FedMentalCare: Towards Privacy-Preserving Fine-Tuned LLMs to Analyze Mental Health Status Using Federated Learning Framework

S M Sarwar
University of Maryland Baltimore County
smsarwar96@gmail.com

Abstract

With the increasing prevalence of mental health conditions worldwide, AI-powered chatbots and conversational agents have emerged as accessible tools to support mental health. However, deploying Large Language Models (LLMs) in mental healthcare applications raises significant privacy concerns, especially regarding regulations like HIPAA and GDPR. In this work, we propose FedMentalCare, a privacy-preserving framework that leverages Federated Learning (FL) combined with Low-Rank Adaptation (LoRA) to fine-tune LLMs for mental health analysis. We investigate the performance impact of varying client data volumes and model architectures (e.g., MobileBERT and MiniLM) in FL environments. Our framework demonstrates a scalable, privacy-aware approach for deploying LLMs in real-world mental healthcare scenarios, addressing data security and computational efficiency challenges.

1 Introduction

Mental health encompasses cognitive processing, emotional regulation, behavioral actions, and mood stability. Achieving mental well-being allows individuals to handle daily stress, work effectively, and contribute to their communities. In 2019, around 970 million people globally experienced mental disorders, affecting one in eight individuals (of Health Metrics and Evaluation, 2023). To address these challenges, AI-powered chatbots and conversational agents (CAs) are being developed to provide fast, accessible, and confidential mental health support (Healthcare, 2023). The global market for mental health chatbots was valued at $0.99 billion in 2022 and is projected to reach $6.51 billion by 2032, driven by a 21.3% annual growth rate (Healthcare, 2023).

Recent developments in Large Language Models (LLMs) have significantly improved the effectiveness of chatbots in delivering psychological support. However, deploying LLMs in healthcare introduces significant data privacy and security concerns (May and Denecke, 2022; Nicholas et al., 2020). These concerns are compounded by stringent regulations like the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the EU (104th United States Congress, 1996; Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016). This raises an essential research question **(RQ1)**: How can LLMs be fine-tuned in a Federated Learning setting to ensure data privacy while complying with HIPAA and GDPR?

Another limitation of LLMs is their vulnerability to producing false or inaccurate information, which can negatively impact mental health (Hua et al., 2024; Guo et al., 2024; Marrapese et al., 2024). Collecting large-scale conversational data is essential for improving LLM accuracy, but privacy concerns often deter users from sharing data. This leads to incomplete datasets, hindering robust model development for mental health analysis (Yu et al., 2023; Han et al., 2023). Consequently, we explore the research question **(RQ2)**: How does the volume of client data affect the performance of fine-tuned LLMs in generating accurate mental health insights?

The key contributions of this paper are:

- We propose **FedMentalCare**, a Federated Learning framework, for fine-tuning LLMs to preserve data privacy and comply with HIPAA and GDPR.
- We investigate the impact of client data volume on LLM performance in sensitive mental health applications.
- We analyze the impact of small language models like MobileBERT and MiniLM on performance in Federated Learning environments.

2 Backgrounds

2.1 Large Language Models

Large Language Models (LLMs) revolutionized the field of NLP with the introduction of the transformer architecture by Google Brain in 2017 (Vaswani et al., 2017). This innovation led to the development of Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2018) for contextual understanding and GPT (Radford et al., 2018) for coherent text generation. LLMs vary in size, from smaller models like TinyBERT (14.5M parameters) (Jiao et al., 2019) to models exceeding a trillion parameters like GPT-4 (1.7T) (OpenAI, 2023). Depending on architecture, LLMs can be encoder-based (BERT (Devlin et al., 2018)), decoder-based (LLaMa (Touvron et al., 2023)), or encoder-decoder (T5 (Raffel et al., 2020)). In our work, we utilize BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), MobileBERT (Sun et al., 2020) and MiniLM (Wang et al., 2020) for their effectiveness in text generation tasks.

