Evaluating Lightweight open-source LLMs for Mental Health Counselling: A Comparative Study

Nikhil Rajput, Diviit MG
Department of Artificial Intelligence
Amity University Noida

Abstract—**Advancements in Large Language Models(LLMs) have led to substantial progress in the field of text generation and understanding human cognition. However, there's still a significant gap in the research regarding the capabilities of LLMs in the field of Mental Health Counselling. This work presents a comparative analysis of different lightweight open-source LLMs such as Google-T5, BART, FLAN-T5, and Microsoft-Godal fine-tuned on a custom dataset. In this study, we compiled diverse mental health counselling datasets from reputable online sources, which were meticulously pre-processed to meet ethical, privacy, and technical requirements for effective model training and evaluation. The result indicates the superior performance of the BART model across all evaluated metrics, with notably higher ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores compared to T5 and Flan-T5. We further evaluated each model's ability to generate responses based on coherence, practicality, and emotional support, as well as their level of empathy. Our findings showed that BART outperformed the other models, demonstrating the best overall performance. Hence BART, among the tested models without prompting, holds the greatest promise for supporting mental health counselling applications through AI-driven conversational agents.**

I. Introduction

Recently, the use of large language models (LLMs)[1], such as T5 (Google)[2], Flan-T5 (Google research)[3] and BART[4] has expanded significantly in tasks such as question-answering, natural language understanding, text generation, and summarization. Studies have shown that LLMs which are built on hundreds of billions of parameters are increasingly capable of understanding human logic and doing proper reasoning and inference.

However, despite these advancements, there remains a gap in applying LLMs specifically for mental health counselling, where their potential for therapeutic communication is still under-explored. Mental health is recognized as a critical global challenge, with the WHO reporting that 1 in 8 individuals worldwide suffers from mental disorders[5]. According to the National Mental Health Survey of India (2015-16), 13.7% of the population has a diagnosable mental illness, with 10.6% requiring urgent care[6]. Despite growing awareness and advancements in India, the country faces a significant shortfall in mental health professionals, with only 0.75 psychiatrists per 100,000 people, far below the WHO-recommended 3 per 100,000.[7]

Recent advancements in AI have led to the development of several mental health counselling tools, primarily in the form of chatbots[8]. However, most of these systems are rule-based or rely on traditional machine learning models, limiting their adaptability. These models are typically domain-specific, focusing on particular issues such as depression, anxiety, suicide prevention, or stress management. While useful for targeted interventions, their lack of flexibility and generalization across broader mental health challenges makes them less reliable. Additionally, these systems often struggle with the nuanced understanding required for effective therapeutic communication, which raises concerns about their scalability and effectiveness in real-world clinical settings.

Since natural language plays a crucial role in understanding, assessing and treating mental health conditions, Large Language Models (LLMs) have the potential to become a tool of importance in mental health care[9]. These models, whether fine-tuned for certain specific tasks or general use can handle a wide range of inputs, eliminating the necessity to train separate models from scratch for different tasks. In simple words, a single LLM could be utilized for multiple mental health applications, like question-answering, reasoning and inference. Thus, this leads to the creation of virtual support systems in the form of chatbots, monitoring systems and counselling support systems. Furthermore, while some LLMs have shown promise in tasks like language generation and sentiment analysis, little is known about how fine-tuned LLMs perform in actual mental health counselling scenarios. There is a lack of comparative studies analyzing the performance of different LLMs across key counselling dimensions such as empathy, context retention, and adherence to therapeutic guidelines.

While advancements in artificial intelligence, specifically Large Language Models (LLMs), offer potential solutions by automating aspects of mental health support, their effectiveness in delivering empathetic, safe, and clinically appropriate responses in counselling settings remains under-explored. Existing AI models and mental health chatbots primarily rely on predefined rules, often lacking the flexibility and depth necessary for effective therapeutic engagement. Hence to achieve our vision of having a mental health counselling system by leveraging the potential of LLMs, we need to address the research question: **How do various language models compare in accuracy and relevance for mental health counselling tasks?**

In this study, we fine-tuned several open-source Large Langauge Models(LLMs) available such as T5[10], BART[11], FLAN-T5[12] and Godel[13] which are widely used for text generation and other related tasks. For the data, we utilized the publicly available mental health counselling conversations of patients and experts in the form of Questions and Responses from multiple sources. Our methodology involved fine-tuning each model under consistent experimental conditions, followed by a detailed comparative analysis to evaluate their performance in generating contextually relevant and accurate responses within the mental health domain.

Among the models, BART consistently delivered the best performance across multiple evaluation metrics. It achieved the highest ROUGE-1, ROUGE-2, and ROUGE-L scores, indicating a superior ability to capture context and generate coherent, relevant responses. Additionally, BART outperformed the other models in terms of BLEU score, reflecting its enhanced linguistic accuracy in generating fluent and natural language outputs. While the perplexity score for BART was average compared to the other models, its overall performance suggests that BART is the most effective model for mental health counselling applications, balancing both response quality and contextual relevance.

II. Related Works

We briefly provide here the overview of related works on the application of chatbots and LLMs in the field of mental health.

A. AI and Mental Health

The advancement in artificial intelligence and its integration across various sectors has opened up new possibilities, particularly in healthcare. Fusing this rapidly growing advancement in the medical sector can ease the growing mental health crisis. The potential of AI, Large Language Models, and Chatbots have been explored in numerous studies, shedding light on their potential, future scope and sometimes the challenges.

