generic response lacks empathy in its tone and depth in understanding.

5.4 Experiment / Case Study 1

Question

```
  Another trigger for my anxiety is crowded places. Whenever I'm in a crowded
environment, such as shopping malls or public transportation, I feel overwhelmed
and suffocated. The noise, the people, and the lack of personal space make me
extremely anxious. Just last week, I had to leave a store because it was too
crowded, and I felt trapped and panicked.
```

Model Responses

Llama-2:

> It can be challenging to navigate crowded places when you're experiencing anxiety. One approach that may help is gradually exposing yourself to these environments in a controlled way. You could start by visiting less crowded places during off-peak hours or during the week, when the crowds may be smaller. As you become more comfortable, you can gradually work your way up to busier locations. Deep breathing exercises can also be beneficial in managing anxiety symptoms in crowded places. Taking slow, deep breaths can help activate your body's relaxation response and reduce feelings of tension. You might find it helpful to practice these techniques before entering a crowded environment and carry a deep breathing exercise worksheet with you as a reminder. Additionally, it may be useful to explore any underlying beliefs or thoughts that contribute to your anxiety in crowded places. Are there specific fears or concerns that arise when you're in these situations? Identifying and challenging any irrational or unhelpful thoughts can help alleviate anxiety. Engaging in regular physical activity can also have a positive impact on overall mental well-being and anxiety levels. Consider incorporating exercise into your routine, whether it's going for a walk, practicing yoga, or participating in a sport you enjoy.

GPT-Neo

> It can be challenging to navigate crowded places, especially when they trigger anxiety for you. The noise, people, and lack of personal space can indeed make such environments overwhelming and suffocating. It's important to remember that you're not alone in feeling this way, as many people experience similar feelings when they are in crowded places. One suggestion I have is to try practicing deep breathing exercises in these situations. When you start feeling anxious, take slow, deep breaths in through your nose and out through your mouth. This can help calm your nervous system and reduce the intensity of your anxiety. You might find it helpful to focus on your breath as you inhale and exhale, allowing yourself to relax and regain a sense of control. Another strategy that might be beneficial is to gradually expose yourself to crowded places. Start by visiting smaller, less overwhelming environments, like a shopping mall or public transportation. As you become more comfortable, gradually increase the size of the crowd and the level of sensory overload. This gradual exposure can help desensitize you to the triggers and build your confidence in managing your anxiety in these situations. Additionally, consider finding ways to incorporate small, manageable steps into your daily routine. For example, instead of trying to navigate a crowded store all at once, you could start by taking short strolls or engaging in brief conversations with others. By taking small steps, you can build your confidence over time and reduce the overall anxiety associated with crowded places.

Gemma-2

I can understand how overwhelming and suffocating it must be to be in crowded places like shopping malls or public transportation. It sounds like these environments trigger a lot of anxiety for you, with the noise, people, and lack of personal space all contributing to your feelings of panic. Leaving a store due to feeling trapped is a clear indication that your anxiety is impacting your daily life and causing distress. One suggestion I have is to try using deep breathing techniques when you find yourself in crowded places. Deep breathing can help regulate your breathing and reduce feelings of anxiety. You could try inhaling slowly through your nose, holding your breath for a few seconds, and then exhaling slowly through your mouth. Repeat this process several times until you start to feel more calm and centered. Additionally, it might be helpful to practice grounding exercises in these situations. Grounding involves focusing on your senses to bring your attention back to the present moment. For example, you could try identifying five things you can see, four things you can touch, three things you can hear, two things you can smell, and one thing you can taste. Engaging your senses can divert your attention away from anxious thoughts and provide a sense of comfort. Another strategy that may be beneficial is gradually exposing yourself to crowded environments. Start by visiting less crowded places, such as small shops or quiet parks, and gradually work your way up to larger stores or busy streets. This gradual exposure can help desensitize you to the triggers and build your confidence.

6 Conclusion and Future Work

This study adopted a secondary approach in the realm of Psychological Mental Health support, bridging the gap between conversation based psychological therapy and large language models. By introducing a methodology to utilize LLMs in a resource-constrained environment demonstrated the use of Parameter efficient Fine-tuning technique like LoRa and Memory Efficient strategy including Quantization. Empirical analysis and quantitative evaluation (where MODEL outperformed other models with THIS ROUGE-1 score), reaffirmed the robustness of the PEFT Instruction-tuned model, with a comparative inspection of 5 language models, where llama-2 7b, gemma-2 2b and GPT-Neo 2.7b performed with excellence generating coherent and empathetic advised responses for psychological health support.

In future, research can focus on two major aspects. First, modeling and leveraging a more comprehensive data with extensive psychological therapy like CBT, ACT and DBT based conversations to enhance language model's performance. Secondly, to mimic a real world therapy session transitioning from single turn question-answer to multi turn question-answer enhancing the realism and depth of virtual mental health system.

