2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)

2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N) | 979-8-3503-5681-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAC2N63387.2024.10895249

Expert Patient Interaction Language Model (EPILM)

Pratham Gala
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J Sanghvi College of Engineering
Mumbai, India
galapratham03@gmail.com

Pooja Vartak
Assistant Professor of Computer Science
Dept. of Computer Science and Engineering (Data Science)
Mumbai, India
pooja.vartak@djsce.ac.in

Jay Vithlani
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J Sanghvi College of Engineering
Mumbai, India
jvithlani4@gmail.com

Naman Chheda
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J Sanghvi College of Engineering
Mumbai, India
namanchheda3@gmail.com

***Abstract* — Today's youth encounter numerous mental health challenges, many of which can be effectively addressed through professional intervention. However, there is a notable reluctance among young individuals to seek professional mental health services. The integration of language models in providing mental health support offers a non-intrusive and accessible approach for young people to engage with professional assistance for their mental health concerns. Although existing language models like GPT 3.5, Falcon-7B model have been effective, they also present certain challenges like high dependability of data from sources like Reddit, computational obstacles hindering access, and a preference for certain models over superior alternatives have impeded advancements in this area. that require thorough examination. Issues related to biases and the model's performance across diverse scenarios underscore the necessity for more intelligent strategies. The proposed method involves enhancing the Falcon-7B model through the application of quantization low rank adaptation (Q-LoRA). This technique reduces the computational requirements and simplifies the accessibility for individuals equipped with the necessary technology while providing the average bertSCORE 85%. By addressing these issues, paper's objective is to refine mental health support systems to benefit a broader audience.**

***Index Terms*—Fine-tuning techniques, Data credibility, Reddit, Computational inefficiency, Falcon-7B model, Quantization low rank adaptation.**

I. INTRODUCTION

The global landscape of mental health is experiencing a significant shift, with an alarming increase in mental health issues across diverse populations [11]. The aforementioned trend highlights the pressing want for inventive approaches to improve the availability and efficiency of mental health treatments. Technology-driven mental health platforms have developed as viable options in response to this expanding demand, especially those that make use of machine learning and natural language processing (NLP). Among these technological interventions, chatbots are one of the most recognized examples and provide a new way in which to represent therapist-patient dialogue, simulating it and making big data transformed into personal exchanges. For this AI conversational agent, stigma in the process of help-seeking may be reduced, support provided at all hours, and a scaleable response to the shortage of specialist workers in mental health offered [16].

But regardless of its promise, there are a number of serious challenges facing current methodologies in chatbots, most notably those of biased responses, usually because of limited or prejudiced training sets. This would then make a chatbot, as such, inept as a tool for mental health support if some advice bases for such biases are inappropriately given and thus unsafe. Rigidity is another major constraint with several of the current systems. Chatbots should be able to assimilate and put into practice new information with the change of people's awareness in mental health and emergence of new therapy modalities [2,8]. Unfortunately, these are all aspects that many systems in place today do not integrate since they lack the flexibility required to keep up with the most recent advances in treatment for mental health. An additional challenge to wide usability is that advanced language models generally have a huge computing demand. However, most potential effects remain minimized because many of these mental health organizations and researchers are beyond reach by the very hard costs and technical competence required for training and deployment [14, 15].

In particular, what is urgently needed in tailored therapeutic approaches is hybrid chatbot systems that can be embedded with wide functionalities. Such systems try to balance adaptability with personalization by offering subtle and contextually appropriate responses to users' mental health concerns.[3] The promise of recent developments may hold toward overcoming part of these limitations by using automated data compilation and model enhancement techniques. Besides Q-LoRA, other fine-tuning methodologies have shown to be effective and point to a future where mental health chatbots could both have significantly improved computational efficiency and accessibility. Another trend is the quality and diversity of training data, which has been gaining interest in recent times. Datasets should be curated from reliable sources, such as real-world therapist-patient interactions, to enable the development of more reliable and effective chatbots.[11] These also will contribute toward reducing biases and increasing the accuracy and relevance of the chatbot responses. Research in this domain is beginning to show that for developing successful chatbots in mental health, a multidisciplinary effort is warranted. Besides user-experience design, it goes into advances in NLP and machine learning with a deep understanding of psychological concepts and ethical issues [16].