D'Alfonsa et al. (2020) [14]provided us with a foundational perspective of how utilizing digital data and integrating them with AI-incorporated web and smartphone apps can be useful in predicting/detecting mental health conditions. This study is further supported by the work of Feng et al. (2022)[15]who discussed about the application of AI in mental health diagnosis and treatment, acknowledging the ethical and practical challenges.

The immense and vast capacity that Artificial intelligence holds enables it to be integrated as a therapeutic tool which was examined by Darzi (2023)[16], highlighted its potential to identify mental illness at early stages thus, leading to transformation in mental health care. The study dwells into the further development of AI-based therapeutic platforms to diagnose people and reduce the burden on the traditional healthcare system.

Similarly, Shimada (2023)[17], emphasised the importance of artificial intelligence advancement and the promises it holds in mental health care. In a recent study, Kim (2023)[18] demonstrated how much potential personalised AI mental health counselling systems are in helping individuals in crisis.

The growing prevalence of chatbots has started playing a vital role in mental health support systems. Alrazaq et al. (2019)[19], conducted a comprehensive review of 41 chatbots. Similarly, Bendig et al. (2019)[20] examined the limitations of these chatbots concerning their effectiveness and acceptance and stressed the need for further refinements.

Recent studies have also focused on the potential of AI-powered chatbots to serve as a therapeutic tool. A behavioural Activation-based AI chatbot by Rathnayaka et al. (2022) [21] aimed to provide emotional support for its patients. More recently, van der Schyff et al. (2023) [22]in their paper, investigated the efficacy of AI-powered chatbot Leora in supporting patients with depression and anxiety, while addressing the ethical challenges surrounding its deployment.

B. Large Language Models and Health Applications

The integration of AI, particularly the adaptation of Large Language Models in Mental health counselling has proven to be useful after the success of Transformer-based language models like T5, GPT, and BART. The ability to fine-tune such models has provided several frameworks and solutions aimed at addressing therapeutic needs while offering cost-effective and scalable approaches to deal with mental health crises.

In the border healthcare domain, LLMs have shown immense promise. Sathe et al. (2023)[23] explored the potential of LLMs like ChatGPT in healthcare communication. GatorTronGPT by Peng et al. (2023)[24] showed its capability in biomedical natural language processing tasks with an F1 score increment of 3.8% for clinical document classification tasks and a 6.1% increase in accuracy for entity recognition tasks. Another example can be Med-PaLM 2 by Singhal et al., (2023)[25] achieved an accuracy of 86.5% on the MedQA dataset. Similarly, Yang et al., 2023 [26]put forward the LLM-Synergy framework that demonstrated remarkable accuracy (96.21%) on MedMCQA and PubMedQA showcasing the potential for LLMs to tackle clinical questions.

While significant advancements have been made in healthcare applications, using LLMs in mental healthcare is limited. Most existing works focus on classifying tasks, sentiment analysis, entity recognition task and emotional reasoning. According to Stade et al. (2024)[27], the use of LLMs for mental health support systems is increasing. Still, there are certain drawbacks due to the research being in its early stages, dropout rates and user engagement persistence. Supporting this, Lai et al. (2023)[28], revealed that LLMs can alleviate the burden on mental health systems by automating certain aspects of therapy like providing emotional support and identifying suicide risks. Yadav et al. (2024)[29]evaluated fine-tuning LLMs for generating diagnostic screening summaries in mental health care and achieving ROUGE-1 and ROUGE-L values of 0.810 and 0.764 with their top-performing fine-tuned model. Yang et al. (2023) [26]introduced MentalLLaMA, a fine-tuned model for mental health analysis on social media. In another notable study, Liu et al. (2023)[30] provided an innovative framework, Psy-LLM that utilizes Chinese LLMs PanGu and WenZhong to assist in real-time consultations. Zheng et al. (2023) [31]demonstrated the lack of data to optimize models for mental health support systems and created the ExTES dataset which was used to fine-tune certain models like LLaMA, and DialoGPT, further emphasizing the importance of domain-specific data.

Most LLM-based approaches for mental health excel in tasks like sentiment analysis and entity recognition but struggle with unanticipated user inputs and complex diagnoses. While these systems automate aspects of therapy, they often rely on rule-based methods and structured inputs, limiting their effectiveness in real-world scenarios. Challenges such as dropout rates, a shortage of specialized datasets, and user engagement issues hinder the adoption of LLMs for sustained mental health support. While the current systems have shown success in addressing specific issues like depression and eating disorders, they remain largely rule-based and less capable of handling complex, unstructured conversations. Our work aims to bridge this gap and explore the broader potential of LLMs in mental health care, paving the way for their use in more advanced, real-time counselling settings.

