Smaller Models, Deeper Impact: Empathy-Driven AI for Low-Resource Mental Health Counseling

Abstract

Access to quality mental health care remains a global challenge, especially in underserved regions. While large language models (LLMs) show promise in providing AI-driven therapeutic support, their massive size often limits practical deployment due to high computational and memory requirements. This study presents a fine-tuned version of Phi-3, a compact yet powerful language model, tailored specifically for empathetic mental health counseling. By leveraging Low-Rank Adaptation (LoRA) and the Unsloth framework, we significantly improved response relevance and emotional sensitivity, while achieving a 2x increase in training speed and reducing memory consumption by 70%.

The model was fine-tuned on three diverse dataset of publicly available mental health dialogues, enabling it to generate supportive and context-aware responses. Evaluation using BERTScore metrics demonstrated strong performance, with an F1-score of 0.8429 during training and 0.8385 on unseen test data.

Our results suggest that small-scale, efficiently fine-tuned LLMs like Phi-3 can offer accessible, accurate, and scalable mental health support—bridging the gap between technological capability and real-world usability for the average user.

Keywords: Mental health, Anxiety, Depression, Stigma, Small Language Models, AI-driven Therapy, Fine-tuning, Emotional Intelligence, Personalized Interaction, Digital Mental Health, AI-driven Therapy, Low-Resource Deployment.

1 Introduction

Mental health disorders, such as anxiety and depression, have the highest prevalence rates globally, surpassing other major health conditions like cardiovascular diseases and stroke. According to the World Health Organization (WHO), approximately 25% of people worldwide will experience a mental health issue in their lifetime [19], a staggering statistic compared to the 6.2% prevalence of cardiovascular diseases [25] and the 5% prevalence of stroke [22]. The situation has worsened in recent years due to global crises, such as the COVID-19 pandemic, wars, and economic downturns, which have led to a 25.6% increase in anxiety and a 27.6% rise in depression since 2020 [9]. Despite the growing prevalence of mental health disorders, access to medical treatment remains alarmingly low. In the United States, only 36.9% of adults with a mental health disorder received treatment in the past few year [20], and this rate is even lower for teenagers. The situation is dire in low- and middle-income countries, where over 75% of people with severe mental health disorders do not receive any form of care [20].

The primary barrier to delivering mental health care is stigma, which refers to the negative attitudes and discrimination directed at individuals with mental illnesses. This stigma creates a reluctance to seek help, as revealed by a comprehensive global survey in which 60% of participants were hesitant to pursue professional treatment due to fear of stigmatization[9]. Furthermore, the diagnostic and therapeutic processes for mental illness often require substantial user cooperation through conversation, which is difficult to achieve given the high non-response rates driven by stigma. As a result, many individuals opt to self-diagnose and seek help through the internet, where they feel safer and stigma-free. However, existing online mental health tools often rely on closed-ended questions and lack personalized interaction, continuous monitoring, and tailored interventions.

These widespread barriers to mental health care, coupled with the chronic and long-term nature of mental illnesses, highlight the urgent need to address this significant public health challenge. One promising approach is leveraging artificial intelligence (AI) to provide accessible mental health support. AI-driven solutions like fine-tuned language models offer opportunities to overcome the stigma-related barriers by enabling anonymous, personalized, and continuous interaction with users.

Among these innovations, the fine-tuning of large language models (LLMs) like Phi-3.5 [1] has garnered attention for its potential to enhance empathy and personalization in specific tasks. Fine-tuning LLMs is a highly effective strategy for improving performance and modifying behaviours to align with desired outcomes. Low-Rank Adaptation (LoRA) [11] is one of the most widely adopted methods for this purpose, demonstrating significant promise in enabling smaller, specialized models to outperform larger, generalized models on specific tasks while utilizing a fraction of the trainable parameters. This challenges the notion that larger models are always superior, highlighting the potential of fine-tuning to deliver tailored solutions in specialized domains, such as mental health counseling.

The Phi-3.5 small language model represents a significant advancement in the field of AI. Built on a transformer architecture, Phi-3.5 is designed to understand and generate human-like text efficiently. Its smaller size allows for rapid deployment and reduced operational costs, making it an attractive choice for applications in mental health. However, effectively fine-tuning such models to ensure they comprehend user inputs and respond in ways that resonate emotionally with individuals seeking support remains a challenge.