Indeed, the authors review the current status of mental health chatbots, which hold promises but at the same time do not yet reflect their potentially major benefits; very real are their current limitations. This study investigates novel approaches to model construction, data curation, and system design that may serve in creating AI-driven mental health support systems which function more effectively, ethically, and ubiquitously. The preceding discussion, while certainly not intended to address all the challenges facing mental health, presents an attempt to contribute toward the development of useful chatbot augmentations to conventional mental health services, with the ultimate goal of improving the quality and access to mental healthcare for citizens all over the world.

Generally, the paper is organized so that an overview of EPILM is given. Furthermore, it provides an extended literature review based on recent works in the areas of language modeling and AI-powered mental health support systems. This chapter details the methodology of the new approach of using the Quantization Low-Rank Adaptation techniques and application of Sharded architecture for fine-tuning the Falcon-7B model. Results and Analysis: This section compares EPILM with larger models in terms of BertScore and Sentence Transformer Score, among others, such as GPT-3.5 and Gemini. Conclusion: Summarizing the major contributions that have been made possible with respect to mental health assistance, this model has proven effective and available. Future scope will deal with what kind of enhancements and newer uses can be pursued in the near future with EPILM: integration into pre-existing mental health platforms and further enhancement of the model. Finally, an exhaustive reference list is included for proper citation and further research in this rapidly evolving field.

II. LITERATURE REVIEW

Xuhai Xu, Bingsheng Yao, et al. suggested a huge language model in 2023 that makes use of datasets from Reddit, including SDCNL (Suicide vs. depression classification dataset), Deraddit (10 subreddits on abuse, PTSD, anxiety, etc.), and community discussion forums like r/depression. Using zero-shot prompting, Alpaca and Google Flan-T5 were adjusted on four datasets as part of the methodology. More contextual information about the task and input can consistently increase performance, as demonstrated by the analysis, with dialogue-focused models with bigger trainable parameters outperforming Flan-T5 [1]. The study did, however, have some drawbacks, including low data trustworthiness from using social media in place of qualified specialists, computational inefficiency (Alpaca was trained on eight 80GB A100s for three hours)[15], and extremely biased results with little ability to be generalized. In addition, even though there were better models on the OpenLLM leaderboard, only Alpaca and Flan-T5 were optimized.

In 2022 M. A. N. Siddik, S. Arifuzzaman, and A. Kalam proposed a chatbot which used a dataset from Zenodo, comprising 50,000 statements from Reddit conversations about mental health across 28 subreddits. The methodology started with logging user activity across social media platforms like Twitter, WhatsApp, and Facebook. Various models, including BERT, were used for sentiment classification based on user data, and a chatbot was implemented to conduct the GAD-7 questionnaire and respond with relevant domain knowledge. The analysis showed accuracy rates of 74% with Conv-LSTM (Fasttext embedding method), 75% with Sandwich (Fasttext embedding method), and 83% with BERT (word2vec) [2]. Limitations included output restricted to sentiment classification and the chatbot providing inhuman responses that were difficult to understand and implement.

In 2023 Desiree Bill, Theodor Eriksson proposed a finetuning technique which used a dataset of 4,000 therapy question-answer pairs from counselchat.com, containing expert-client conversations on topics like depression and anxiety. The methodology involved selecting a large language model (LLM) with frozen weights, generating outputs from the raw pre-trained model, combining them with the reward model output, and applying the PPO update rule iteratively [3]. The study did not provide detailed result analysis or specific limitations.

In 2021 S. Prabakeran, S. Suryavanshi proposed a chatbot for women which utilized datasets from r/depression_help, r/Stress, and r/talktherapy. The methodology involved building two fully connected LSTM layers and implementing seq2seq learning with attention using an encoder-decoder architecture and the adam optimizer. The analysis achieved a BLEU score of 0.65 [3]. However, the outputs were limited and redundant due to constrained weights and parameters.