III. Methodology

A. Dataset

Data Collection: Mental health counselling datasets used in this project were collected from various publicly accessible sources, primarily on platforms like Hugging Face and GitHub. After gaining access to these datasets, the data was extracted and organised in a standardised format of Question-Response pairs to facilitate consistent fine-tuning and evaluation of the language models. Key datasets utilised include:

- Aditya Mental Health Counselling Dataset [32]
- Mental Health Counselling Chat [33]
- Counsel Chat Dataset [34]
- Amod-Mental Health Counselling Conversations [35]

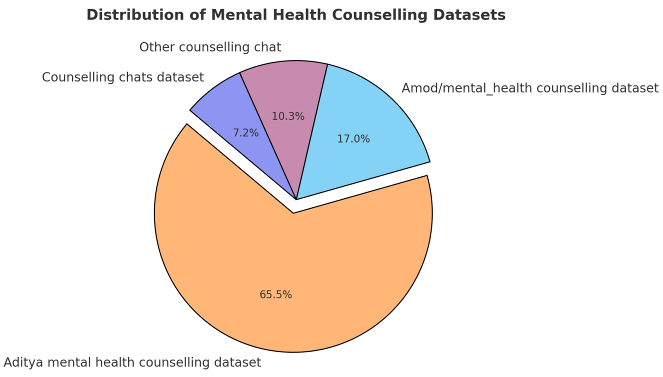


Fig. 1: Data Distribution

These datasets provide various mental health dialogue scenarios and counselling interactions, essential for training models to generate contextually accurate and empathetic responses. Each dataset was reviewed to ensure it aligns with the project's goal of enhancing mental health support via language models, capturing diverse counselling topics and varying levels of emotional support. Table I below provides the summary of the datasets.

1) Dataset Description:: The collected mental health counselling datasets consist of 20,620 dialogue instances, each formatted into Question-Response pairs. These datasets are entirely in the English language and encompass a wide array of mental health topics, including but not limited to anxiety, depression, stress management, relationship issues, self-esteem, and coping mechanisms.

Each dataset provides unique scenarios, enabling the fine-tuned language models to capture the nuances of various mental health concerns and respond with contextually appropriate and empathetic dialogue. The diversity in topics and conversational styles within the datasets supports a comprehensive training environment, promoting the development of language models capable of addressing complex mental health issues.

2) Data Pre-processing:: A systematic data preprocessing approach was implemented to prepare the mental health counselling dataset for fine-tuning the language models, utilizing each model's tokenizer to convert the text data into embeddings compatible with the respective models. The preprocessing pipeline involved the following steps:

1) **Tokenization:** Each question-response pair was formatted for tokenization. For example, in the case of T5, the questions were prefixed with the text *”question: ”* to contextualize the input for the model. Questions and responses were tokenised using the T5Tokenizer from the Hugging Face transformers librar to create model inputs and labels, respectively. Tokenization was configured with a maximum sequence length of 512 tokens to ensure that long conversations were truncated and padded as needed, maintaining consistency in input size.

2) **Input Preparation:** The inputs were created by appending the prefix to each question and then tokenizing these inputs with the *T5Tokenizer* in case of T5 fine-tuning. The tokenized input sequences were truncated to 512 tokens and padded to ensure uniform length across all examples.

3) **Target Encoding:** The responses were similarly tokenized to generate the target sequences (labels). Tokenization was applied with the same settings as the inputs, including truncation and padding. The encoded responses were then assigned to the *”labels”* attribute within the dataset for supervised learning.

4) **Dataset Mapping:** This preprocessing function was applied to both the training and validation datasets. The Hugging Face *Dataset.map()* function was used to batch-

TABLE I: Summary of Four Mental Health Datasets Used in Our Experiment.

Dataset Name	Source	Task	Dataset Size	Text Length (Tokens)	Description
Aditya Mental Health Counselling[32]	Hugging Face, GitHub	Mental health question-answering	13.5k	Varies	Contains Q&A pairs focused on mental health concerns like anxiety and stress.
Mental Health Counselling Chat[33]	Hugging Face	Mental health dialogue generation	100	Varies	Conversations on depression, anxiety, and emotional challenges for therapy models.
Counsel Chat Dataset[34]	Counsel Chat, Hugging Face	Mental health Q&A	3k	Not specified	Licensed counsellors respond to user-submitted questions on mental health.
Amod-Mental Health Counselling[35]	Hugging Face	Mental health conversation modeling	3.5k	Varies	Features dialogues on depression, mood swings, and self-care.

process the tokenization and mapping for all question-response pairs, optimizing for computational efficiency.

Table 1 summarizes the information from all four datasets used for training the models. The final combined data set has been prepared and a 70% - 30% train validation split has been used for this study.

The methodology applied is systematically represented in the following flowchart(Figure 2), which outlines each procedural stage in detail.

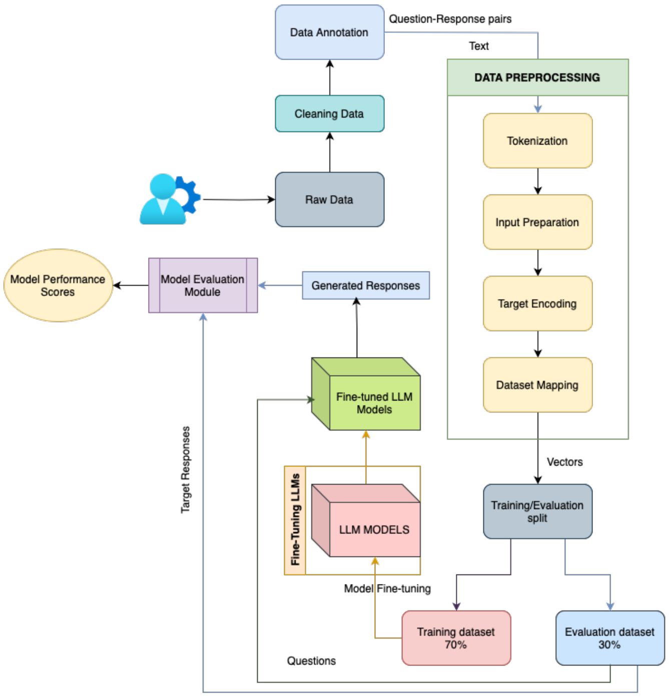


Fig. 2: Methodology flow chart

B. Models

We experimented with different LLMs of different configurations, model sizes, availability and trainable on our resources.

