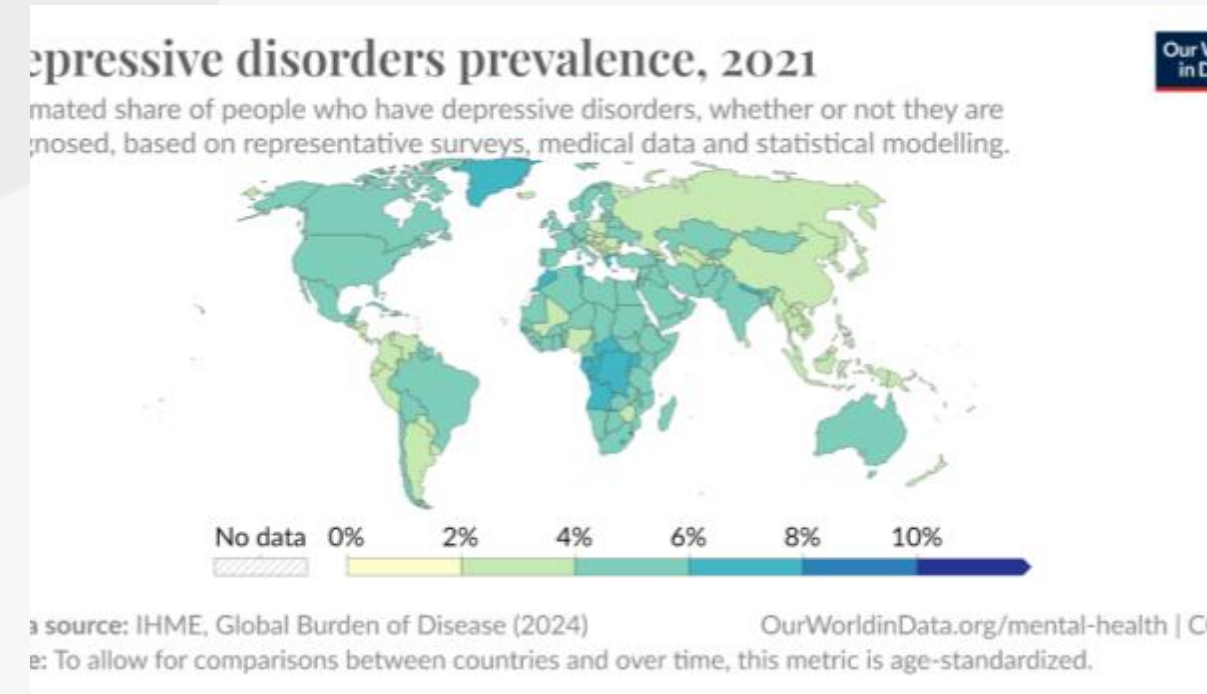




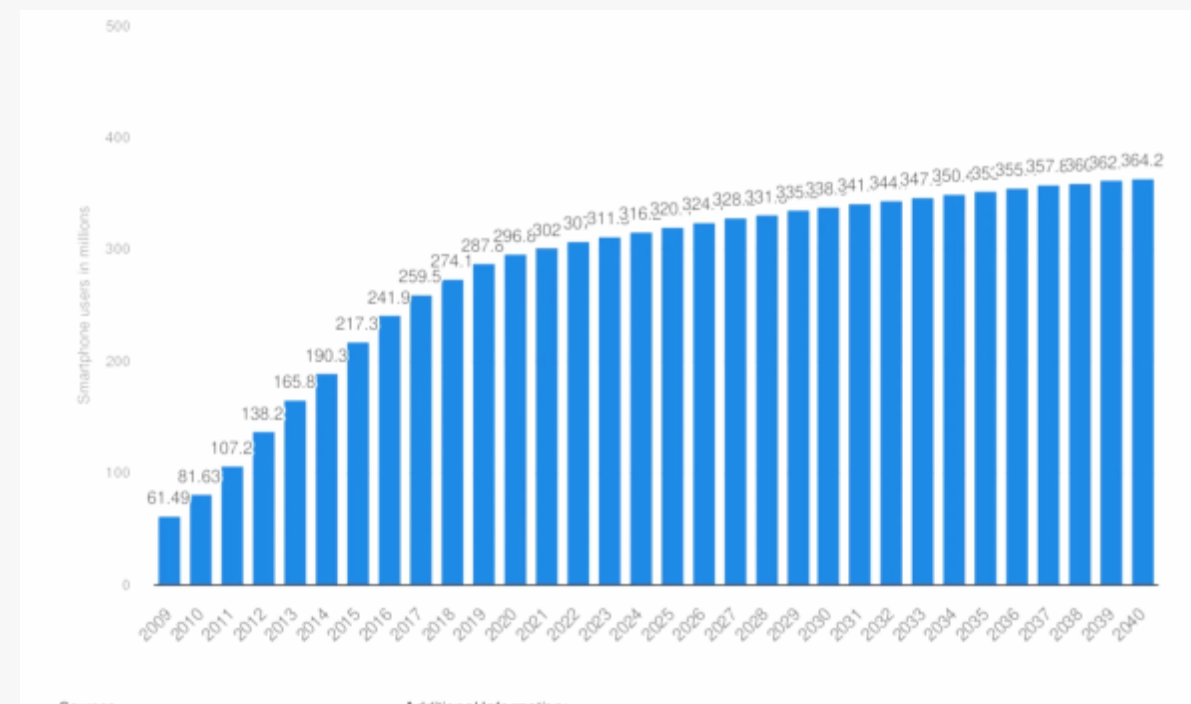
P A P E R R E V I E W



FedTherapist: Mental Health Monitoring with User-Generated Linguistic Expressions on Smartphones via Federated Learning



EXPRESSIVE DISORDERS PREVEALENCE



POTENTIAL OF SMARTPHONE DATA

Introduction

01

IMPORTANCE OF MENTAL HEALTH MONITORING

- Need for early detection and treatment of mental illnesses.
- Most patients have been aware of their disorder for years which delays streatement while the symptoms worsen

02

POTENTIAL OF SMARTPHONE DATA

- Over 6 billion people worldwide use smartphones.
- Potential to utilize linguistic data (speech and keyboard inputs).
- (LiKamWa et al., 2013; Wang et al., 2014, 2018; Li and Sano, 2020; Tlachac et al., 2022a).

Previous Research

- Social Media-Based Studies:



- Examples: Reddit, Twitter, Facebook, Instagram.
- Limitations: Only active users are covered, and posts often reflect idealized self-representations.

- Non-Text Data-Based Studies:



- Use of sensor data (e.g., location, heart rate).
- Lack of text data utilization and sophisticated language models.

- Smartphone-Based Mental Health Monitoring:



- Features Used: Phone usage patterns, location, and activity.
- Studies: LiKamWa et al., 2013; Wang et al., 2014, 2018; Li and Sano, 2020; Tlachac et al., 2022a.
- Limitation: These features do not capture the diagnostic methods used by licensed psychiatrists, who assess mental disorders through patient conversations.