Simultaneously, the biomedical field is experiencing explosive growth in textual data, with resources like PubMed [26] adding thousands of scientific papers daily. The digitization of patient records has created vast stores of clinical text, with millions of new cancer patients generating extensive clinical notes annually. This accumulation of data presents opportunities to accelerate clinical research and enhance patient care. However, manual curation of this information is not scalable due to the time-consuming nature of annotation and the domain-specific expertise required.

Natural language processing (NLP) has emerged as a promising solution to automate the extraction of key findings from biomedical texts, facilitating human validation. Nevertheless, the effectiveness of standard supervised learning in this domain often hinges on the availability of large training datasets, which are scarce in biomedical contexts. Task-agnostic self-supervised learning is gaining traction, allowing large language models to facilitate transfer learning and achieve remarkable success across a range of NLP applications.

To bridge existing gaps in the literature, this study conducts an extensive analysis of LoRA-based fine-tuning across multiple models and tasks, utilizing standardized prompting techniques to enable direct comparisons. By exploring the viability of serving multiple LoRA models in real-world applications, we leverage the LoRAX framework, which allows for the efficient deployment of fine-tuned LLMs while measuring latency and concurrency metrics.

This research aims to improve AI-driven therapy by enhancing the ability to offer personalized support, thus addressing the nuanced needs of individuals facing mental health challenges. The findings have the potential to drive the development of more impactful AI-based counseling systems that not only establish authentic emotional engagement with users but also optimize cost-effectiveness in real-world implementation.

2 Related Work

Recent research has focused on using conversational agents, or chatbots, to provide mental health support. These chatbots offer a promising way to deliver accessible and affordable help to those in need. According to Martinengo [16] found that chatbots can provide anonymous, empathetic, and non-judgmental interactions, similar to traditional therapy. Utilizing Natural Language Processing (NLP) techniques, these chatbots can engage users in therapeutic conversations and offer tailored support. As shown in prior studies [7], chatbots have demonstrated effectiveness in delivering mental health interventions. Pre-trained language models (PLMs), especially GPT-3, have become significant in mental health counseling. However, there are concerns about controlling outputs and adhering to ethical standards.

2.1 NLP Techniques and Mental Health Analysis

Research on using NLP to analyse mental health-related text is growing. For example, Gonzalez-Hernandez [8], Sadler [21] and Ji [12] used machine learning to identify mental health conditions, predict suicidal thoughts, and find linguistic markers related to mental well-being. As per De Choudhury [6], social media data can be effectively analyzed to predict depression, showing potential for early detection of individuals at risk. Other studies have integrated modern technologies into mental health interventions, such as Lui [15], who looked into mobile apps for psychotherapy. Further advancements include Shaikh [23], who developed a friendly AI-based chatbot to help individuals with insomnia by addressing negative feelings and encouraging interaction during sad or anxious times. Chatbots have also been widely used in customer service, providing quick responses that improve customer satisfaction [24]. The development of language models like BERT and GPT has greatly improved chatbot accuracy and efficiency [14]. However, challenges still exist in generating coherent and contextually relevant responses.

2.2 Therapeutic Approaches and Counseling Techniques

Psychological counseling methods, like Cognitive Behavioral Therapy (CBT) [10], have proven effective in improving mental health and quality of life by helping individuals recognize, resolve, and manage psychological challenges [17]. Integrating LLMs into mental health counseling has gained popularity due to their ability to produce human-like text and respond to user input. Models like GPT-3.5 and GPT-4 have shown potential in providing real-time emotional support for individuals facing challenges like anxiety and depression. Nevertheless, these models struggle with demonstrating genuine empathy and personalization. For instance, while GPT-3.5 has been evaluated for various mental health tasks, its capacity to provide emotionally resonant responses is limited, often lacking the depth needed for effective therapeutic conversations. To overcome these limitations, domain-specific models like Mental-LLaMA have been developed, utilizing datasets tailored to mental health dialogues. This specialization allows the models to better understand and respond to the specific needs of users. Although models like Mental-LLaMA show improvements over general-purpose LLMs, they still face challenges in fostering meaningful emotional connections and delivering responses that reflect true empathy.