- **T5**: Developed by Google, T5 is a transformer-based open-source model which is preferred for tasks such as text generation, translation, summarization, etc. The model uses a unified framework, treating both input and output as text sequences, which makes it adaptable and versatile. The encoder-decoder architecture makes it much more suitable for sequence-to-sequences tasks. Here we chose the T5-small model with 60 million parameters. [10]
- **BART:** Developed by Facebook, it is a denoising autoencoder for pre-training sequence-to-sequence models. Combines the strengths of bidirectional and autoregressive transformers. BART effectively generates coherent and contextually appropriate responses by reconstructing corrupted text sequences, thus making it a powerful choice for analysis. We used a base model which has 139 million parameters for experimenting.[11]
- **FLAN-T5:** An open-sourced model built upon the T5 architecture and additionally includes instruction-following datasets. It is optimized for better generalization and responsiveness to human-like prompts. The instruction-tuning helps it understand nuanced queries and provide more accurate and context-based responses which is one of the main reasons to choose this model for analysing. We took the small model with 80 million parameters.[12]
- **Microsoft-GODEL:** An open-sourced language model with lightweight architecture, focusing on efficiency and speed is trained on 551M multi-turn dialogues from Reddit discussion threads, and 5M instruction and knowledge-grounded dialogues. Suitable for real-time applications due to its reduced computational requirements. Its lower parameter count reduces latency in generating responses, facilitating smoother interactions. The choice of Godel in this study aims to evaluate its effectiveness in delivering counselling responses.[13]

Each model was chosen based on its architecture, parameter size, and suitability for real-time applications. The smaller versions of T5, FLAN-T5, and BART were selected to accommodate resource constraints while still utilizing their ability to handle natural language tasks. Godel, with its lightweight architecture, was included to assess its performance in delivering real-time responses in environments with limited computational power. This comparative approach provides insights into the trade-offs between model size, efficiency, and the quality of generated responses, particularly in scenarios where empathetic, context-aware interaction is crucial. A comparative summary of the models can be found in Table II.

TABLE II: Comparison of Language Models Used in the Study

Model	Size	Architecture
BART[11]	140M parameters	Encoder-Decoder (Transformer)
FLAN-T5[12]	80M parameters	Encoder-Decoder (Transformer)
T5[10]	60M parameters	Encoder-Decoder (Transformer)
Godel[13]	Approx. 100M parameters	Encoder-Decoder (Transformer)

IV. Results

A. Experiment Setup and Performance Metrics:

The setup includes loading the open-sourced models into the Kaggle platform to fine-tune them. Kaggle platform provides multiple GPU options and computing resources including the one we used 2 X NVIDIA T4 GPUs[36]. All four models with minimum size enabled us to take advantage of free resources provided by Kaggle with a limitation of 30 hours per week allowance for GPU access. All the parameters used are included below in Table III.

TABLE III: Hyper-parameters Used

Hyperparameter	Value
`output_dir`	`./results`
`per_device_train_batch_size`	8
`per_device_eval_batch_size`	8
`num_train_epochs`	50
`learning_rate`	5e-5
`eval_strategy`	`epoch`
`weight_decay`	0.01
`save_total_limit`	3
`save_strategy`	`epoch`
`logging_dir`	`./logs`
`fp16`	`True`
`dataloader_num_workers`	4
`gradient_accumulation_steps`	2

Key parameters included batch sizes of 8 for training and evaluation, a learning rate of 5e-5, 50 training epochs, and strategies for evaluation and saving models at each epoch. Additional settings, such as a weight decay of 0.01, mixed precision (`fp16`) training, a worker pool size of 4 for data loading, and gradient accumulation over 2 steps, ensured efficient training within the platform's constraints.

To assess the performance of the fine-tuned models, we used a test dataset comprising approximately 4,000 mental health counselling dialogues. The evaluation focused on comparing the model-generated responses against the reference responses in this test set, utilizing key metrics such as ROUGE, BLEU, and Perplexity scores.

For evaluating the performance of large language models (LLMs) for mental health counselling applications, we employ multiple metrics, including ROUGE[37] scores, BLEU[38] scores, and perplexity[39], to assess the effectiveness and reliability of the generated responses. ROUGE[37] (Recall-Oriented Understudy for Gisting Evaluation) scores are commonly used in natural language processing (NLP) to measure the overlap between generated text and reference responses, particularly focusing on recall metrics across various n-grams. ROUGE is valuable in our context as it helps evaluate the model's ability to capture essential information and align its output with expected counselling responses. BLEU[38] (Bilingual Evaluation Understudy) scores, on the other hand, assess the precision of n-gram overlaps between the generated and reference responses, providing insight into the linguistic fluency and relevance of the model's outputs. BLEU is especially useful in mental health counselling because it evaluates how closely the generated responses resemble natural, human-like responses, which is crucial for fostering rapport and empathy in counselling scenarios.