References

AXA (2024). Axa mind health report 2024.

Black, S., Leo, G., Wang, P., Leahy, C. and Biderman, S. (2021). GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.
URL: *https://doi.org/10.5281/zenodo.5297715*

Burger, F., Neerincx, M. A. and Brinkman, W.-P. (2021). Natural language processing for cognitive therapy: Extracting schemas from thought records, *PLOS ONE* **16**(7): e0254323.

Carlson, C. G. (2023). Virtual and augmented simulations in mental health, *Current Psychiatry Reports* **25**(9): 365–371.

David, D., Cristea, I. and Hofmann, S. G. (2022). Why cognitive behavioral therapy is the current gold standard of psychotherapy, *Frontiers in Psychiatry* **9**: 4–10.

Hiraoka, T., Hiraoka, T. and et al. (2022). Generating descriptions from historical changes in time-series data, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2225–2240.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

Hugging Face (2023). google/flan-t5-base.

Ji, S., Zheng, X., Sun, J., Chen, R., Gao, W. and Sr, M. (2024). Mindguard: Towards accessible and stigma-free mental health first aid via edge llm. Preprint.

Jia Xu, Tianyi Wei, B. H. P. O. S. Y. R. J. R. P. J. W. G. D. L. S. (2024). Mentalchat16k: A benchmark dataset for conversational mental health assistance.
URL: *https://huggingface.co/datasets/ShenLab/MentalChat16K*

Kim, Y. (2014). Convolutional neural networks for sentence classification.

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y. and Wang, Z. (n.d.). Psy-LLM: Scaling up global mental health psychological services with AI-based large language models.

Liu, H., Lu, Y., Zhang, D., Yang, Y., Ren, X. and He, X. (2021). Clip-event: Connecting text and images with event structures, *arXiv preprint arXiv:2106.01702* .

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G. and White, A. D. (n.d.). PaperQA: Retrieval-augmented generative agent for scientific research.

Madani, A., Krause, B. and et al. (2023). Large language models generate functional protein sequences across diverse families.

Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M. S. and Chakraborty, T. (n.d.). Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations, *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ACM, pp. 735–745.

Morris, R., Schueller, S. and Picard, R. (2015). Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: Randomized controlled trial, *Journal of Medical Internet Research* **17**(3): e72.

Na, H. (n.d.). CBT-LLM: A chinese large language model for cognitive behavioral therapy-based mental health question answering.

Prabhune, S. and Berndt, D. J. (2024). Deploying large language models with retrieval augmented generation.

Radwan, A., Amarneh, M., Alawneh, H., Ashqar, H. I., AlSobeh, A. and Magableh, A. A. A. R. (2024). Predictive analytics in mental health leveraging llm embeddings and machine learning models for social media analysis.

Rojas-Barahona, L., Tseng, B.-H., Dai, Y., Mansfield, C., Ramadan, O., Ultes, S., Crawford, M. and Gasic, M. (n.d.). Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy.

Seighali, N., Abdollahi, A., Shafiee, A. et al. (2024). The global prevalence of depression, anxiety, and sleep disorder among patients coping with post covid-19 syndrome (long covid): a systematic review and meta-analysis, *BMC Psychiatry* **24**: 105.

Sharma, A., Rushton, K., Lin, I., Wadden, D., Lucas, K., Miner, A., Nguyen, T. and Althoff, T. (n.d.). Cognitive reframing of negative thoughts through human-language model interaction, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 9977–10000.

Sun, H., Lin, Z., Zheng, C., Liu, S. and Huang, M. (2021). PsyQA: A Chinese dataset for generating long counseling text for mental health support, *in* C. Zong, F. Xia, W. Li and R. Navigli (eds), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Online, pp. 1489–1503.

Team, G. (2024). Gemma.
URL: *https://www.kaggle.com/m/3301*

Touvron, H., Martin, L., Stone, K. and et. al, P. A. (2023). Llama 2: Open foundation and fine-tuned chat models.

World Health Organization (2023). Mental health: Strengthening our response.

Xin, C. and Zakaria, L. Q. (2024). Integrating bert with cnn and bilstm for explainable detection of depression in social media contents, *IEEE Access* **PP**: 1–1.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K. and Wang, D. (n.d.). Mental-LLM: Leveraging large language models for mental health prediction via online text data, **8**(1): 1–32.

Zhang, P., Zeng, G., Wang, T. and Lu, W. (2024). Tinyllama: An open-source small language model.
URL: *https://arxiv.org/abs/2401.02385*

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J. and Dolan, B. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), System Demonstrations*, Association for Computational Linguistics, Online.

Ziems, C., Li, M., Zhang, A. and Yang, D. (2022). Inducing positive perspectives with text reframing, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, Association for Computational Linguistics, Dublin, Ireland.