2.3 Challenges in Empathy and Personalization

A major challenge in using large language models (LLMs) for mental health counseling lies in fostering authentic empathy. Empathy requires not only understanding surface-level text but also decoding nuanced emotional cues such as sarcasm, hesitation, or distress. General-purpose LLMs, including GPT-3.5, often generate generic or emotionally flat responses that may feel impersonal. For example, if a user says, "I feel like I'm disappearing," a general model might reply with a factual statement like, "You are not invisible," rather than validating the emotional pains [13]. This is especially concerning in therapeutic contexts, where misinterpreting emotions can have harmful consequences for users. Another ongoing issue is the generation of hallucinations responses that sound plausible but are factually incorrect [2]. These inaccuracies can erode the trust and safety essential in mental health counseling, where users seek reliable support during vulnerable times. Consequently, enhancing the personalization and accuracy of LLM responses remains a crucial focus of research.

2.4 LoRA, Unsloth, and Their Role in Mental Health AI

Low-Rank Adaptation (LoRA) is a fine-tuning technique that modifies only a small subset of a model's parameters using low-rank matrices [11]. This allows developers to adapt large models like Phi-3.5 for specific tasks—such as empathetic mental health counseling—without retraining the full model. LoRA focuses on select neural network layers (e.g., query, key, value projections) and introduces trainable parameters that refine how attention is computed in the transformer. The result is a more task-aware, memory-efficient model that retains the benefits of pretraining while incorporating emotional intelligence.

The Unsloth framework further complements LoRA by enhancing training speed and memory optimization [5]. It applies techniques like gradient checkpointing and 4-bit floating point loading to drastically reduce memory usage. When used together, LoRA and Unsloth enable fine-tuning large models on modest hardware (e.g., a single GPU), reducing training time by 50% and memory usage by 70%.
In the context of mental health, this enables the deployment of empathetic AI models on mobile devices or in low-resource settings. Together, these frameworks offer a scalable path for building intelligent, responsive AI therapists that adapt to individual needs.

3 Methodology

This section outlines the methodology used to adapt the Phi-3.5-mini-instruct language model for empathetic mental health counseling. The goal was to produce a lightweight, emotionally intelligent model deployable even in resource-constrained settings. We achieved this by integrating Low-Rank Adaptation (LoRA) for efficient fine-tuning and the Unsloth framework to optimize training time and memory usage.

3.1 Model Architecture

We based our work on the Phi-3.5-mini-instruct model—a compact variant of GPT-3.5 optimized for fast, cost-effective deployment. Built on a decoder-only transformer architecture, it is capable of handling long-range context and generating coherent, context-aware responses. Its structure mirrors that of LLaMA-2, enabling compatibility with modern fine-tuning tools. To adapt the model for mental health tasks, we modified only select attention layers using LoRA modules, which insert small trainable matrices into specific layers like q_proj (query), k_proj (key), v_proj (value), and o_proj (output).

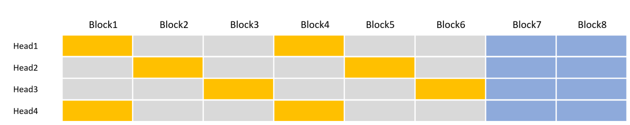


Figure 1: This example demonstrates block-sparse attention in the phi-3-small model, using two local blocks with a vertical stride of 3. The table shows keys and values attended to by a query token in block 8, with blue for local blocks, orange for remote blocks, and gray for skipped blocks.

3.2 Technical Specifications of Phi-3.5-mini

Table 1: Model Specification and Configuration

Feature	Description
Model Type	Transformer Decoder
Base Architecture	Similar to LLaMA-2
Tokenizer	Shared with LLaMA-2 family
Vocabulary Size	320,641 tokens
Context Length	Default: 4,000 tokens; Extended: 128,000 (via LongRope)
Hidden Dimensions	3,072
Number of Heads	32
Number of Layers	32
Training Precision	bfloat16
Total Pretraining Data	3.3 trillion tokens