Perplexity[39] measures the degree of uncertainty in the model's predictions and is calculated as the inverse probability of the test set, normalized by the number of words. In mental health counselling, a lower perplexity score is ideal, as it indicates that the model can generate responses with higher confidence, contributing to more consistent and coherent interactions. Using ROUGE and BLEU, we gauge the informativeness and fluency of responses, while perplexity offers insight into the model's reliability and confidence. Together, these metrics provide a comprehensive evaluation framework, ensuring that the LLM can generate responses that are not only accurate but also contextually relevant, empathetic, and linguistically coherent for effective mental health support.

B. Performance Comparision:

Table IV highlights the performance metrics of various models evaluated on a mental health counselling dataset, focusing on their ability to generate empathetic, contextually appropriate, and semantically coherent responses. These models were assessed using ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores, which reflect their effectiveness in retaining essential linguistic features and psychological context crucial for therapeutic communication.

Among the models, BART demonstrated superior performance across all metrics, with scores of **0.4727** for ROUGE-1, **0.2665** for ROUGE-2, **0.3554** for ROUGE-L, and **25.3993** for the BLEU score. These results underscore BART's exceptional ability to emulate the significant language patterns required for empathetic engagement and effective counselling responses.

In comparison, GODEL achieved moderate performance, with a BLEU score of **6.6183** and ROUGE metrics that, while reasonable, fell short of BART's capabilities. FLAN-T5 and T5 exhibited relatively lower performance, with FLAN-T5 achieving a ROUGE-1 score of **0.2632** and a BLEU score of **3.0431**, while T5 scored **0.2585** for ROUGE-1 and **3.0649** for BLEU. These findings suggest that BART's fine-tuned architecture excels in capturing the linguistic subtleties and psychological depth necessary for generating responses that closely align with the supportive and reflective nature of counselling conversations, positioning it as a promising tool for mental health applications.

TABLE IV: Performance Metrics Comparison

Model	ROUGE-1 (Avg)	ROUGE-2 (Avg)	ROUGE-L (Avg)	BLEU Score (Avg)
T5	0.2585	0.0877	0.1914	3.0649
Flan-T5	0.2632	0.0954	0.1990	3.0431
BART	**0.4727**	**0.2665**	**0.3554**	**25.3993**
GODEL	0.3350	0.1324	0.2328	6.6183

The perplexity evaluation offers deeper insights into the models' performance, particularly their ability to generate linguistically coherent and psychologically meaningful responses essential for mental health support. Table V presents the generated text for a common question *("Can you help me with understanding how to deal with anxiety?")*, along with perplexity scores and an analysis of the responses.

Among the models, T5 achieved the lowest perplexity score of **1.78**, surpassing GODEL, BART, and FLAN-T5, which recorded scores of **3.40**, **3.75**, and **6.20**, respectively. The low perplexity score indicates T5's proficiency in generating smooth and grammatically accurate outputs; however, its tendency to produce repetitive and self-referential responses, such as reiterating the question itself which is evident in Table V, undermines its practical utility in offering therapeutic or actionable guidance, which is critical in counselling contexts.

GODEL, with a perplexity score of **3.40**, demonstrated potential in crafting empathetic and supportive responses, framing answers to anxiety-related inquiries in a comforting and understanding tone. Despite this strength, its outputs often lacked depth and specificity, limiting their applicability in scenarios requiring detailed therapeutic strategies. Similarly, FLAN-T5, which recorded a higher perplexity score of **6.20**, showed competence in generating brief coping strategies but struggled with fluency and psychological nuance, rendering it less effective for addressing complex emotional concerns.

BART, with a perplexity score of **3.75**, showcased a strong alignment between syntactic precision and psychological relevance, consistent with its superior performance in the ROUGE and BLEU metrics. Although its perplexity score was not the lowest, BART's ability to generate contextually appropriate and semantically rich responses positions it as a well-rounded model for mental health counselling. These findings reveal the unique strengths of the models: T5 excels in linguistic fluency but lacks psychological depth, GODEL and FLAN-T5 prioritize empathy but under-deliver on detail and fluency, and BART provides an optimal balance of syntactic precision and actionable, supportive content necessary for effective counselling applications.

Analysis of real-time response generation capability of the fine-tuned LLMs on real-life mental health questions

The responses generated by the fine-tuned models were evaluated for their relevance, coherence, and empathy when addressing questions related to anxiety and depression. The aim was to assess the ability of each model to provide supportive and contextually appropriate answers aligned with the principles of mental health counselling. Two sample questions were posed, as illustrated in Figures 3(a) and 3(b).

In response to the question, *"Is it normal for people to cry during therapy, or is it just me?"*, shown in Figure 3(a), the outputs of the models reveal notable differences. BART produced the most coherent and insightful response, effectively addressing the emotional context of the question while offering practical reassurance in a compassionate tone. This demonstrates BART's capability to deliver responses that align with the needs of mental health counselling. FLAN-T5 also generated an empathetic and contextually appropriate answer, but it repeated certain phrases, resulting in less fluidity compared to BART. In contrast, GODEL and T5 exhibited issues with repetitiveness and contradictions, detracting from their overall effectiveness. These issues highlight the limitations of these models in generating nuanced and well-structured responses.