In 2018 Inkster, Sard, and Subramaniam discussed how chatbots could have a positive impact on mental health. They proposed that the platforms could bridge the gap that exists between the lack of professionals and those in need [16]. They used Wysa, an AI chatbot program, to investigate a random group of people. After messaging, the participants filled out an online form to describe symptoms of depression. Wysa usage led to the emergence of two user categories: high and low users.

Both groups demonstrated improvement when scores were compared before and after the 21-day research period, although the high user group showed a more noticeable improvement. The study's findings regarding the application of AI in mental health were encouraging, but they also highlighted the need for more research to back up the findings [15].

It is imperative to tackle the existing constraints in the domain of AI-powered mental health assistance systems in order to create more dependable, expandable, and morally sound solutions. The computationally inefficient and costly fine-tuning of pre-trained models is one of the key obstacles, and other issues include doubts regarding the variety, accuracy, domain knowledge, real-time interactions, and quality of the data that is taken from social media platforms [1, 2, 4]. Furthermore, generalized learning for a wide range of mental health conditions has a noticeable gap. Existing approaches frequently terminate at model outputs and don't provide reliable therapist recommendations in response to user inquiries [9, 10]. Initiatives such as the Expert-Patient Interaction Language Model (EPILM) are being developed in response to these problems. The aims of these efforts are to revolutionize the delivery of mental health care by bridging the gap between AI-driven support systems and credible, personalized therapeutic recommendations by integrating advanced natural language processing techniques [11–13], improving data quality and diversity, and involving expert knowledge.

Gap Analysis:

- Need for more resource-efficient model adaptation techniques.
- Requirement for more reliable and diverse data sources for training AI models.
- Need to integrate expert knowledge more effectively into AI systems.
- Development of more responsive and dynamic AI interaction models.
- Development of more tailored and individualized AI responses.

III. METHODOLOGY

The fields of machine learning and natural language processing are redefining mental health support systems. This research introduces a novel approach using Quantization Low-Rank Adaptation (Q-LoRA) to fine-tune the Falcon-7B model, aiming to create an efficient and accessible mental health chatbot. The methodology encompasses innovative techniques for model optimization and training on expert-patient interaction data.

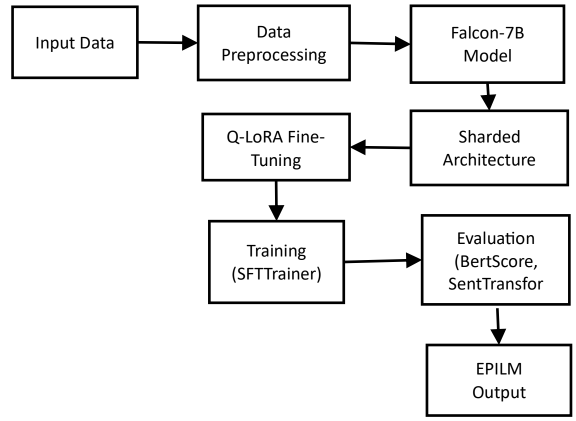


Fig. 1. WorkFlow stages

A. Dataset Description:

The model uses scraped data from various credible and legitimate sources facilitating real case-studies. Counsel-Chat is an example of an `expert community which offers services by licenced therapists. Clients submit queries via Councel-Chat, and therapists respond; users can favorite the answers they find most useful.

B. Data Preprocessing:

Patient names are included in this data. Therapists can promote their clinics on CounselChat.com by offering good, freely accessible counsel to the public. The names of the therapists were retained in the initial dataset. The model uses a lexicon dictionary to remove names from the dataset, considering them as 'stop words' [18]. WebMD, Mayo Clinic and Heatlhline.com - One text exchange between a patient and their therapist is included in this dataset. The dataset was assembled from online FAQs, WebMD, Mayo Clinic, and HealthLine, among other well-known healthcare blogs. The names of the patient and therapist have been eliminated from all questions and replies, and unnecessary characters have been pre-processed [2-4].

C. Model Architecture:

The suggested approach entails utilizing a Sharded architecture to fine-tune the Falcon-7B model by dividing the model and hyperparameter tweaking over several GPUs for parallel computation. Sharded design expedites the training process and enhances overall functionality. The Sharded architecture includes several key components:

1. Sharded Model: The Falcon-7B model is divided into smaller segments and distributed across multiple GPUs, ensuring concurrent training and faster learning.