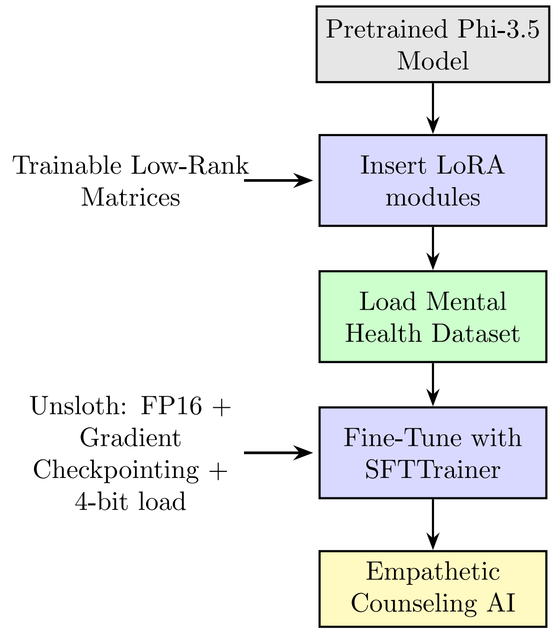
- **Chat Template:**
```
<|user|>
   Question
<|end|>
<|assistant|>
```

3.3 Fine-Tuning Process with LoRA

Low-Rank Adaptation (LoRA) allows us to fine-tune a small portion of the model's weights by adding trainable low-rank matrices. This drastically reduces memory requirements and improves task specificity without sacrificing general model performance.

Steps in Fine-Tuning:

- **Initialization**: Loaded pre-trained Phi-3.5-mini weights.
- **LoRA Injection**: Added trainable matrices into transformer attention layers.
- **Dataset Integration**: Merged diverse, real-world counseling dialogues (see section 4.1).
- **Training Configuration:**
  - Batch size: 16
  - Learning rate: 3e-5
  - Epochs: 35
  - Floating point precision: load_in_4bit for efficient memory usage



3.4 Training Optimization with Unsloth

The Unsloth framework was implemented to speed up training and reduce GPU memory use. It employs the following techniques:

- **Gradient Checkpointing:** Reduces memory load by recomputing activations during backpropagation instead of storing them.
- **Mixed Precision Training (FP16):** Most computations are done in 16-bit precision, reducing memory and speeding up processing.
- **4-bit Precision Loading (load_in_4bit):** Minimizes model weight size, allowing for training on mid-range GPUs (e.g., Tesla T4).

Outcomes:

- Memory usage reduced by 70%.
- Training time cut by 50%.
- Enabled training on a single 16 GB GPU (Google Colab Tesla T4)

3.5 Evaluation Framework

The performance of the fine-tuned model was rigorously assessed using the following metrics:

- **Quantitative Evaluation (BERTScore):** To assess the semantic quality of AI-generated responses, we used BERTScore, a contextual similarity metric based on embeddings from pre-trained BERT models. Unlike traditional lexical metrics, BERTScore captures deeper semantic alignment between the candidate and reference texts, making it particularly suitable for evaluating context retention and emotional nuance—critical aspects in mental health and counseling dialogue systems.

We computed Precision, Recall, and F1 Score using BERTScore to capture different facets of performance. Precision measured the semantic relevance of the generated response, Recall assessed content coverage, and the F1 Score provided a balanced view. This combination allowed for a robust evaluation of both factual and emotional fidelity in comparison to human-written responses.

- **Human-Informed Analysis of Emotional and Therapeutic Appropriateness:** Given the limitations of automated metrics in capturing emotional nuance, we incorporated a human-centered evaluation to complement quantitative results. A representative sample of AI-generated responses was manually reviewed by experts in psychology and counseling, using the following criteria:
  - **Empathy:** Ability to acknowledge and validate emotional expressions.
  - **Tone Appropriateness:** Appropriateness of the language used (supportive, non-judgmental).
  - **Therapeutic Relevance:** Relevance to principles of supportive counseling.

  This qualitative assessment provided critical insight into the model's ability to engage in emotionally attuned and ethically sound interactions, essential for mental health applications.

4 Experimental Details

This section delineates the experimental framework, encompassing the datasets, training parameters, hardware specifications, and evaluation methodology utilized for fine-tuning the Phi-3.5-mini-instruct model for mental health counseling.