Similarly, for the question, *"Give me a solution for my panic attacks,"* as shown in Figure 3(b), BART again provided the most practical and structured response. It offered specific coping mechanisms and actionable advice that would be valuable in a therapeutic setting. FLAN-T5 followed with a compassionate and supportive tone but lacked detailed strategies or solutions, making it less effective than BART in addressing the user's needs. GODEL produced a response that was brief and overly generic, failing to include specific advice or coping strategies. T5, on the other hand, struggled to maintain contextual relevance, often producing contradictory statements.

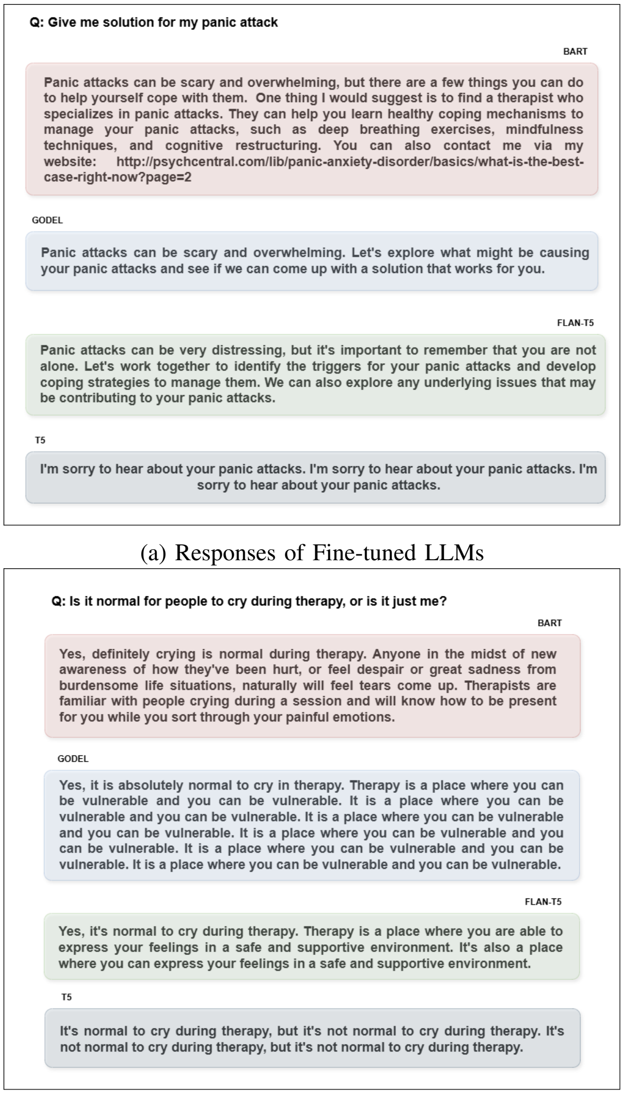
The comparative analysis of these models across the two prompts highlights distinct differences in their capacity to support mental health inquiries. BART consistently emerged as the most effective, delivering detailed, actionable, and emotionally supportive advice. While FLAN-T5 demonstrated empathy, it lacked the depth required for practical application. GODEL and T5 showed significant limitations, with issues such as redundancy, contradictions, and insufficient detail. These findings underscore the need for further refinement in models like GODEL and T5, whereas BART shows strong potential for applications requiring mental health support.

LIMITATIONS

While our study provides valuable insights into applying lightweight Large Language Models (LLMs) for mental health counselling, several limitations must be acknowledged. First,

TABLE V: Generated Response and Perplexity Comparison

Model	Generated Text (Summary)	Perplexity	Analysis
GODEL	”Anxiety can be a difficult thing to deal with... Let’s work together to identify the triggers...”	3.40	Empathetic response generation or human-like anxiety support.
BART	”This question is a great one! Depression is a treatable condition... Talk to a therapist... Practice deep breathing exercises...”	3.75	Structured actionable suggestions or comprehensive practical advice.
Flan-T5	”Anxiety can be a difficult emotion to manage... We can work together to develop coping strategies...”	6.20	Concise coping strategies or less detailed, higher perplexity.
T5	”Can you help me with understanding how to deal with anxiety? Can you help me with understanding how to deal with anxiety?”	**1.78**	Repetitive and nonsensical responses.



Q: Give me solution for my panic attack

BART

Panic attacks can be scary and overwhelming, but there are a few things you can do to help yourself cope with them. One thing I would suggest is to find a therapist who specializes in panic attacks. They can help you learn healthy coping mechanisms to manage your panic attacks, such as deep breathing exercises, mindfulness techniques, and cognitive restructuring. You can also contact me via my website: http://psychcentral.com/lib/panic-anxiety-disorder/basics/what-is-the-best-case-right-now?page=2

GODEL

Panic attacks can be scary and overwhelming. Let's explore what might be causing your panic attacks and see if we can come up with a solution that works for you.

FLAN-T5

Panic attacks can be very distressing, but it's important to remember that you are not alone. Let's work together to identify the triggers for your panic attacks and develop coping strategies to manage them. We can also explore any underlying issues that may be contributing to your panic attacks.