Recent advancements, including FLAN-T5 (Chung et al., 2022), PaLM (Chowdhery et al., 2023), and Gemini (Team et al., 2023), have significantly enhanced NLP capabilities and shown promise in healthcare applications (Nori et al., 2023; Liu et al., 2023b). However, applying LLMs for mental health status analysis within a Federated Learning (FL) framework presents challenges (McMahan et al., 2017). Key drawbacks include the risk of data leakage during fine-tuning, high computational costs, and the difficulty of preserving model performance while ensuring privacy through techniques like Differential Privacy (DP) (Dwork, 2006). Addressing these issues is essential to developing privacy-preserving, accurate LLMs for mental health analysis using fine-tuned or zero-shot methods (Bonawitz et al., 2017).

2.2 Federated Learning for Mental Health

Federated Learning (FL) provides a collaborative machine learning framework that enables model training without exposing private user data (McMahan et al., 2017; Kairouz et al., 2021). Unlike traditional centralized ML, where data is aggregated on a central server, FL allows clients to contribute model updates (weights and gradients) to a central server, minimizing data leakage and enhancing privacy (Bonawitz et al., 2017). Foundational algorithms like FedAvg (McMahan et al., 2017) aggregate client updates by averaging, while methods like FedProx (Li et al., 2020) address data heterogeneity. Techniques such as FedSVRG (Konečnỳ et al., 2016) improve training efficiency, and Secure Aggregation (SecAgg) (Bonawitz et al., 2017) ensures privacy-preserving update aggregation.

In mental health, FL has been leveraged for privacy-preserving emotion and depression analysis. CAFed (Li et al., 2021) proposes an asynchronous FL model for depression analysis that improves communication efficiency and convergence over FedAvg. Similarly, Cui et al. (Cui et al., 2022) propose a speech-based FL model incorporating norm bounding, differential privacy, and secure aggregation to protect user data. By leveraging its decentralized structure, FL enables secure and privacy-aware collaboration among healthcare stakeholders, including hospitals, institutions, and patients (Rieke et al., 2020; Antunes et al., 2022). Recent works have explored FL for developing LLMs to address privacy and data security concerns in distributed settings (Fan et al., 2023; Ye et al., 2024; Vu et al., 2024). These efforts point to FL as a transformative tool for privacy-centric, efficient, and personalized healthcare solutions (Rauniyar et al., 2023; Peng et al., 2023).

3 LLMs in Healthcare and Mental Healthcare Solutions

Integrating LLMs in healthcare, particularly mental health support, is an area of increasing interest and rapid development. The research in this field highlights the transformative potential of LLMs, from detecting early signs of depression and anxiety through textual analysis to providing therapeutic conversation agents for mental health support. The potential applications of LLMs in healthcare are diverse, including medical information extraction and analysis (Singhal et al., 2022), drug discovery-related research (Liang et al., 2023), personalized medicine development (Yang et al., 2023a), etc. The possibilities of LLMs-powered healthcare chatbots (Bernstein et al., 2023), virtual health assistants (Zhang et al., 2023; Li et al., 2023b; Gao et al., 2023), and clinical decision-making tools (Li et al., 2023a) are also promising.

The application of LLMs in mental health support is a rapidly growing area. Some mental health solutions based on LLMs have already been introduced (Yang et al., 2023c; Xu et al., 2023; Chen et al., 2023; Liu et al., 2023a; Yang et al., 2023b).