4.1 Datasets

To enhance the model's capabilities, we leveraged three publicly available datasets focusing on mental health counseling conversations:

- **Amod Mental Health Counseling Conversations [3]:** This dataset contains real, anonymized mental health counseling dialogues, reflecting a range of emotional contexts including anxiety, grief, and trauma.

Table 2: Example Dialogues Used for Fine-Tuning

User Input	**AI Response**
"I feel like a burden to everyone."	"I'm really sorry you're feeling this way. You're not alone, and your feelings matter. Can you tell me more about what's been going on?"
"I can't sleep anymore. Everything feels pointless."	"That sounds incredibly hard. Sometimes our thoughts can get overwhelming. Would you like to explore ways to ease some of that stress?"
"My parents don't believe in mental health."	"That must feel isolating. It's important to be heard, and I'm here to support you in this space."

- **Mayuresh Anand Pingale mental-health-chat-dataset [18]:** This resource includes anonymized community mental health chat transcripts, covering concerns like anxiety, depression, stress, and emotional states such as panic and loneliness.
- **Arun Brahma (heliosbrahma) mental_health_chatbot_dataset [4]:** A collection of simulated therapeutic conversations designed to train AI models in providing structured, empathetic, and personalized mental health support.

These datasets were integrated to construct a rich and diverse training corpus, ensuring the model could generalize across various mental health interactions effectively.

4.2 Hardware and Compute Resources

The fine-tuning experiments were conducted using a Colab Tesla T4 GPU-based compute engine, which provided robust computational power. This setup facilitated optimized memory utilization through the integration of the LoRA and Unsloth frameworks, significantly reducing memory consumption and enabling efficient resource management throughout the training process.

4.3 Training Setup

The model was fine-tuned utilizing the following hyperparameters and configuration settings:

- **Model Name:** `unsloth/Phi-3.5-mini-instruct`
- **Maximum Sequence Length:** `max_seq_length`
- **LoRA Parameters:**
  - **Rank (r):** 16
  - **Target Modules:** ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]
  - **LoRA Alpha:** 16
  - **LoRA Dropout:** 0 (optimized for zero dropout)
  - **Bias:** "none" (configured for no bias)
  - **Gradient Checkpointing:** `"unsloth"` (employed for managing very long context sequences, resulting in a 30% reduction in VRAM usage)
  - **Random State:** 3407
  - **Use RSLora:** False (rank-stabilized LoRA not applied)
  - **LoftQ Configuration:** None
  - **Unsloth Version:** 2024.9.post1, patched for 32 layers, including 32 QKV layers, 32 output layers, and 32 MLP layers.
- **Projection Modules:**
  - **q_proj (Query Projection):** Transforms input data into query vectors, which are pivotal in determining the focus on different parts of the input sequence during attention calculation.
  - **k_proj (Key Projection):** Converts input data into key vectors that are compared against query vectors to derive attention scores, influencing which information is prioritized.
  - **v_proj (Value Projection):** Generates value vectors representing the actual information output based on attention scores derived from the queries and keys.
  - **o_proj (Output Projection):** Processes the output from the attention mechanism, typically projecting it back to the original input dimension for further processing.
  - **gate_proj (Gate Projection):** Works with gating mechanisms to manage the flow of information, aiding the model in deciding which elements to retain and which to disregard.
  - **up_proj (Upward Projection):** Increases the dimensionality of the data, facilitating more complex representations when necessary.
  - **down_proj (Downward Projection):** Decreases the dimensionality of the data, often employed to compress information and lessen computational load.

Training was conducted using the `SFTTrainer` framework with the following parameters:

- **Training Batch Size:** 2 per device, with gradient accumulation steps set to 4
- **Learning Rate:** 2e-4
- **Warmup Steps:** 5
- **Maximum Steps:** 60
- **Optimizer:** `adamw_8bit` (optimized for 8-bit precision)
- **Weight Decay:** 0.01
- **Scheduler:** Linear learning rate scheduler
- **Mixed Precision:** FP16 where applicable, with BF16 fallback for environments supporting BF16
- **Random Seed:** 3407
- **Output Directory:** "outputs"

To further optimize memory utilization, the model was trained with floating point `load_in_4bit`, significantly minimizing the memory footprint while preserving accuracy.