2. Tokenizer: It converts words into individual tokens which are then distributed to GPUs for parallel processing.
3. Data Loading and Computation: Data is loaded in 4-bit precision, and computations are performed in 16-bit precision to optimize memory usage which also speed up training.

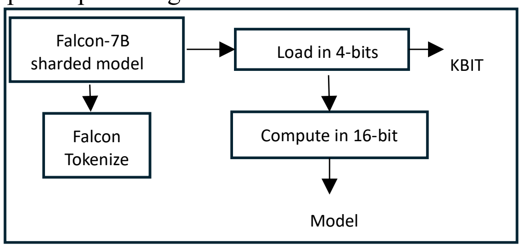


Fig. 2. Sharded Accelerate and Distribute among GPUs

4. Learning Rate: The learning rate, determined by the SFTTrainer, is distributed across all GPUs for consistent training.
5. Model and Hyperparameter Tuning: The SFTTrainer optimizes the model and hyperparameters, performing operations like gradient clipping, weight decay, and learning rate scheduling to enhance performance.

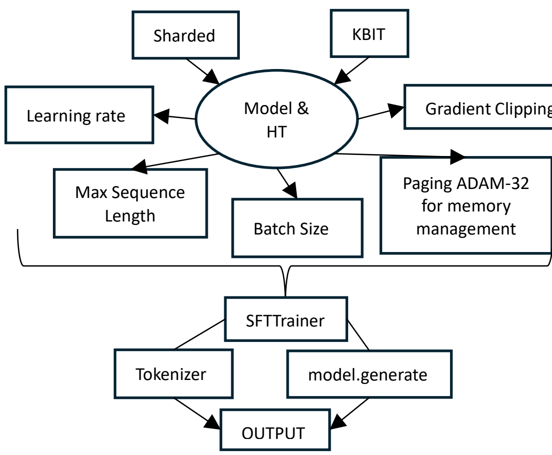


Fig. 3. Model hyperparameters

6. Output and Inference: Once trained, the model generates output using the model generate method and supports knowledge-based inference training (KBIT) to improve output quality.

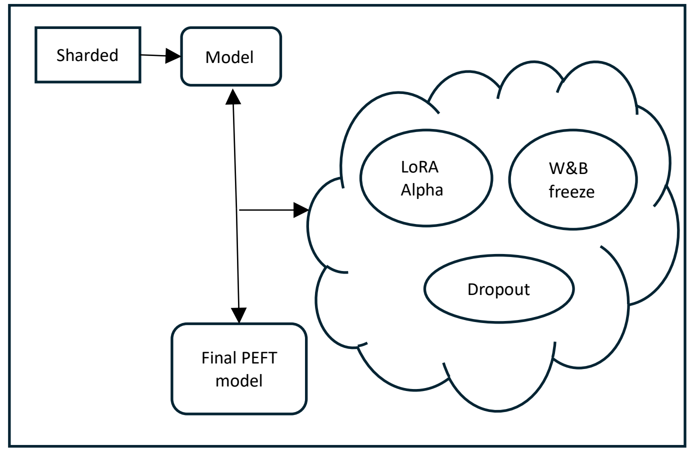


Fig. 4. Knowledge-Based Inference Learning (KBIT)

1) Low-Rank Adaptation (LoRA): The core idea of LoRA is to represent the weight updates during fine-tuning as low-rank decompositions. For a pre-trained weight matrix W ∈ R^(d×k), the update is represented as:

$$W' = W + \Delta W \qquad \Delta W = BA \tag{1}$$

Where:

- B ∈ R^(d×r) and A ∈ R^(r×k) are low-rank matrices
- r is the rank, typically much smaller than d and k

In addition, Quantized Low-Rank Adaptation (Q-LoRA) emerges as a refined solution to optimize deep learning models, specifically Transformer architectures. Q-LoRA builds on Low-Rank Adaptation (LoRA), introducing quantization techniques to reduce memory consumption while maintaining or enhancing model performance. Q-LoRA incorporates several enhancements:

1. 4-bit NormalFloat: Q-LoRA uses 4-bit NormalFloat, an optimal quantization data type for normalized data, offering significant memory savings and superior results compared to traditional quantization methods. For 4-bit NormalFloat:

$$Q(x) = round\ (x / \Delta) * \Delta \tag{2}$$

Where:

- x is the original float value
- Δ is the step size (quantization scale)
- Round () is the rounding function
- The step size Δ is determined based on the standard deviation σ of the weight distribution:

$$\Delta = 2\sigma / (2^{\wedge}(b-1) - 1) \tag{3}$$

Where b is the number of bits (4 in this case).