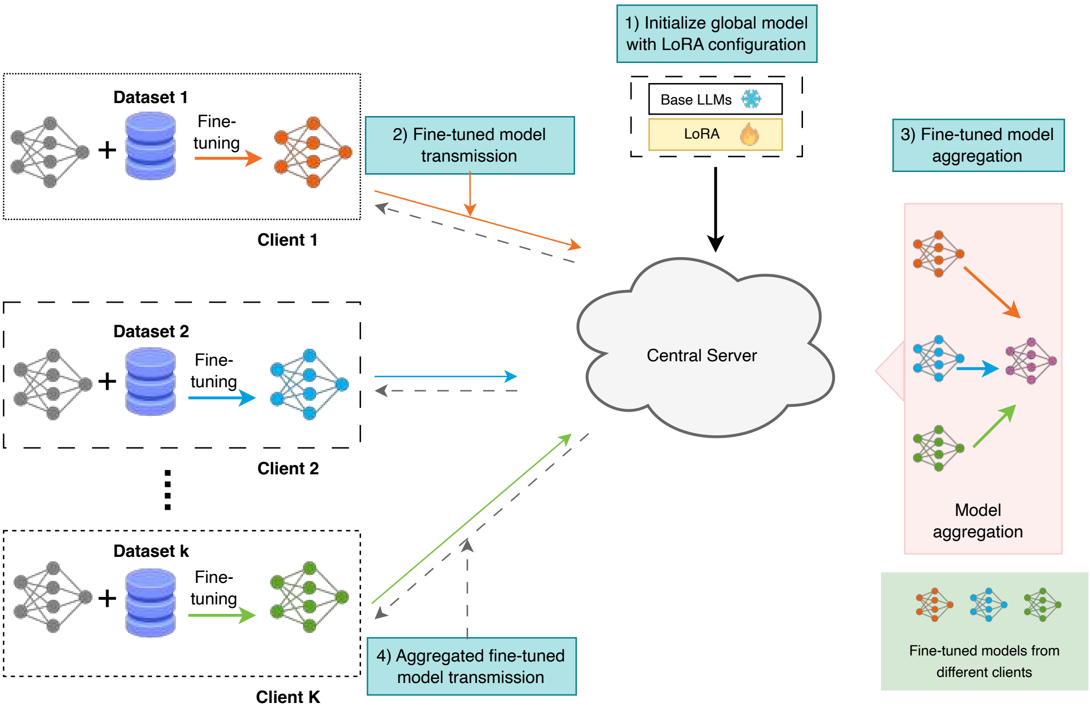
1) Initialize global model with LoRA configuration
Base LLMs
LoRA
Dataset 1
Fine-tuning
Client 1
2) Fine-tuned model transmission
3) Fine-tuned model aggregation
Dataset 2
Fine-tuning
Client 2
Central Server
Dataset k
Fine-tuning
Client K
4) Aggregated fine-tuned model transmission
Model aggregation
Fine-tuned models from different clients

Figure 1: Proposed **FedMentalCare** Framework: The architecture integrates **Federated Learning (FL)** with **Low-Rank Adaptation (LoRA)** to efficiently fine-tune **Large Language Models (LLMs)** on client devices for mental healthcare applications. The server initializes a global model with LoRA configuration, clients fine-tune the global model locally, and the server aggregates client updates via **Federated Averaging (FedAvg)**, ensuring privacy and optimizing computational efficiency.

These models have the capability to engage in therapeutic interactions and evaluate patient communication and actions, which helps in the early detection of mental health conditions. These LLMs enhance the support provided in mental healthcare. However, the effectiveness of state-of-the-art LLMs in FL settings remains an open question. This gap motivates our research to explore FL-enabled LLMs for detecting mental health indicators from social media text, offering new insights into their downstream applications.

4 Method

4.1 Problem Statement

We aim to fine-tune LLMs for mental healthcare applications in a FL setting, addressing both computational efficiency and privacy constraints. Given a set of $\mathcal{K}$ clients, each with a local dataset D_k for client k, the goal is to collaboratively train a global model $\mathcal{M}_{\text{global}}$ without sharing raw data (Equations 3).

Each client k locally trains a copy of the global model $\mathcal{M}_k$ for E epochs using gradient descent. After local training, clients send their model parameters θ_k to the server. The server aggregates these updates using Federated Averaging (FedAvg) (McMahan et al., 2017) and updates the global model $\mathcal{M}_{\text{global}}$ accordingly (Equations 4).

To address the computational constraints of FL, we incorporate Low-Rank Adaptation (LoRA) (Hu et al., 2021), which reduces the number of trainable parameters by decomposing the weight update ΔW into low-rank matrices A and B (Equations 1 and 2). This approach minimizes communication costs and enables efficient model adaptation on resource-constrained devices like mobile phones, ensuring privacy preservation and scalability in real-world mental healthcare applications (Pfeiffer et al., 2024; Li et al., 2020; Kairouz et al., 2021).