4.4 Evaluation

For evaluation, the **BERTScore** was computed using the model `allenai/longformer-base-4096`. BERTScore measures the semantic similarity between the model-generated responses and human-authored reference responses, making it an apt metric for assessing the quality and empathetic nature of AI-generated dialogues in mental health counseling.

The model's performance was evaluated on its capacity to generate empathetic and relevant responses based on inputs from the mental health counseling datasets. BERTScore precision, recall, and F1-score metrics were calculated, offering valuable insights into the model's ability to emulate expert-level, contextually appropriate counseling responses.

5 Results

5.1 Training Loss

The training performance of the model was assessed over 35 epochs, with training loss recorded at each step. The overall trend indicated a progressive reduction in loss, despite intermittent fluctuations.

The initial training loss was 2.0646 in the first epoch, which generally decreased over time, reaching its lowest value of 0.8933 at epoch 33. However, certain epochs exhibited deviations from this downward trend, such as epoch 3 (1.6567), epoch 8 (1.7425), and epoch 14 (1.7073), potentially indicating learning challenges or early signs of overfitting. Despite these fluctuations, the overall trajectory suggests effective model convergence. By the later epochs, training loss stabilized within the range of 1.3–1.5, signifying that the model successfully optimized its parameters throughout the training process.

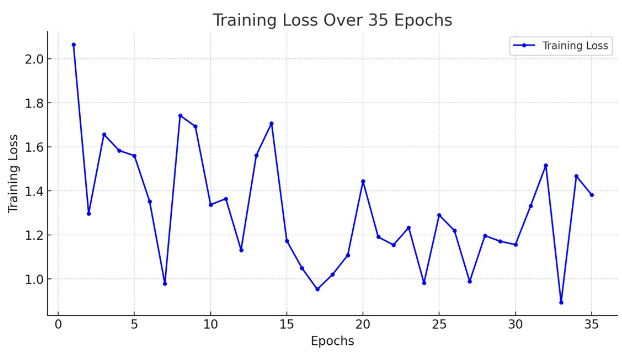


Figure 2: Training Loss: A visual representation illustrating the progressive diminution of error over epochs, capturing the model's gradual refinement and enhanced predictive accuracy as it undergoes iterative learning.

Summary of Key Observations:

- Initial training loss: 2.8832 at epoch 1.
- Lowest training loss: 1.9589 at epoch 39.
- Notable fluctuations in training loss at epochs 4, 12, and 17, where the loss briefly increased before stabilizing.

This decline in loss illustrates that the model was able to gradually optimize its parameters through the training process.

5.2 Memory Consumption

Memory usage during the training process remained well within the available computational limits of the Tesla T4 GPU. Peak reserved memory during training was 3.338 GB, which represented 22.634% of the total memory capacity (14.748 GB). The memory specifically reserved for training peaked at 1.053 GB, amounting to 7.14% of the maximum available memory.

These metrics indicate that memory utilization was highly efficient, largely due to the optimization techniques implemented through the LoRA and Unsloth frameworks. This allowed the training process to be conducted smoothly without exceeding memory limitations, ensuring stable performance throughout the training period.

5.3 Model Performance: Precision, Recall, and F1 Score

The model's performance was evaluated using precision, recall, and F1 score metrics, both on the training set and the test set. The scores are as follows:

Dataset	Precision	Recall	F1 Score
Training Set	0.8663	0.8213	0.8429
Test Set	0.8655	0.8139	0.8385

Table 3: Model Performance Metrics

The training results reflect strong precision and recall, leading to an overall F1 score of 0.8429, demonstrating the model's ability to accurately predict and generalize unseen data. The slight reduction in F1 score on the test set (0.8385) suggests some degree of overfitting, but overall, the model maintained robust performance across both training and test datasets.

6 Conclusion

This study presents a fine-tuned version of the Phi-3.5-mini-instruct model, designed for delivering empathetic and personalized mental health counseling. By using Low-Rank Adaptation (LoRA) and the Unsloth framework, we achieved a significant reduction in memory usage and training time, making the model lightweight and efficient for real-world applications.

The model was able to produce responses that closely match expert-written counseling replies, as shown by strong BERTScore evaluations. These responses reflected both relevance and empathy-two key qualities in effective mental health communication. The performance demonstrates that with proper fine-tuning, even smaller language models can deliver meaningful and supportive interactions.