- Linguistic Analysis for Mental Health Monitoring:




- Advantage: Analyzing users' language use is ideal for monitoring mental health.
- Challenge: Significant privacy concerns hinder the collection of sufficient data needed to train advanced Natural Language Processing (NLP) neural networks (e.g., Devlin et al., 2019).






Purpose of the Study

To develop and evaluate FedTherapist, a privacy preserving mobile mental health monitoring system that leverages user-generated text data from smartphones.

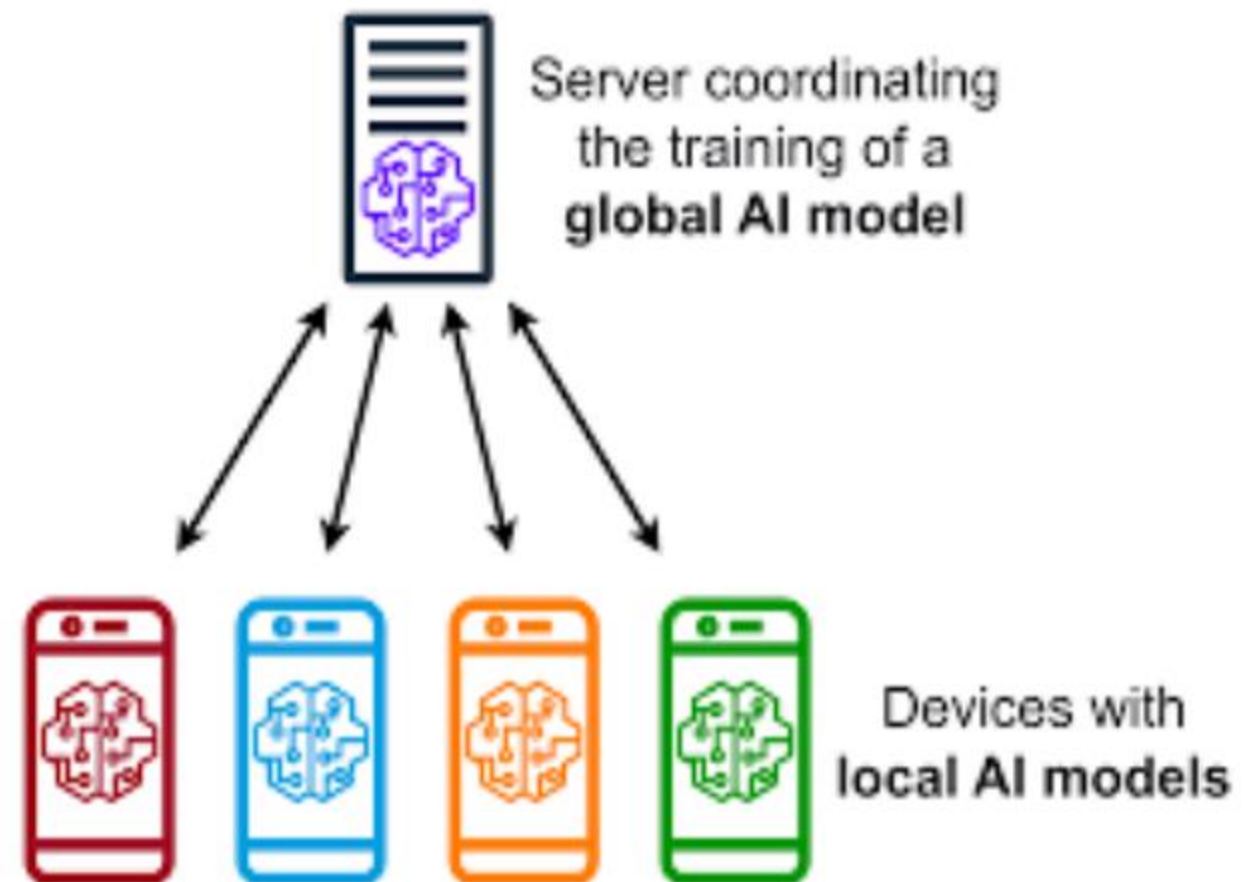
- Improve Mental Health Monitoring
 - Protect User Privacy
 - Utilize Context-Aware Insights
 - To explore the most suitable design
 - Evaluating smartphone resource consumption
- 

Key Feature FedTherapist

-  Privacy-Preserving Mechanism: Federated Learning (FL)
Description: Models are trained locally on user devices without transmitting personal data to a central server. Only encrypted model updates are aggregated, ensuring the security of individual information.
-  Enhanced Analysis: Context-Aware Language Learning (CALL)
Description: Integrates user context information—such as time, location, mobility status, and app usage—to perform deeper analyses. Employs multiple contextual models and ensemble learning techniques to accurately extract mental health signals.
-  Comprehensive Data Utilization
Description: Incorporates both speech and keyboard input data from all smartphone applications to evaluate users' mental health status accurately.

FL

(Federated Learning)



- Definition: A decentralized approach to model training where the model is trained across multiple client devices (e.g., smartphones) using locally stored data, ensuring that personal data remains on the device.

FedAvg

Algorithm 1 FederatedAveraging. The K clients are indexed by k ; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:

initialize w_0

for each round $t = 1, 2, \dots$ **do**

$m \leftarrow \max(C \cdot K, 1)$

$S_t \leftarrow$ (random set of m clients)

for each client $k \in S_t$ **in parallel do**

$w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$

$w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w): // Run on client k

$\mathcal{B} \leftarrow$ (split \mathcal{P}_k into batches of size B)