4.2 Parameter-Efficient Fine-Tuning

We adopt Low-Rank Adaptation (LoRA) (Hu et al., 2021) as a Parameter-Efficient Fine-Tuning (PEFT) approach (Houlsby et al., 2019; Dettmers et al., 2024; Liu et al., 2024) to efficiently fine-tune LLMs in a FL setting for mental healthcare applications. LoRA reduces the number of trainable parameters by decomposing the weight update matrix ΔW into two low-rank matrices A and B, such that:

$$\Delta W = BA \tag{1}$$

where $A \in \mathbf{R}^{r \times k}$ and $B \in \mathbf{R}^{d \times r}$, with $r \ll d$. During training, the pre-trained weights W_0 remain frozen, and the modified weight becomes:

The updated model weights are expressed as:

$$W = W_0 + \Delta W = W_0 + BA, \tag{2}$$

where $W_0 \in \mathbf{R}^{d \times k}$ represents the frozen pre-trained weights. This decomposition reduces the number of trainable parameters from dk to $r(d + k)$, significantly lowering computational overhead while maintaining the expressiveness of the model.

By integrating LoRA into the FL framework, we enable efficient adaptation of transformer models such as MobileBERT (Sun et al., 2020) and MiniLM (Wang et al., 2020). This approach preserves privacy by keeping the data on-device and reduces communication costs by limiting the size of model updates (Pfeiffer et al., 2024).

Algorithm 1 Server Aggregation

Inputs: $\mathcal{K}$ (number of clients), E (number of local training epochs), $\mathcal{R}$ (number of global aggregation rounds), η (learning rate), $\mathcal{M}$ (pre-trained transformer models such as BERT, RoBERTa)

Output:

$\mathcal{M}_{\text{global}}$ (Updated global model)

Server executes:

```
Initialize global model $\mathcal{M}_{\text{global}}$ with LoRA configuration.
Load tokenizer $\mathcal{T}$ corresponding to the transformer models.
for each round $r = 1, 2, \ldots, \mathcal{R}$ do
    for each client $k = 1, 2, \ldots, \mathcal{K}$ in parallel do
        $\mathcal{M}_k \leftarrow$ Copy of $\mathcal{M}_{\text{global}}$
        $\theta_k \leftarrow$ ClientUpdate($\mathcal{M}_k, D_k, \mathcal{T}, E$)
    end for
    Aggregate client models:
        $\theta_{\text{global}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$
    Update $\mathcal{M}_{\text{global}}$ with $\theta_{\text{global}}$
end for
return $\mathcal{M}_{\text{global}}$
```

4.3 FedMentalCare Framework

FedMentalCare Framework integrates Federated Learning (FL) with LoRA (Hu et al., 2021) to efficiently fine-tune LLMs for mental healthcare applications. This approach ensures data privacy by keeping user data on-device and reduces computational overhead, making it suitable for resource-constrained devices.

The training process consists of two primary steps: Server Aggregation (Algorithm 1) and Client Training (Algorithm 2). Figure 1 provides a detailed illustration of the interactions between the server and clients during these steps.

Server Aggregation: The server initializes the global model $\mathcal{M}_{\text{global}}$ with a LoRA configuration and distributes it to $\mathcal{K}$ clients. After each round r, clients return their locally trained model parameters θ_k. The server aggregates these updates using Federated Averaging (FedAvg):

$$\theta_{\text{global}} = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \theta_k, \tag{3}$$

and updates the global model $\mathcal{M}_{\text{global}}$ accordingly.

Algorithm 2 Client Training

Inputs: D_k (local dataset for each client k), E (number of local training epochs), η (learning rate), $\mathcal{T}$ (tokenizer corresponding to the transformer models)

Output:

θ_k (Updated model parameters for client k)

ClientUpdate($\mathcal{M}_k, D_k, \mathcal{T}, E$):

```
Split $D_k$ into training and evaluation sets.
Tokenize data using tokenizer $\mathcal{T}$.
for each local epoch $e = 1, 2, \ldots, E$ do
    for each batch $b$ in $D_k$ do
        Compute gradients $g \leftarrow \nabla \ell(\mathcal{M}_k; b)$
        Update model $\mathcal{M}_k \leftarrow \mathcal{M}_k - \eta g$
    end for
end for
return model parameters $\theta_k$ to the server.
```

Client Training: Each client k receives the global model $\mathcal{M}_k$ and trains it locally for E epochs on their dataset D_k. The client performs gradient descent to update the model weights:

$$\mathcal{M}_k \leftarrow \mathcal{M}_k - \eta \nabla \ell(\mathcal{M}_k; b), \tag{4}$$

where η is the learning rate and ℓ is the loss function computed over a batch b of D_k.