2. Double Quantization: This technique quantizes the quantization constant, providing additional memory savings and maximizing efficiency without compromising accuracy. If q represents the quantized weights and s the quantization scale: q = round (W / s) Double quantization introduces another quantization step:

$$s' = round\ (s\ /\ s\_s)\ *\ s\_s \quad (4)$$

$$q' = round\ (q\ *\ s\ /\ s') \quad (5)$$

Where s_s is the scale factor for the scales.

3. Paged Optimizers: Utilizing NVIDIA unified memory, paged optimizers facilitate seamless transfers between CPU and GPU, mitigating memory spikes during gradient checkpointing and enhancing scalability and performance. They use a form of gradient checkpointing:

$$\partial L/\partial \theta = \sum (\partial L/\partial y\_i) * (\partial y\_i/\partial \theta) \quad (6)$$

Where θ represents the model parameters, and y_i are intermediate activations.

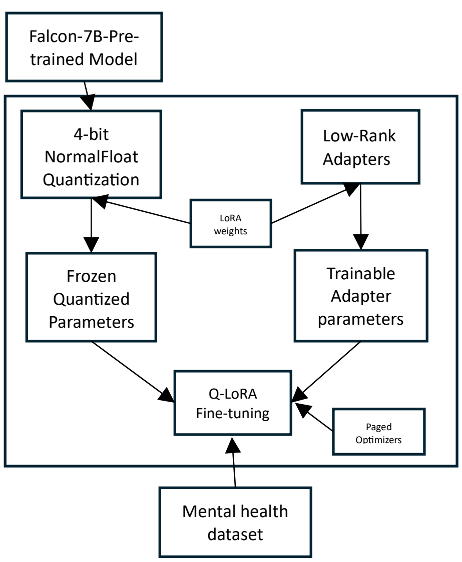


Fig. 5. Quantization Low Rank Adaptation

2) Loss Function and Optimization: The model is trained to minimize a loss function L, typically cross-entropy for language models:

$$L = -\sum y\_i\ log(p\_i) \quad (7)$$

Where y_i are the true labels and p_i are the model's predictions. The optimization process updates only the LoRA parameters A and B:

$$A\_t + 1 = A\_t - \eta * \partial L/\partial A \quad (8)$$

$$B\_t + 1 = B\_t - \eta * \partial L/\partial B \quad (9)$$

Where η is the learning rate.

D. Training and Evaluation:

Memory-Efficient Fine-tuning Process:

a) Quantization of Falcon-7B Base Model:

$$Wq = Q(W) \quad (10)$$

Where: W are the original Falcon-7B weights Wq are the quantized weights

b) LoRA Rank Selection: Choose r << min(d,k) to maintain low rank

c) Initialization of LoRA matrices:

$$A \sim N(0, \sigma^2)\ B \sim N(0, \sigma^2) \quad (11)$$

Where: σ^2 = 1/√r

d) Forward Pass:

$$y = f(x;\ Wq + Q(BA)) \quad (12)$$

Where: f is the Falcon-7B architecture x is the input

e) Backward Pass and Update: Compute ∂L/∂A and ∂L/∂B using backpropagation Update only A and B using the optimization step

Fine-tuning Procedure:

a) Load quantized Falcon-7B model

b) Initialize LoRA matrices A and B

c) For each batch in the dataset:

- Perform forward pass
- Compute loss
- Perform backward pass
- Update A and B using paged optimizer d) Repeat for specified number of epochs