for each local epoch i from 1 to E **do**

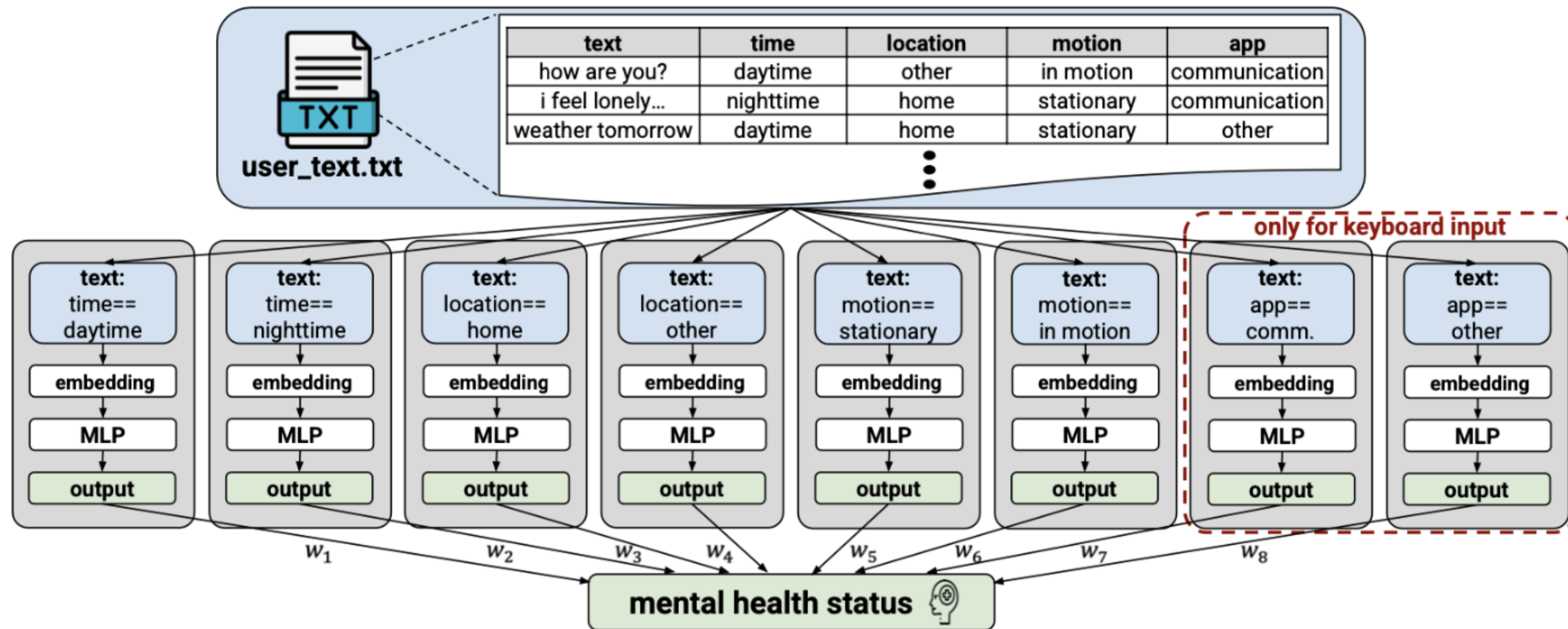
for batch $b \in \mathcal{B}$ **do**

$w \leftarrow w - \eta \nabla \ell(w; b)$

 return w to server

1. Client Sampling
2. Model Download and Local Training
3. Model Update and Upload
4. Server Model Aggregation

Context-Aware Language Learning(CALL)

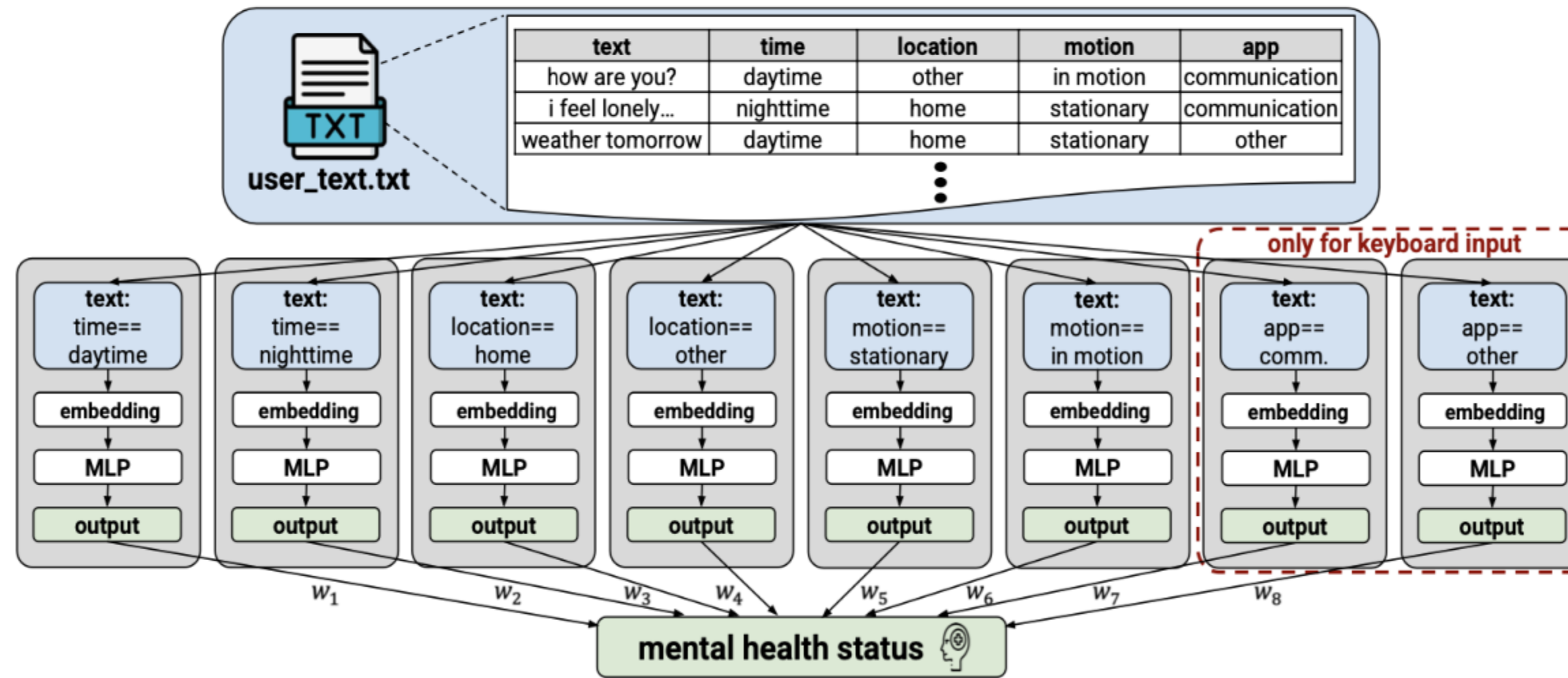


Context-Aware Language Learning (CALL)



Call integrates various temporal contexts of users (e.g., time, location) captured on smartphones to enhance the model's ability to sense mental health signals from the text data.

CALL - (1)



- The core idea of CALL is to enable neural networks to focus on text generated in contexts where users are likely to reveal their mental health status.
- CALL is designed to classify and analyze text data based on four contexts: time, location, movement, and application, to effectively detect mental health signals. This integration helps better reflect emotions and mental states in user-generated data.
- CALL performs ensemble learning on N context models and takes a weighted sum on the models' outputs to determine the user's mental health status.

CALL - (2)

1. Input Data Processing : User-generated text is stored in user_text.txt.
2. Contextual Model Training: The text is embedded and processed through a Multi-Layer Perceptron (MLP) based on four contexts: time, location, movement, and application.
3. Output Integration : Outputs from each contextual model are combined using specific weights (e.g., w_1 , w_2 , ..., w_8) to reflect their importance.
4. Keyboard Input Processing : The area marked with a red dashed line indicates models related to contexts applicable only to keyboard inputs (e.g., application usage type).

CALL - (3)

- Model Settings

- N Value Configuration:

- For keyboard input: Two contexts from each of T, L, M, A, totaling $N = 8$.
 - For speech input: Excluding application contexts, totaling $N = 6$.
 - For both speech and keyboard input: Including all contexts, totaling $N = 14$.

- Ensemble Weight Types:

- EA (Ensemble Averaging): Averages the model outputs with equal weights: $w_1 = w_2 = \dots = w_N = 1/N$.
 - EE (Weight Sum of model output): Applies trained ensemble weights over Federated Learning (FL)

to compute a weighted sum of model outputs.