By leveraging the LoRA decomposition described earlier (Equations 1 and 2), the FedMentalCare framework optimizes both computation and communication. This enables efficient model adaptation while preserving privacy, making it ideal for deploying transformer models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and MiniLM (Wang et al., 2020) in real-world mental healthcare applications.

5 Experiment

5.1 Datasets

In this experiment, we utilize the Dreaddit dataset (Turcan and McKeown, 2019), a large-scale collection of 190k Reddit posts across five distinct domains: interpersonal conflict, mental illness, financial need, PTSD, and Social. The dataset includes 3,553 annotated segments (∼100 tokens each) labeled for binary stress classification (stressful vs. non-stressful). The detailed and context-rich posts, combined with diverse expressions of stress, make it a valuable resource for benchmarking stress detection models and uncovering implicit stress indicators. This dataset supports the development of robust models with potential applications in mental health monitoring, social behavior analysis, and stress assessment in online communities. Figure 2 presents a word cloud visualization highlighting key terms prevalent in the dataset, while Table 1 provides a concise breakdown of the dataset structure.

Table 1: Dreaddit Dataset Overview

Domain	Posts	Labeled Segments
Interpersonal Conflict	2,901	703
Mental Illness	59,208	728
Financial Need	12,517	717
PTSD	4,910	711
Social	107,908	694
Total	**187,444**	**3,553**

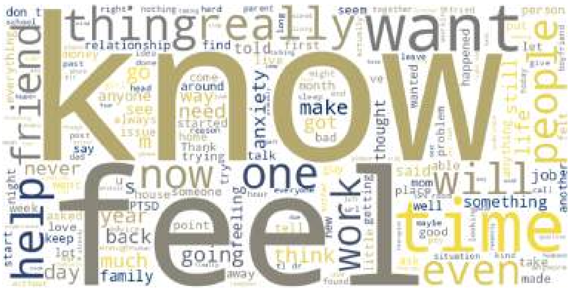

Figure 2: Word Cloud for Dreaddit Dataset.

5.2 Implementation Details

The experiments are executed on Google Colab utilizing an NVIDIA T4 GPU, which provides 16 GB of GDDR6 memory and 320 Tensor Cores. The T4 GPU supports mixed-precision computation (FP16/FP32), making it well-suited for accelerating deep learning tasks. This hardware configuration enables efficient model training and inference, facilitating the experimentation and evaluation of computationally intensive models in a cloud-based environment.

5.3 State-of-the-Art Performance Comparison

We compare state-of-the-art models for stress classification on the Dreaddit dataset (Turcan and McKeown, 2019), focusing on centralized and FL approaches. The evaluated models include variations of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), as well as domain-specific models like MentalBERT and MentalRoBERTa (Ji et al., 2021). Table 2 summarizes the F1 scores and model sizes, while Figure 3 visualizes the accuracy of these models. MentalBERT achieves the highest performance with an F1 score of 94.62, demonstrating the effectiveness of domain-specific pre-training for stress detection tasks. Among general-purpose models, RoBERTa-base outperforms others with an F1 score of 82.38.

In the FL setting, models generally exhibit slightly lower performance compared to their centralized counterparts due to the decentralized nature of training. For example, BERT-base-uncased$_{\text{FL}}$ achieves an F1 score of 80.42, compared to 81.88 for the centralized version. The reduced performance in FL is partially attributed to the limitations of distributed devices, such as mobile phones, which often have constrained computational power, memory, and bandwidth. As a result, smaller models like MobileBERT$_{\text{FL}}$ and MiniLM$_{\text{FL}}$ are employed for FL to ensure feasibility on resource-constrained devices. These models achieve lower accuracy scores of around 51%, reflecting the trade-offs between model size, computational efficiency, and performance in FL scenarios (Li et al., 2020; Kairouz et al., 2021).