During inference, combine the quantized base model with the fine-tuned LoRA parameters:

$$W\_inference = Wq + Q(BA) \quad (13)$$

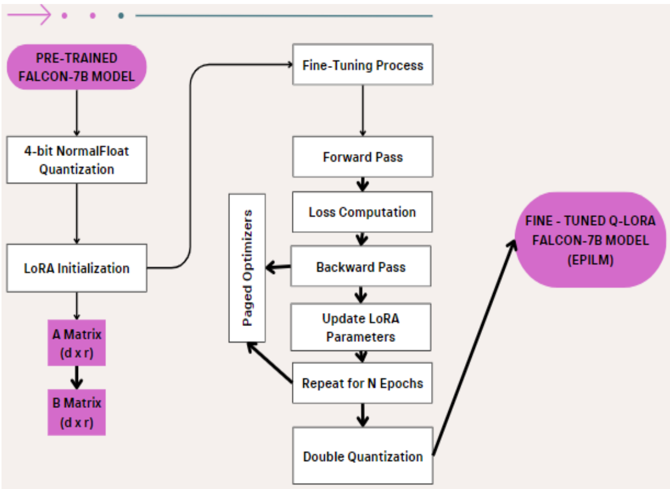
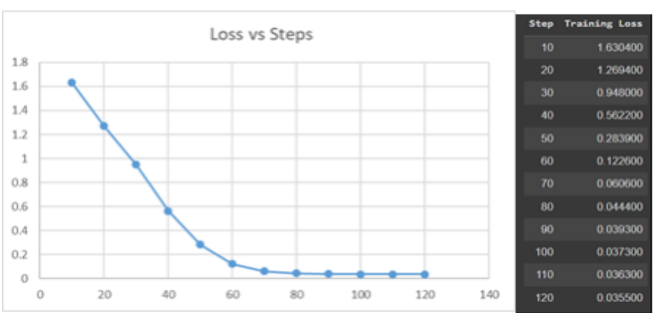


Fig. 6. Q-LoRA Finetuning

Comparison of three different language models – GPT-3.5, Gemini, and EPILM – across two key characteristics: model size and training data size [1]. GPT-3.5 has a massive 175 billion parameter model size, having been trained on an enormous 45 terabytes of data. Gemini is slightly smaller at 137 billion parameters but was still trained on a very large 750 gigabytes of data. In contrast, EPILM is a much smaller 7 billion parameter model. However, the notable aspect of EPILLM is that it underwent "fine-tuning" on just 6 megabytes of data, an incredibly small training dataset compared to the others [6].

Parameter 1: Loss Function: EPLIM used a paged_adamw_32bit optimizer for additional memory efficiency, and correspondingly Cross-Entropy Loss function [5].



Step	Training Loss
10	1.630400
20	1.269400
30	0.948000
40	0.562200
50	0.283900
60	0.122600
70	0.060600
80	0.044400
90	0.039300
100	0.037300
110	0.036300
120	0.035500

Fig. 7. Plots of Training and Validation losses over epochs

Parameter 2: BertScore: identifies words in candidate and reference phrases using cosine similarity, utilizing the pre-trained contextual embeddings from BERT [13]. It has been demonstrated to correlate with human judgment in evaluations at the sentence and system levels.

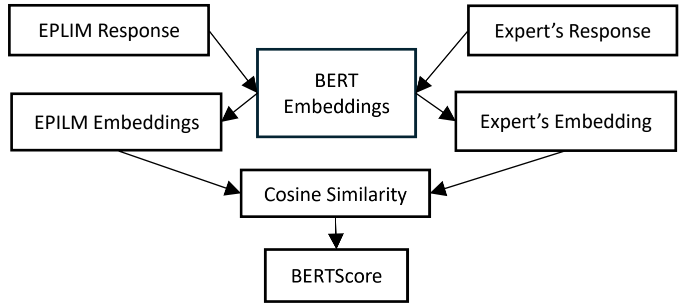


Fig. 8. BertSCORE

Parameter 3: Sentence Transformer Score Currently, the most easy and successful approach involves encoding phrases using a strong model to obtain their embeddings, and then computing the similarity score using a similarity metric [13]. The similarity score indicates how similar or different the meanings of the two texts are from one another.