- Conclusion: By integrating various temporal, spatial, and behavioral contexts into text data analysis, CALL accurately detects mental health signals. This enables FedTherapist to monitor users' mental health status more effectively.

CALL - (4)

Aspect	Standrard Federated Learning	Secure Aggregation Protocol
Individual Gradient Exposure	The server can access individual users' gradients.	The server cannot access individual users' gradients.
Privacy Protection Level	Low; potential risk of data leakage.	High; data is protected through encryption.
Computational Complexity	Low; no additional encryption processes involved.	High; additional encryption and decryption processes required.
Application Purpose	General federated learning environments.	Handling sensitive data where privacy is crucial (e.g., healthcare, mental health).

Data Collection

Collection method

- Purpose: To evaluate FedTherapist using real user data, a data collection study was conducted with IRB approval.
- Participants: 52 native English speakers recruited via Amazon Mechanical Turk (MTurk).
- Procedure: Participants used an Android application to collect data over 10 days.
- Data Collected: Naturalistic data generated during participants' daily activities.



Input data

- Speech: user-spoken words on the microphone-equipped smartphones
- Keyboard: typed characters on smartphone keyboards
- Used in a mobile depression detection study (Wang et al., 2018)
- The count of the participant responses on each mental health status

Data Collection

Collection
method

Input data

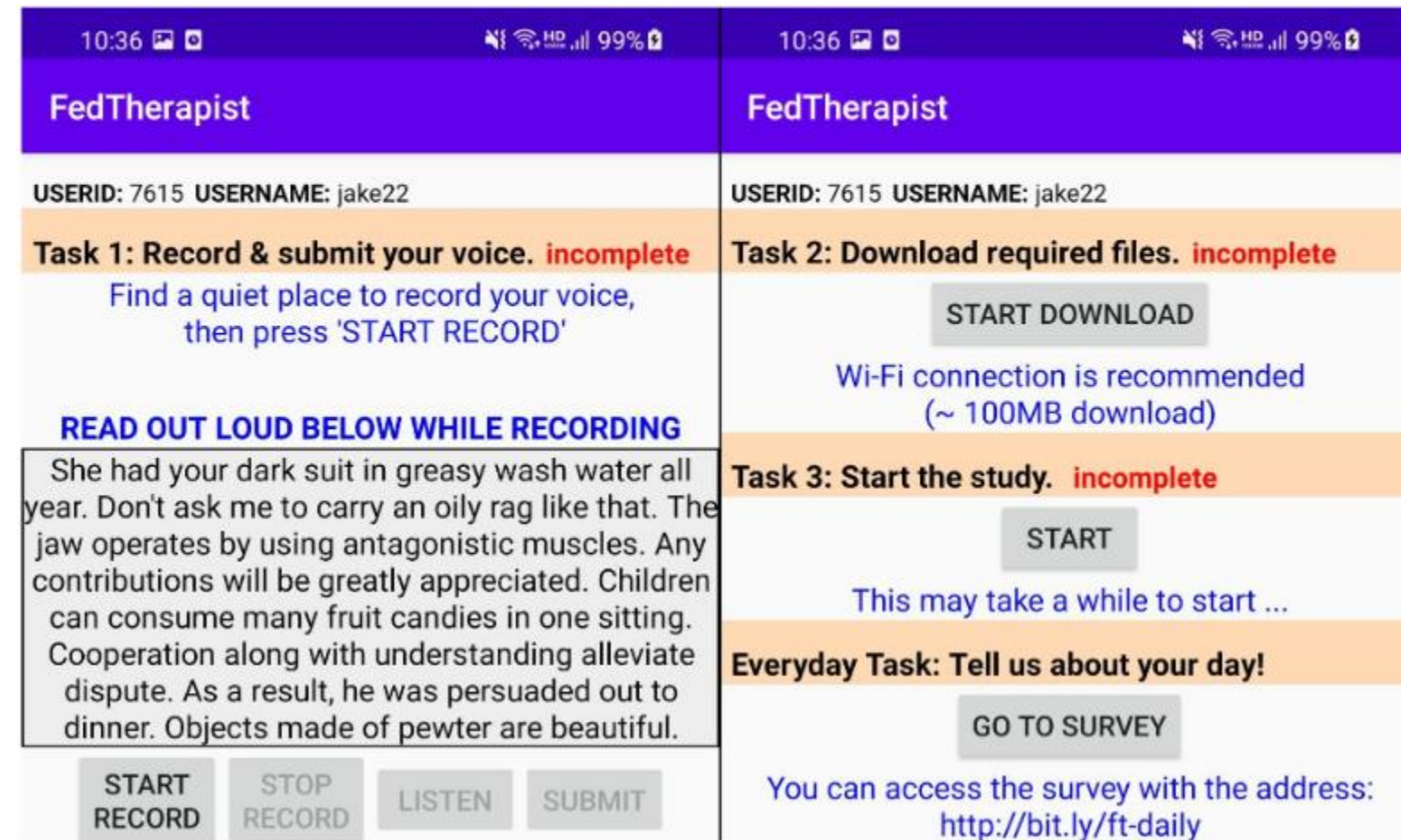


Figure 3: Screenshots of our data collection application. Participants completed three tasks to join the study. First, participants were asked to upload their voice sample (Task 1) to use VoiceFilter (Wang et al., 2019) (Section C.1). Participants then downloaded the model files used in our local data collection module (Task 2), and started the study (Task 3). We asked the participants for a daily mental health status report during the study.