Our approach also enables deployment in resource-limited environments, such as mobile devices or regions with limited access to computational infrastructure. This makes the solution not only technically robust but also practically scalable.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Saleh Afroogh, Yasser Poreesmaiel, and Junfeng Jiao. Hallucinations vs. predictions: Reframing uncertainty in llm-generated medical responses. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

[3] Antonio Agliata, Antonio Pilato, Sorrentino Mariacarmen, Salvatore Bottiglieri, Emanuel Di Nardo, and Angelo Ciaramella. Generative ai and emotional health: Innovations with haystack. In *2024 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–4. IEEE, 2024.

[4] Arun Brahma. Mental health chatbot dataset, 2023.

[5] Michael Han Daniel Han and Unsloth team. Unsloth, 2023.

[6] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137, 2013.

[7] Kerstin Denecke, Alaa Abd-Alrazaq, and Mowafa Househ. Artificial intelligence for chatbots in mental health: opportunities and challenges. *Multiple perspectives on artificial intelligence in healthcare: Opportunities and challenges*, pages 115–128, 2021.

[8] Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O'Connor, and Guergana Savova. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227, 2017.

[9] Jagoda Grzejszczak, Dominik Strzelecki, Agata Gabryelska, and Magdalena Kotlicka-Antczak. Evaluation of covid-19 effect on mental health, self-harm, and suicidal behaviors in children and adolescents population. *Journal of clinical medicine*, 13(3):744, 2024.

[10] Steven D Hollon and Aaron T Beck. Cognitive therapy of depression. *Cognitive-behavioral interventions: Theory, research, and procedures*, pages 153–203, 1979.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[12] Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226, 2020.

[13] Sijie Ji, Xinzhe Zheng, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. Mindguard: Towards accessible and sitgma-free mental health first aid via edge llm. *arXiv preprint arXiv:2409.10064*, 2024.

[14] Nikita Kanodia, Khandakar Ahmed, and Yuan Miao. Question answering model based conversational chatbot using bert model and google dialogflow. In *2021 31st International Telecommunication Networks and Applications Conference (ITNAC)*, pages 19–22. IEEE, 2021.

[15] Joyce HL Lui, David K Marcus, and Christopher T Barry. Evidence-based apps? a review of mental health mobile applications in a psychotherapy context. *Professional Psychology: Research and Practice*, 48(3):199, 2017.

[16] Laura Martinengo, Elaine Lum, and Josip Car. Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis. *Journal of affective disorders*, 319:598–607, 2022.

[17] Augustine Meier, Micheline Boivin, and Molisa Meier. Working through the transference of an unresolved separation/individuation pattern: a case study using theme-analysis. *The helping relationship: Healing and change in community context*, pages 101–129, 2010.

[18] Mpingale. Mental health chat dataset, Year.

[19] World Health Organization. The world health report 2001: Mental health: new understanding, new hope. 2001.

[20] Aravind Pillai. *Screening for common mental disorders in primary care in low and middle income countries: A rational approach to address the mental health treatment gap?* PhD thesis, Columbia University, 2020.

[21] Eugene Sadler-Smith and Vita Akstinaite. Human hubris, anthropogenic climate change, and an environmental ethic of humility. *Organization & Environment*, 35(3):446–467, 2022.

[22] Vasu Saini, Luis Guada, and Dileep R Yavagal. Global epidemiology of stroke and access to acute ischemic stroke interventions. *Neurology*, 97(20_Supplement_2):S6–S16, 2021.

[23] Talha Abdul Hakeem Shaikh and Manisha Mhetre. Autonomous ai chat bot therapy for patient with insomnia. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–5. IEEE, 2022.

[24] Manoj Kumar Tamrakar and Abhishek Badholia. Scientific study of technological chatbot adoption in customer service. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1117–1123. IEEE, 2022.

[25] Adam Timmis, Nick Townsend, Chris P Gale, Aleksandra Torbica, Maddalena Lettino, Steffen E Petersen, Elias A Mossialos, Aldo P Maggioni, Dzianis Kazakiewicz, Heidi T May, et al. European society of cardiology: cardiovascular disease statistics 2019. *European heart journal*, 41(1):12–85, 2020.

[26] Jacob White. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387, 2020.