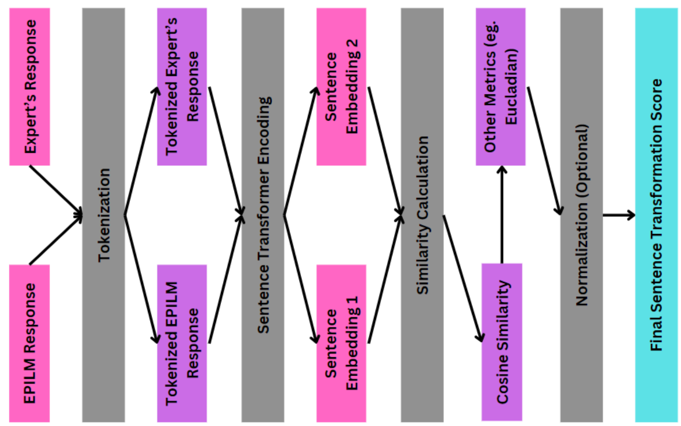


Fig. 9. Sentence Transformer Model [13]

Table. 1. Q-LoRA fine-tuning BertScore using different optimizers for Falcon-7B

Optimizer Name	BertScore (%)
AdamW	**84.79**
AdaFactor	78.32
Lion	82.15
SGD	81.93

Table. 2. AdamW optimizer with different learning rates for Q-LoRA fine-tuning of Falcon-7B

Learning Rate	BertScore (%)	Loss
1e-3	79.65	2.25
5e-4	81.36	1.96
1e-4	83.12	1.72
5e-5	**84.79**	**1.59**
1e-5	83.45	1.81
5e-6	82.78	1.95
1e-6	81.92	2.16

5e-7	80.66	2.38
1e-7	79.58	2.57

IV. RESULTS AND DISCUSSION

Assessing the potential impact of AI-driven mental health support systems requires evaluating their performance. Using metrics like BertScore and Sentence Transformer Score, this section compares the performance of the Expert Patient Interaction Language Model (EPILM) against larger models like GPT-3.5. The results provide insight into the efficacy and efficiency of EPILM in mental health discussions.

Table. 3. Overview of Models

	GPT-3.5	Gemini	EPILM
Model Size	175B	137B	7B
Training Data	45TB	750B	6MB (Fine Tuning)

Table. 4. Comparison of EPILM and GPT-3.5

Questions	GPT-3.5		EPILM	
	bertSCORE	SentTr. score	bertSCORE	SentTr. Score
1.What is panic attack?	77.45	78.34	85.22	85.96
2.What are symptoms of panic attack vs anxiety attack?	84.83	81.91	85.22	84.53
3.What are types of mental illness?	73.72	79.28	83.75	77.82
4.What does mental illness mean?	80.24	82.29	85.21	70.72
5.How can you treat mental illness?	89.09	85.23	84.55	62.71

V. CONCLUSION

In fact, LLMs have not achieved universal application yet due to the high training cost, not to mention tuning them with relevant information regarding mental health. EPILM makes use of a few state of art techniques in the process of solving this problem: model splitting, paging, quantization, LoRA. Such methods fine-tune the huge Falcon-7B model on a Google Colab GPU. Very few of these clever techniques save memory, hence it can only be trained on one extra GPU which has very little resources. In addition to that, EPILM is built using a high-quality dataset obtained from several reliable sources. Fine-tuning the Falcon-7B model on this dataset with QLoRA and PEFT methods generated a unique model concerned with mental health conversations. This fine-tuned model entails precise and insightful recommendations, hence extracting more knowledge about mental health concepts and symptoms and ways of coping with them. This novelty will increase the use of mental health chatbots, culminating in high performance. EPILM takes mental health chatbots to every corner of the world by use of opensource and free technology. It enables people across the world to seek aid in an easy and flawless manner. The EPILM is a big leap toward making mental health disorder chatbots trusted for self-care and accessible to everyone in the field for proper self-care or calling out for help where necessary.