Data Collection

Collection method

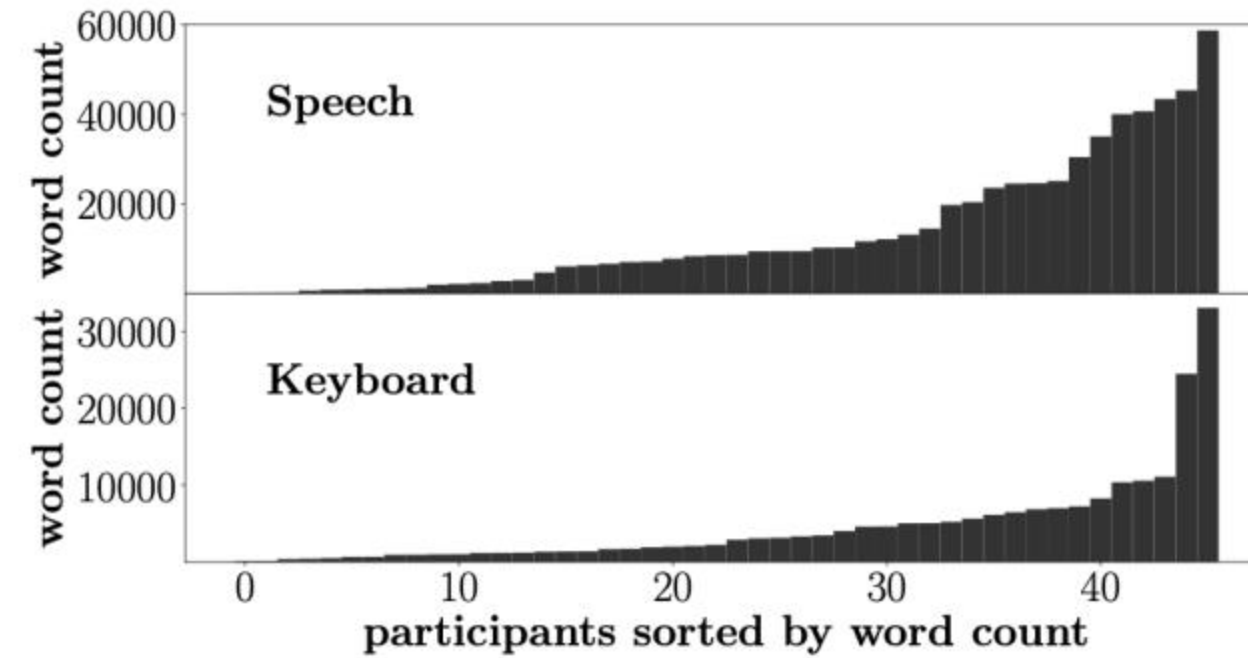


Figure 4: Word count of each participant's collected speech and keyboard input from the 10-day study.

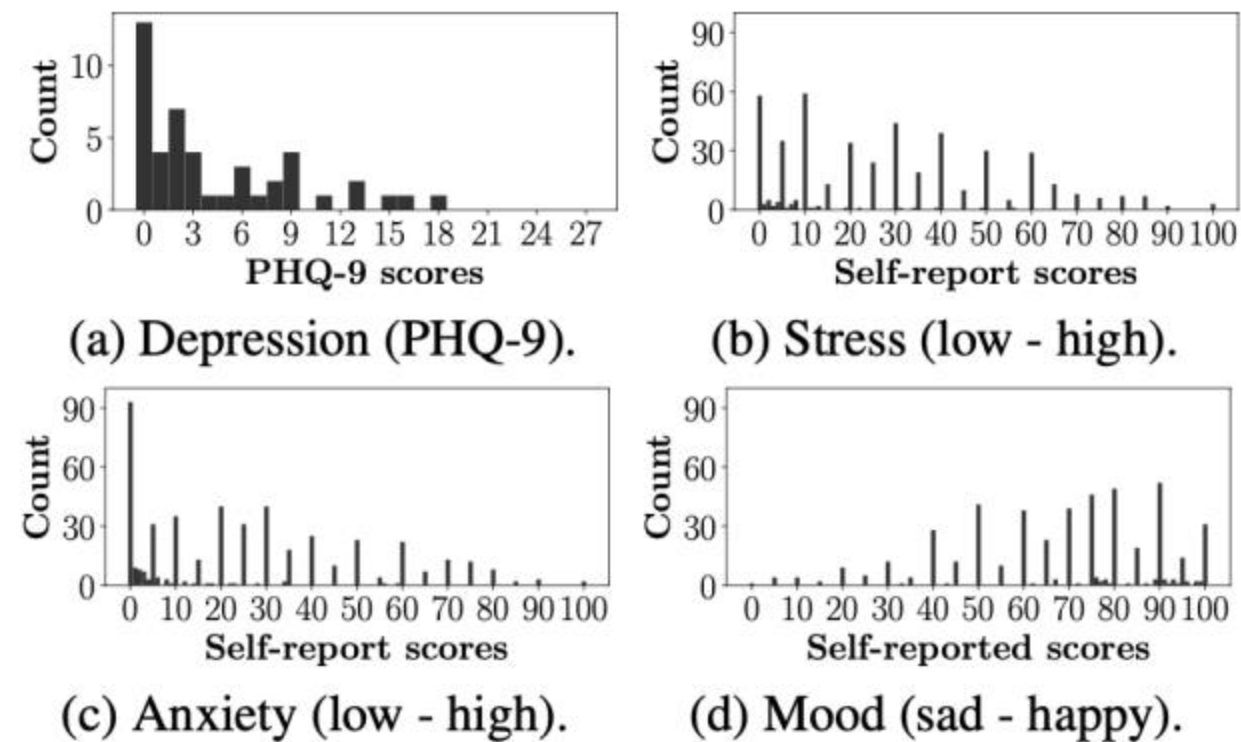


Figure 5: Count of participant responses on different mental health status scores.

Data Collection (Speech)

Challenges

- Significant smartphone battery drain
- The voice of non-target users could be included in the audio
- The user voice could be recorded in low quality with noise, as the recording occurs in unconstrained, real-world environments



Solution

- To avoid continuous recording, we adopted a duty cycling system with a Voice Activity Detector (VAD) model
- The module continues recording if a conversation is detected, where the recorded audio is inferred on VoiceFilter
- Designed the module to leverage a language classifier that outputs the language of a given audio file

Data Collection (Keyboard)

- The keyboard text collection module captures all the input text the user enters with their smartphone keyboard. Our implementation on our app tracked the text input events on Android smartphones using the Android Accessibility Service (Android, 2021).
1. Removal of Specific Elements: Eliminated emails, hashtags, links, mentions, punctuation, and numbers.
 2. Emoji and Abbreviation Handling: Replaced emojis with their short names from the Common Locale Data Repository (CLDR). Expanded abbreviations to their full meanings.
 3. Excessive Character Repetition Handling : Simplified overly repeated characters (e.g., "loooove" to "love").
 4. Typographical Error Correction: Utilized an auto-correction API to correct spelling errors.

Handling Long Input Sequences

FULL + PULL

- Chunking
- Embedding and Max Pooling
- Final Embedding Vector
- GPU Memory Consideration



SINGLE

- Individual Chunk Processing
- Logit Averaging

Comparison of Text Embedding Models

Methods Parameter Size	ALBERT-base 11M	MobileBERT 25M	DistilBERT-base 66M	RoBERTa-base 110M	BERT-base 110M	BERT-large 340M	BART-base 140M	BigBird 128M
Depression (AUROC \uparrow)	0.526 \pm 0.007	0.319 \pm 0.003	0.775 \pm 0.010	0.729 \pm 0.002	0.560 \pm 0.011	0.622 \pm 0.004	0.636 \pm 0.005	0.626 \pm 0.009
Stress (MAE \downarrow)	24.75 \pm 0.02	23.98 \pm 0.01	24.78 \pm 0.03	24.59 \pm 0.05	24.80 \pm 0.06	24.76 \pm 0.04	24.29 \pm 0.05	24.95 \pm 0.05
Anxiety (MAE \downarrow)	26.95 \pm 0.02	25.99 \pm 0.01	27.30 \pm 0.03	27.05 \pm 0.06	27.07 \pm 0.04	27.05 \pm 0.03	26.74 \pm 0.04	27.25 \pm 0.04
Mood (MAE \downarrow)	23.57 \pm 0.02	22.82 \pm 0.01	23.67 \pm 0.02	23.79 \pm 0.02	23.54 \pm 0.02	23.01 \pm 0.03	23.09 \pm 0.02	23.58 \pm 0.04

Table 3: Mental health monitoring performance on different text embedding methods at epoch 200.