T5

I'm sorry to hear about your panic attacks. I'm sorry to hear about your panic attacks. I'm sorry to hear about your panic attacks.

(a) Responses of Fine-tuned LLMs

Q: Is it normal for people to cry during therapy, or is it just me?

BART

Yes, definitely crying is normal during therapy. Anyone in the midst of new awareness of how they've been hurt, or feel despair or great sadness from burdensome life situations, naturally will feel tears come up. Therapists are familiar with people crying during a session and will know how to be present for you while you sort through your painful emotions.

GODEL

Yes, it is absolutely normal to cry in therapy. Therapy is a place where you can be vulnerable and you can be vulnerable. It is a place where you can be vulnerable and you can be vulnerable. It is a place where you can be vulnerable and you can be vulnerable. It is a place where you can be vulnerable and you can be vulnerable. It is a place where you can be vulnerable and you can be vulnerable. It is a place where you can be vulnerable and you can be vulnerable.

FLAN-T5

Yes, it's normal to cry during therapy. Therapy is a place where you are able to express your feelings in a safe and supportive environment. It's also a place where you can express your feelings in a safe and supportive environment.

T5

It's normal to cry during therapy, but it's not normal to cry during therapy. It's not normal to cry during therapy, but it's not normal to cry during therapy.

(b) Responses of Fine-tuned LLMs

Fig. 3: LLMs Analysis for Mental Health Counselling Task

the datasets used in this research were sourced from various online repositories, including GitHub and Hugging Face, without thorough verification of their quality or generalizability. The lack of comprehensive data validation may limit the applicability of the models to a broader range of mental health contexts. Ensuring a more rigorously vetted and diversified dataset would be crucial for future work to enhance the robustness and effectiveness of the models.

Second in our study, we chose to fine-tune small, lightweight versions of the LLMs, allowing us to achieve efficient, resource-friendly results. These models showed promising outcomes; however, we recognize that larger, parameter-rich versions of models like BART and T5 may offer further performance gains. Additionally, by fine-tuning for a limited number of epochs, we could identify trends and patterns quickly, though extended training might reveal deeper insights. Future work with larger-scale models, prolonged training, and increased computational resources holds the potential for even stronger model performance and broader applicability across various datasets and contexts.

V. CONCLUSION

In this study, we explored the potential of various lightweight Large Language Models (LLMs) for AI-driven mental health counselling, fine-tuning four models—GODEL, BART, T5, and Flan-T5—on diverse counselling conversations. Among these, BART emerged as the best performer, consistently achieving the highest ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores, demonstrating its superior ability to generate coherent, contextually relevant, and linguistically accurate responses. Despite an average perplexity score, BART’s overall performance indicates that it is the most effective model for this application.

Moreover, the dataset we compiled from reputable online sources proved valuable in fine-tuning these pre-trained LLMs for mental health counselling tasks. This suggests that with further refinement and more rigorous data validation, the dataset can serve as a foundation for developing AI-powered mental health counselling tools. Our research highlights the potential for lightweight, accessible AI models to provide support in mental health interventions, paving the way for future work on more generalizable and scalable solutions in this sensitive domain.

REFERENCES

[1] "Large language model wikipedia," https://en.wikipedia.org/wiki/Large_language_model.

[2] "T5 model overview."

[3] "Flan-t5 model overview," https://huggingface.co/docs/transformers/en/model_doc/flan-t5.

[4] "bart model overview," https://huggingface.co/docs/transformers/en/model_doc/bart.

[5] "Who mental health report," https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report.

[6] "National mental health survey," https://indianmhs.nimhans.ac.in/.

[7] "National mental health survey," https://science.thewire.in/health/the-case-to-expand-psychiatric-education-for-mbbs-students/.

[8] "The rise of ai in mental health care," https://trendsresearch.org/insight/smart-therapy-solutions-the-rise-of-ai-in-mental-health-care/#:~:text=In%20the%20realm%20of%20mental,enhancing%20treatment%20accessibility%20and%20effectiveness.

[9] K. Denecke, A. Abd-Alrazaq, and M. Househ, *Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges*. Cham: Springer International Publishing, 2021, pp. 115–128. [Online]. Available: https://doi.org/10.1007/978-3-030-67303-1_10

[10] Google, "T5 small," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/google-t5/t5-small

[11] Facebook, "Bart base," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/facebook/bart-base

[12] Google, "Flan-t5 small," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/google/flan-t5-small

[13] Microsoft, "Godel v1.1 large seq2seq," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/microsoft/GODEL-v1_1-large-seq2seq

[14] S. D'Alfonso, "Ai in mental health," *Current Opinion in Psychology*, vol. 36, pp. 112–117, 2020, cyberpsychology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352250X2030049X

[15] X. Feng, M. Hu, and W. Guo, "Application of artificial intelligence in mental health and mental illnesses," in *Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*, ser. ISAIMS '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 506–511. [Online]. Available: https://doi.org/10.1145/3570773.3570834

[16] P. Darzi, "Could artificial intelligence be a therapeutic for mental issues?" *Science Insights*, vol. 43, no. 5, p. 1111–1113, Nov. 2023. [Online]. Available: https://www.bonoi.org/index.php/si/article/view/1209

[17] K. Shimada, "The role of artificial intelligence in mental health: A review," *Science Insights*, vol. 43, no. 5, p. 1119–1127, Nov. 2023. [Online]. Available: https://www.bonoi.org/index.php/si/article/view/1211

[18] H. Kim, M. Choi, and K. Kim, "A study on the design and effectiveness of a ai-based hyper-personalized mental health counseling system," *The Journal of Humanities and Social Science*, vol. 14, no. 3, pp. 3777–3790, 2023.