VI. FUTURE SCOPE

Keeping this in mind, this study demonstrated the way large language models could be fine-tuned with respect to picked mental health datasets like Falcon-7B with the use of different techniques including QLoRA and PEFT. In light of the foregoing, several areas for future improvement were identified:

1. Dataset Refinement: With increased refinement and the addition of more detailed and diverse samples to the mental health dataset, the model will hold the capability to handle a wide variety of queries and scenarios related to mental health.
2. Prompt Engineering: Understand fancy prompt engineering tricks like few-shot learning, prompt tuning, and prompt augmentation that greatly improve model performance and give contextually relevant responses.
3. Transfer learning: Fine-tuned model transfer, where possible, into subcategories or groups of mental health for personalization in conversational speech.
4. Ethical Consideration: Direct strict ethical standards that will help mitigate biases and privacy while deploying the use of conversational AI for practical mental health use cases.

REFERENCES

[1] K. Yang, S. Ji, T. Zhang, et al., "Towards Interpretable Mental Health Analysis with Large Language Models," Conference on Empirical Methods in Natural Language Processing, 2023.

[2] M. A. N. Siddik, S. Arifuzzaman, and A. Kalam, "Psyche Conversa - A Deep Learning Based Chatbot Framework to Detect Mental Health State," in 2022 10th International Conference on Information and Communication Technology (ICoICT), IEEE, 2022, pp. 146-151.

[3] S. Prabakeran, S. Suryavanshi, "Women's Mental Health Chatbot Using Seq2seq With Attention," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 10, pp. 6873-6881, 2021.

[4] 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, 2019.

[5] T. Dettmers, P. Pagnoni, A. Holtzman, et al., "QLoRA: Efficient Finetuning of Quantized LLMs," 37th conference on Neural Information Processing Systems (NeurIPS), 2023.

[6] E. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR 2022.

[7] A. Ghandeharioun, D. McDuff, M. Czerwinski, et al., "EMMA: An Emotion-Aware Wellbeing Chatbot," in 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 1-7.

[8] K. Oh, D. Lee, B. Ko, and H. Choi, "A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation," in 2017 18th IEEE International Conference on Mobile Data Management (MDM), IEEE, 2017, pp. 371-375.

[9] A. Rakib, E. Rumky, A. Ashraf, M. Hillas, and M. Rahman, "Mental Healthcare Chatbot Using Sequence-to-Sequence Learning and BiLSTM," in Advances in Intelligent Systems and Computing, 2021, pp. 378-387.

[10] T. Sharma, J. Parihar, and S. Singh, "Intelligent Chatbot for Prediction and Management of Stress," in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2021, pp. 937-941.

[11] I. Hungerbuehler, K. Daley, K. Cavanagh, H. Garcia Claro, and M. Kapps, "Chatbot-Based Assessment of Employees' Mental Health: Design Process and Pilot Implementation," JMIR Formative Research, vol. 5, no. 4, p. e21678, 2021.

[12] F. Patel, R. Thakore, I. Nandwani, et al., "Combating Depression in Students using an Intelligent ChatBot: A Cognitive Behavioral Therapy," in 2019 IEEE 16th India Council International Conference (INDICON), IEEE, 2019, pp. 1-4.

[13] X. Sun, Y. Meng, X. Ao, F. Wu, T. Zhang, J. Li, and C. Fan, "Sentence Similarity Based on Contexts," Transactions of the Association for Computational Linguistics, vol. 10, pp. 573-588, 2022.

[19] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassami, Anind K. Dey and Dakuo Wang, "Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data", Proceedings of ACM on Interactive, Mobile, wearable and Ubiquitous Technologies (IMWUT) 2024.

[20] Desiree Bill, Theodor Eriksson, "Fine-tuning a LLM Using Reinforcement Learning from Human Feedback for Therapy Chatbot Application", TRITA-EECS-EX; 2023

[21] B. Inkster, S. Sarda, V. Subramanian, et al., "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study", JMIR mHealth and uHealth, vol. 6, no. 11, e12106, 2018.

[22] Edward J. Hu, Yelong Shen, Philip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models", ICLR 2022 Poster

[23] Nicolas Bertagnolli, 2020, N. Counsel Chat: Bootstrapping High-Quality Therapy Data. https://towardsdatascience.com/counsel-chat-bootstrapping-high-quality-therapy-data-971b419f33da

[24] Arjun Gupta, Himanshu, Aeshna Singh, Palash Baranwal, 2021, TSAI Group Assignment, https://towardsdatascience.com/?gi=bc9a124e96df