- DistilBERT: Demonstrated superior performance in depression detection tasks.
- MobileBERT: Achieved the highest performance in stress, anxiety, and mood tasks.
- FedTherapist Model Selection: Due to MobileBERT's lower performance in depression detection, we selected DistilBERT for the FedTherapist evaluations
- These performance variations are likely due to differences in pre-training data and methodologies among the models.

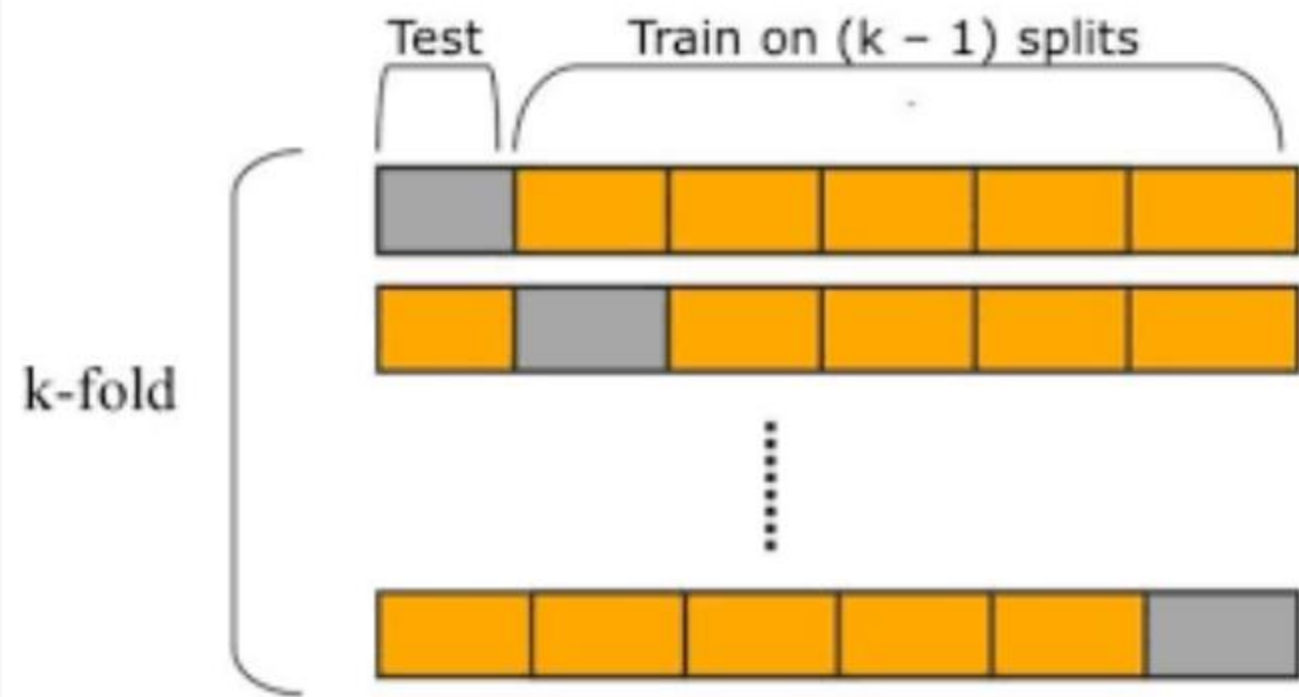
Fine-tuning Bert

Methods	BERT-base	RoBERTa-base	DistilBERT-base
Pretrained	0.560	0.729	0.775
+ SODA	0.757	0.696	0.701
+ SODA-single	0.744	0.665	0.714

Table 4: Effect of fine-tuning on text embedding methods on mental health monitoring performance, measured in AUROC on *depression* task.

- Three pretrained : Bert-base, RoBERTa, DistilBert
- Two Fine-tuning : SODA, SODA-single
- BERT-base: Exhibited significant performance improvements after fine-tuning. Achieved the highest AUROC in the depression detection task.
- RoBERTa-base and DistilBERT-base: Both models demonstrated similar performance levels. Recorded an AUROC of approximately 0.7.
- SODA and SODA-single resulted in similar performances

CrossValidation (LOUO)



- Data Partitioning: Designated one participant as the test user and the remaining participants as training users.
- Model Training: For CL and FL, trained the model for 1,000 epochs , round using the training users' data.
- Iterative Testing: Sequentially designated each participant as the test user, repeating the process until every participant had been tested.
- Performance Evaluation: Repeated the experiments with three different random seeds.
- Averaged the performance metrics across these runs to report the final results.

Training

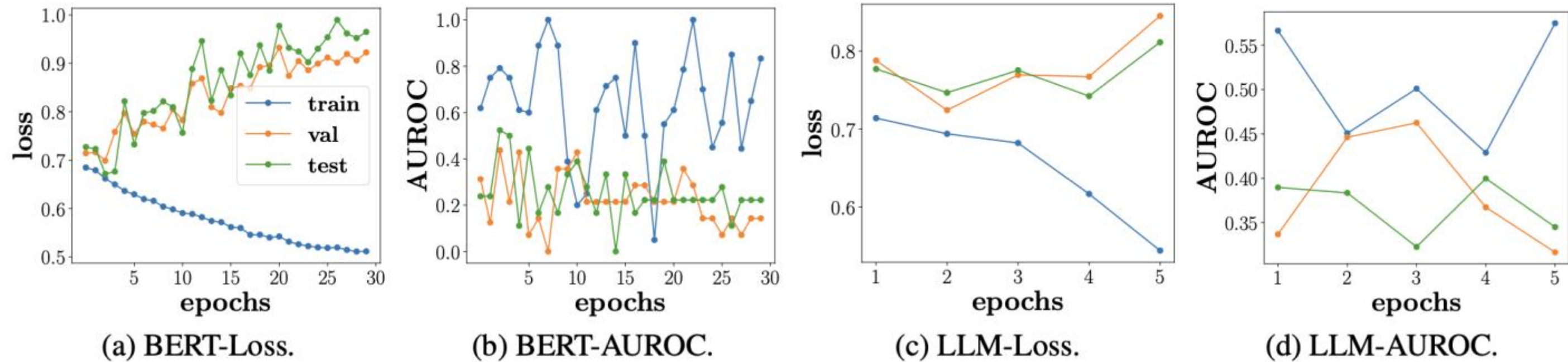


Figure 6: Loss and AUROC results on train, validation, and test samples over epochs from fine-tuning experiments in Section 3.1. Figure 6a and 6b indicate fine-tuning experiment of *End-to-End BERT + MLP* using a pre-trained RoBERTa-base (Liu et al., 2019) model with a learning rate of 0.0001. Figure 6c and 6d depict fine-tuning experiment of *LLM* on a pre-trained LLaMa-7B model with a learning rate of 0.0002. More details on experimental methods are found in Section 3.1 and Appendix D. All subgraphs share the same legend with Figure 6a.