These results highlight the importance of balancing model performance and deployment constraints when applying stress detection models in real-world applications. Selecting appropriate model architectures and training methodologies is crucial for achieving optimal performance, par-

Table 2: Model Performance Comparison on Dreaddit dataset

Model	**Parameters**	**F1 Score**
BERT-base (Turcan and McKeown, 2019)	110M	80.65
BERT-base-uncased	110M	81.88
RoBERTa-base	125M	82.38
MentalBERT (Yang et al., 2023c)	110M	94.62
MentalRoBERTa (Yang et al., 2023c)	110M	81.76
BERT-base-uncased$_{\text{FL}}$	110M	80.42
RoBERTa-base$_{\text{FL}}$	125M	77.91
MobileBERT$_{\text{FL}}$	25M	67.11
MiniLM$_{\text{FL}}$	22M	68.08

Table 3: Evaluation of BERT-base-uncased Across Federated Learning Ablation Scenarios on Dreaddit Dataset

Num Clients	**Client Epochs**	**Global Epochs**	**Eval Accuracy**	**Eval F1**
1	3	10	0.6713	0.6999
1	10	3	0.7832	0.7947
3	10	3	0.7944	0.8043

ticularly in distributed environments with limited computational resources.

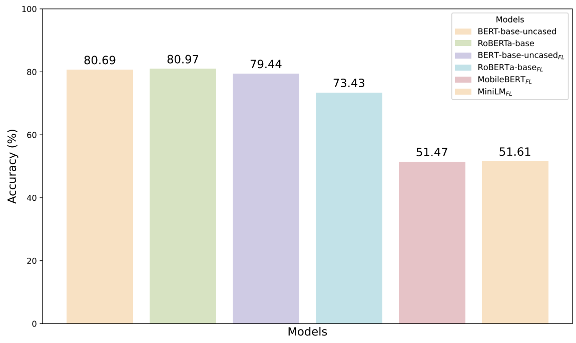

Figure 3: Model Accuracy Comparison.

5.4 Ablation Study

We conduct an ablation study to evaluate BERT-base-uncased performance under different FL configurations on the Dreaddit dataset (Turcan and McKeown, 2019). The study explores variations in the number of clients, local training epochs, and global aggregation rounds. When training with a single client for 3 local epochs and 10 global epochs, the model achieves an accuracy of 67.13% and an F1 score of 69.99%. Increasing the local epochs to 10 while reducing the global epochs to 3 improves the accuracy to 78.32% and the F1 score to 79.47%, indicating the benefits of extended local training. Incorporating 3 clients with the same configuration yields the best results, with an accuracy of 79.44% and an F1 score of 80.43%. These results, summarized in Table 3, demonstrate that increasing the number of clients and local training epochs enhances model performance by leveraging diverse data distributions.

6 Conclusion

In this paper, we introduced FedMentalCare, a privacy-preserving Federated Learning framework that fine-tunes LLMs for mental healthcare applications. By integrating Low-Rank Adaptation (LoRA), FedMentalCare effectively minimizes computational and communication overhead, enabling deployment on resource-constrained devices while ensuring compliance with data privacy regulations such as HIPAA and GDPR. Experimental results on the Dreaddit dataset demonstrate that federated learning supports privacy-aware mental health analysis with only minor performance trade-offs compared to centralized training. Despite these promising results, limitations remain, including the need for larger federated datasets and potential performance degradation caused by heterogeneous data distributions among clients. Future work will focus on enhancing model robustness in non-IID data settings and integrating differential privacy techniques to further bolster data security.

7 Acknowledgement

Author Sarwar gratefully acknowledges the UMBC CS Department for providing financial support through a Graduate Assistantship.

References

104th United States Congress. 1996. H.R.3103 - Health Insurance Portability and Accountability Act of 1996.

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. 2022. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23. Doi: 10.1145/3501813.

Isaac A Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T Chang, Yang Sun, Ann Shue, Jonathan C Chou, Emily Schehlein, Karen L Christopher, et al. 2023. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open*, 6(8):e2330320–e2330320.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Yue Cui, Zhuohang Li, Luyang Liu, Jiaxin Zhang, and Jian Liu. 2022. Privacy-preserving speech-based depression diagnosis via federated learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1371–1374. IEEE.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. 2023. FATE-LLM: A Industrial Grade Federated Learning Framework for Large Language Models. *arXiv preprint arXiv:2310.10049*.

Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, et al. 2023. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*.

Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language model for mental health: A systematic review. *arXiv preprint arXiv:2403.15401*.

Shanshan Han, Baturalp Buyukates, Zijian Hu, Han Jin, Weizhao Jin, Lichao Sun, Xiaoyang Wang, Wenxuan Wu, Chulin Xie, Yuhang Yao, et al. 2023. FedSecurity: A Benchmark for Attacks and Defenses in Federated Learning and Federated LLMs.

Towards Healthcare. 2023. Chatbots For Mental Health and Therapy Market Size Envisioned at USD 6.51 Billion by 2032. Accessed: Apr 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

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

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large Language Models in Mental Health Care: a Scoping Review. *arXiv preprint arXiv:2401.02984*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv preprint arXiv:1909.10351*.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and Open Problems in Federated Learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv preprint arXiv:1610.02527*.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023a. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Jinli Li, Ran Zhang, Mingcan Cen, Xunao Wang, and M Jiang. 2021. Depression detection using asynchronous federated optimization. In *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 758–765. IEEE.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. 2023. Drugchat: towards enabling chatgpt-like capabilities on drug molecule graphs. *arXiv preprint arXiv:2309.03907*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. ChatCounselor: A Large Language Models for Mental Health Support. *arXiv preprint arXiv:2309.15461*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023b. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. *arXiv preprint arXiv:2303.11032*.

Alexander Marrapese, Basem Suleiman, Imdad Ullah, and Juno Kim. 2024. A Novel Nuanced Conversation Evaluation Framework for Large Language Models in Mental Health. *arXiv preprint arXiv:2403.09705*.

Richard May and Kerstin Denecke. 2022. Security, privacy, and healthcare-related conversational agents: a scoping review. *Informatics for Health and Social Care*, 47(2):194–210.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Jennifer Nicholas, Sandersan Onie, and Mark E Larsen. 2020. Ethics and Privacy in Social Media Research for Mental Health. *Current psychiatry reports*, 22:1–7.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375*.

Institute of Health Metrics and Evaluation. 2023. Global Health Data Exchange (GHDx). Accessed: Apr 2024.

OpenAI. 2023. Gpt-4 technical report. (arXiv:2303.08774).

Le Peng, Gaoxiang Luo, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Rui Zhang, and Ju Sun. 2023. An In-Depth Evaluation of Federated Learning on Biomedical Natural Language Processing. *medRxiv*, pages 2023–11.

Kilian Pfeiffer, Mohamed Aboelenien Ahmed, Ramin Khalili, and Jörg Henkel. 2024. Efficient federated finetuning of tiny transformers with resource-constrained devices. *arXiv preprint arXiv:2411.07826*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B Rawat, and Vladimir Vlassov. 2023. Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. *IEEE Internet of Things Journal*.

Regulation (EU) 2016/679 of the European Parliament and of the Council. 2016. General Data Protection Regulation.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7. Doi: `10.1038/s41746-020-00323-1`.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Minh N Vu, Truc Nguyen, Tre'R Jeter, and My T Thai. 2024. Analysis of Privacy Leakage in Federated Large Language Models. *arXiv preprint arXiv:2403.04784*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2023. Mental-llm: Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Hao Yang, Jiaxi Li, Siru Liu, Lei Du, Xiali Liu, Yong Huang, Qingke Shi, and Jialin Liu. 2023a. Exploring the potential of large language models in personalized diabetes treatment strategies. *medRxiv*, pages 2023–06.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023b. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023c. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning. *arXiv preprint arXiv:2402.06954*.

Sixing Yu, J Pablo Muñoz, and Ali Jannesari. 2023. Federated Foundation Models: Privacy-Preserving and Collaborative Learning for Large Models. *arXiv preprint arXiv:2305.11414*.

Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Mahta Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, and Yike Guo. 2023. The potential and pitfalls of using a large language model such as chatgpt or gpt-4 as a clinical assistant. *arXiv preprint arXiv:2307.08152*.