[19] A. Abd-Alrazaq, M. Alajlani, A. Alalwan, B. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: A scoping review," *International Journal of Medical Informatics*, vol. 132, p. 103978, Dec 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1386505619307166?via%3D

[20] E. Bendig, B. Erb, L. Schulze-Thuesing, and H. Baumeister, "The Next Generation: Chatbots in Clinical Psychology and Psychotherapy to Foster Mental Health – A Scoping Review," *Verhaltenstherapie*, vol. 32, no. Suppl. 1, pp. 64–76, 08 2019. [Online]. Available: https://doi.org/10.1159/000501812

[21] P. Rathnayaka, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, and R. Gray, "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring," *Sensors*, vol. 22, no. 10, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/10/3653

[22] E. van der Schyff, B. Ridout, K. Amon, R. Forsyth, and A. Campbell, "Providing self-led mental health support through an artificial intelligence-powered chat bot (leora) to meet the demand of mental health care," *Journal of Medical Internet Research*, vol. 25, p. e46448, Jun 19 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10337342/

[23] T. S. Sathe, M. A. Flitcroft, and A. N. Kothari, "Democratizing scientific and healthcare communication with large language models," *Cancer Research, Statistics, and Treatment*, vol. 6, no. 2, pp. 333–334, Apr–Jun 2023. [Online]. Available: https://journals.lww.com/crst/fulltext/2023/06020/democratizing_scientific_and_healthcare.36.aspx

[24] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, G. Lipori, D. A. Mitchell, N. S. Ospina, M. M. Ahmed, W. R. Hogan, E. A. Shenkman, Y. Guo, J. Bian, and Y. Wu, "A study of generative large language model for medical research and healthcare," *npj Digital Medicine*, vol. 6, no. 1, p. 210, Nov 16 2023. [Online]. Available: https://www.nature.com/articles/s41746-023-00958-w

[25] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, "Towards expert-level medical question answering with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.09617

[26] H. Yang, M. Li, H. Zhou, Y. Xiao, Q. Fang, and R. Zhang, "One llm is not enough: Harnessing the power of ensemble learning for medical question answering," *medRxiv*, 2023. [Online]. Available: https://www.medrxiv.org/content/early/2023/12/24/2023.12.21.23300380

[27] E. C. Stade, S. W. Stirman, L. H. Ungar, C. L. Boland, H. A. Schwartz, D. B. Yaden, J. Sedoc, R. J. DeRubeis, R. Willer, and J. C. Eichstaedt, "Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation," *npj Mental Health Research*, vol. 3, no. 1, p. 12, Apr 2 2024. [Online]. Available: https://www.nature.com/articles/s44184-024-00056-z

[28] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang, "Supporting the demand on mental health services with ai-based conversational large language models (llms)," *BioMedInformatics*, vol. 4, no. 1, pp. 8–33, 2024. [Online]. Available: https://www.mdpi.com/2673-7426/4/1/2

[29] M. Yadav, N. K. Sahu, M. Chaturvedi, S. Gupta, and H. R. Lone, "Fine-tuning large language models for automated diagnostic screening summaries," 2024. [Online]. Available: https://arxiv.org/abs/2403.20145

[30] A. Li, Y. Lu, N. Song, S. Zhang, L. Ma, and Z. Lan, "Understanding the therapeutic relationship between counselors and clients in online text-based counseling using llms," 2024. [Online]. Available: https://arxiv.org/abs/2402.11958

[31] Z. Zheng, L. Liao, Y. Deng, and L. Nie, "Building emotional support chatbots in the era of llms," 2023. [Online]. Available: https://arxiv.org/abs/2308.11584

[32] Aditya, "Mental health counselling dataset," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/datasets/Aditya149/Mental_Health_Counselling_Dataset

[33] A. Pandey, "Chatbot for mental health: Dataset," 2023, accessed: 2024-10-22. [Online]. Available: https://github.com/pandeyanuradha/Chatbot-for-mental-health/blob/main/Dataset/mentalhealth.csv

[34] N. Bertagnolli, "Counsel chat: A dataset for counseling dialogues," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/datasets/nbertagnolli/counsel-chat

[35] Amod, "Mental health counseling conversations," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/datasets/Amod/mental_health_counseling_conversations

[36] D. Becker, "Running kaggle kernels with a gpu," 2023, accessed: 2024-10-22. [Online]. Available: https://www.kaggle.com/code/dansbecker/running-kaggle-kernels-with-a-gpu

[37] H. Face, "Rouge metric," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/spaces/evaluate-metric/rouge

[38] GeeksforGeeks, "Nlp - bleu score for evaluating neural machine translation in python," 2023, accessed: 2024-10-22. [Online]. Available: https://www.geeksforgeeks.org/nlp-bleu-score-for-evaluating-neural-machine-translation-python/

[39] H. Face, "Perplexity in transformers," 2023, accessed: 2024-10-22. [Online]. Available: https://huggingface.co/docs/transformers/en/perplexity