FL

(Federated Learning)

Methods	Performance	Smartphone Overhead*	
	Depression [†]	CPU [‡]	Memory
Fixed-BERT + MLP	0.716	35%	219MB
End-to-End BERT + MLP	0.524	68%	864MB
LLM	0.406	N/A [§]	N/A [§]

Table 1: Comparison of candidate mental health prediction methods. *: averaged measurement on Google Pixel (2016) and Samsung Galaxy S21 (2021). [†]: measured in AUROC; [‡]: measured in average per-core utilization; [§]: failed loading on test smartphones.

1. Fixed-BERT + MLP: The model consists of a text embedding model followed by a multilayer perceptron (MLP).
2. End-to-End BERT + MLP: We use the same model architecture as the former but perform end-to-end training, including BERT and MLP.
3. LLM: Given recent breakthroughs and state-of-the-art performance of Large Language Models (LLMs) (OpenAI, 2023)
4. LOUO(Leave-One-User-Out)

FL

(Federated Learning)

Methods	Performance	Smartphone Overhead*	
	Depression [†]	CPU [‡]	Memory
Fixed-BERT + MLP	0.716	35%	219MB
End-to-End BERT + MLP	0.524	68%	864MB
LLM	0.406	N/A [§]	N/A [§]

Table 1: Comparison of candidate mental health prediction methods. *: averaged measurement on Google Pixel (2016) and Samsung Galaxy S21 (2021). [†]: measured in AUROC; [‡]: measured in average per-core utilization; [§]: failed loading on test smartphones.

- AUROC (Area Under the Receiver Operating Characteristic Curve):
 - Fixed-BERT + MLP: Achieved the highest AUROC, outperforming End-to-End BERT + MLP by 0.192 and LLM by 0.310 in depression detection.
- Findings:
 - Fixed-BERT + MLP is more suitable for smartphone deployment due to its superior performance in depression detection and efficient resource utilization.
 - The subpar performance of End-to-End BERT + MLP and LLM is hypothesized to result from overfitting on the limited training sample of 40 users.

Model Evaluation - (1)

Methods	CL+NonText	FL+Text			<i>FedTherapist</i> (FL+Text+ <i>CALL</i>)		
Data Type	Non Text Data	Speech (S)	Keyboard (K)	S+K	Speech (S)	Keyboard (K)	S+K
Depression (AUROC \uparrow)	0.625 \pm 0.010	0.627 \pm 0.004	0.710 \pm 0.007	0.775\pm0.010	0.571 \pm 0.003	0.746 \pm 0.000	0.721 \pm 0.008
Stress (MAE \downarrow)	20.83 \pm 0.03	23.44 \pm 0.02	24.21 \pm 0.05	24.78 \pm 0.03	21.34 \pm 0.01	20.07 \pm 0.01	19.12\pm0.01
Anxiety (MAE \downarrow)	20.95 \pm 0.06	25.80 \pm 0.03	26.85 \pm 0.08	27.30 \pm 0.03	22.56 \pm 0.05	21.39 \pm 0.01	20.56\pm0.02
Mood (MAE \downarrow)	18.76 \pm 0.09	22.85 \pm 0.03	23.21 \pm 0.02	23.67 \pm 0.02	19.11 \pm 0.02	19.18 \pm 0.00	18.57\pm0.02

Table 2: Mental health monitoring performance on different methods.

- CL + NonText : Centralized Learning (CL) -> Utilizes non-textual features to train an MLP model.
- FL + Text: Federated Learning (FL)-> Employs text data from speech and keyboard inputs to train a Fixed-BERT + MLP model.
- FedTherapist : Enhanced FL + Text -> Applies Context-Aware Language Learning (CALL) to the FL + Text approach.
- Data Variants Tested:
 - S (Speech Input): Utilizes only speech data.
 - K (Keyboard Input): Utilizes only keyboard data.
 - S + K (Both): Combines both speech and keyboard data.

Model Evaluation - (2)

Methods	CL+NonText	FL+Text			<i>FedTherapist</i> (FL+Text+ <i>CALL</i>)		
Data Type	Non Text Data	Speech (S)	Keyboard (K)	S+K	Speech (S)	Keyboard (K)	S+K
Depression (AUROC \uparrow)	0.625 \pm 0.010	0.627 \pm 0.004	0.710 \pm 0.007	0.775\pm0.010	0.571 \pm 0.003	0.746 \pm 0.000	0.721 \pm 0.008
Stress (MAE \downarrow)	20.83 \pm 0.03	23.44 \pm 0.02	24.21 \pm 0.05	24.78 \pm 0.03	21.34 \pm 0.01	20.07 \pm 0.01	19.12\pm0.01
Anxiety (MAE \downarrow)	20.95 \pm 0.06	25.80 \pm 0.03	26.85 \pm 0.08	27.30 \pm 0.03	22.56 \pm 0.05	21.39 \pm 0.01	20.56\pm0.02
Mood (MAE \downarrow)	18.76 \pm 0.09	22.85 \pm 0.03	23.21 \pm 0.02	23.67 \pm 0.02	19.11 \pm 0.02	19.18 \pm 0.00	18.57\pm0.02

Table 2: Mental health monitoring performance on different methods.

- Depression Detection (10-Day Data Input): The FL + Text method outperformed CL + NonText, achieving a 0.15 increase in AUROC.
- Other Mental Health Tasks (1-Day Data Input): CL + NonText demonstrated superior performance over FL + Text, with a reduction in Mean Absolute Error (MAE) ranging from 2.61 to 6.35
- These results suggest that longer-duration text inputs are more effective in capturing mental health signals, likely due to the complexity of detecting these signals within noisy text data. Models may struggle with shorter-term data, highlighting the importance of extended observation periods for accurate mental health assessment.

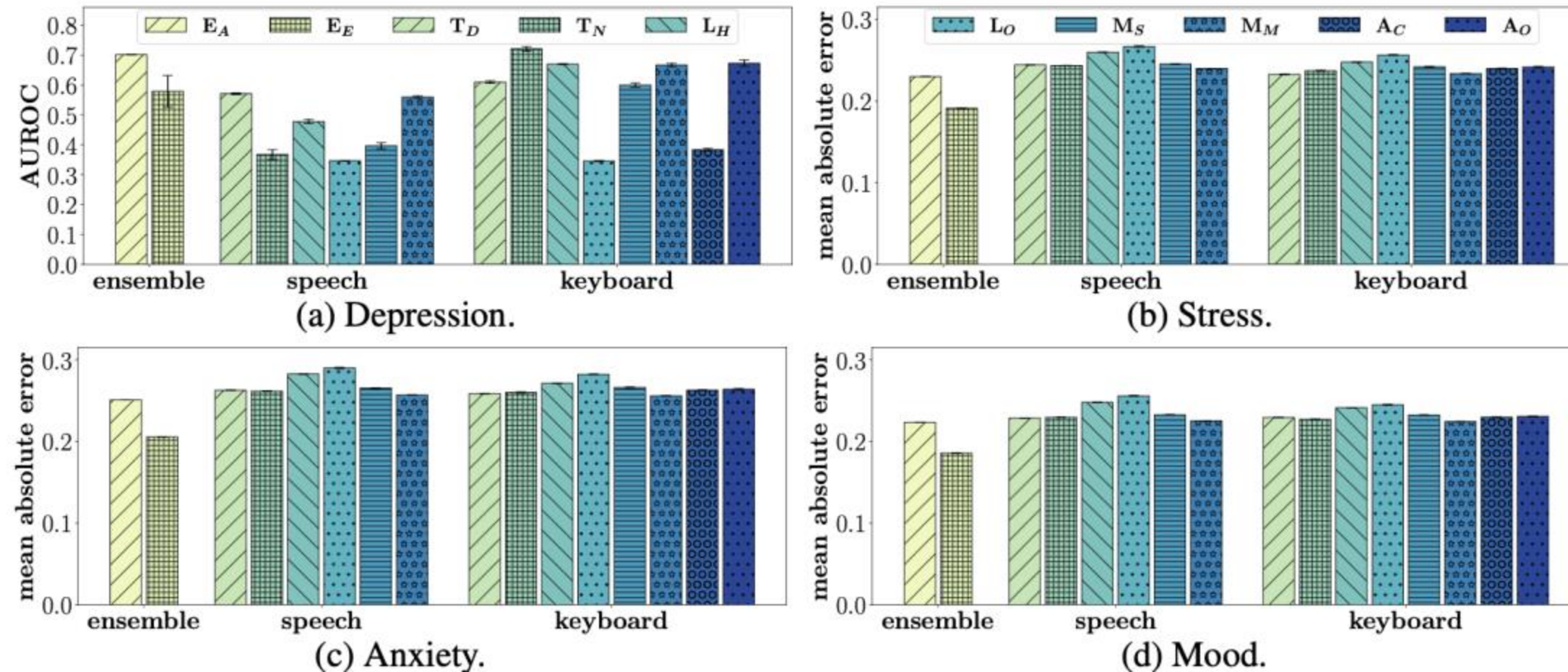
Model Evaluation - (3)

Methods	CL+NonText	FL+Text			<i>FedTherapist</i> (FL+Text+ <i>CALL</i>)		
Data Type	Non Text Data	Speech (S)	Keyboard (K)	S+K	Speech (S)	Keyboard (K)	S+K
Depression (AUROC \uparrow)	0.625 \pm 0.010	0.627 \pm 0.004	0.710 \pm 0.007	0.775\pm0.010	0.571 \pm 0.003	0.746 \pm 0.000	0.721 \pm 0.008
Stress (MAE \downarrow)	20.83 \pm 0.03	23.44 \pm 0.02	24.21 \pm 0.05	24.78 \pm 0.03	21.34 \pm 0.01	20.07 \pm 0.01	19.12\pm0.01
Anxiety (MAE \downarrow)	20.95 \pm 0.06	25.80 \pm 0.03	26.85 \pm 0.08	27.30 \pm 0.03	22.56 \pm 0.05	21.39 \pm 0.01	20.56\pm0.02
Mood (MAE \downarrow)	18.76 \pm 0.09	22.85 \pm 0.03	23.21 \pm 0.02	23.67 \pm 0.02	19.11 \pm 0.02	19.18 \pm 0.00	18.57\pm0.02

Table 2: Mental health monitoring performance on different methods.

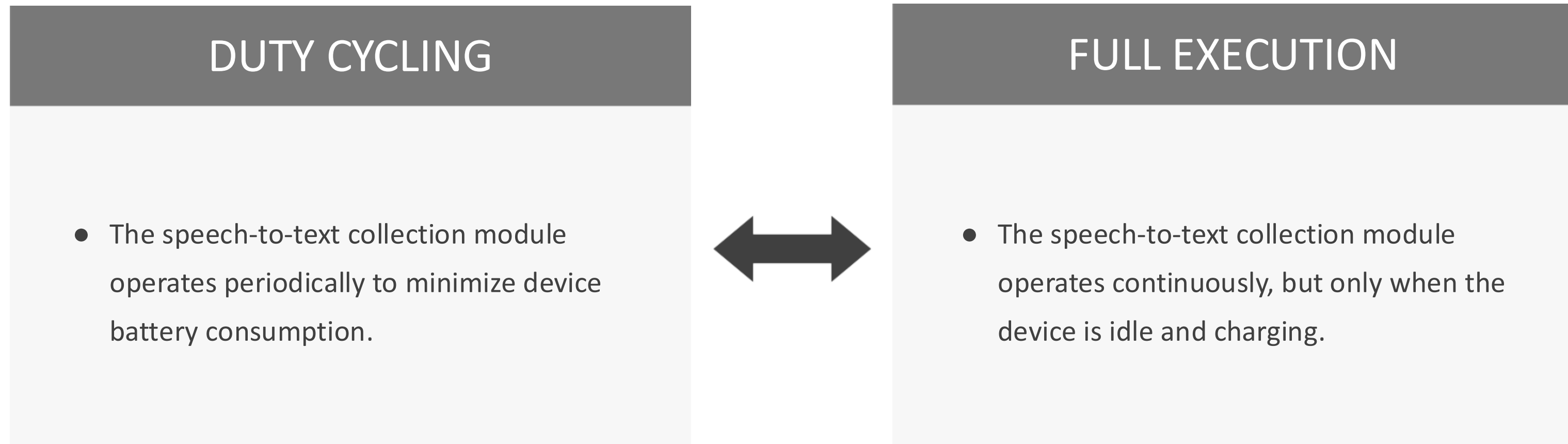
- Depression Detection: Achieves a 0.121 higher AUROC compared to CL+NonText. And Exhibits comparable AUROC to FL+Text (0.746 vs. 0.775).
- Stress, Anxiety, and Mood Assessment : Demonstrates a reduction in MAE by 0.19 to 1.71 compared to CL+NonText. And Shows a significant MAE reduction of 4.32 to 6.74 compared to FL+Text.
- S(Speech) outperforms in mood with FedTherapist
- K(Keyboard) shows superior performance in other tasks

Model Evaluation - (4)



- Ensemble by Averaging (E_A) : Method: Averages the outputs of all contextual models with equal weights.
- Ensemble by Weighted Sum (E_E): Method: Calculates a weighted sum of the outputs from contextual models, assigning different weights to each.
- In tasks related to stress, anxiety, and mood, certain contextual models may not consistently contribute positively.
- The Weighted Ensemble (E_E) method outperforms Ensemble Averaging (E_A) in these three tasks.

System Overhead of FedTherapist



- The measurement scope focused solely on voice data collection, excluding keyboard data collection.
- Since keyboard data collection is a lightweight task that does not involve running deep learning models, it was not measured separately.

System Overhead of FedTherapist

Scenario	Sub-Scenario	Device	CPU (%)	Memory (MB)	Latency (sec)
Data Collection (Section C)	Duty Cycling	Pixel	1.16	119	N/A [†]
		Galaxy S21	0.29	191	N/A [†]
	Speech Text Collection Module	Pixel	49.21	1245	11.20
		Galaxy S21	15.84	1104	9.76
On-Device Training	Fixed-BERT + MLP	Pixel	27.05	215	0.08
		Galaxy S21	42.80	222	0.03
	End-to-End BERT + MLP	Pixel	85.59	842	13.63
		Galaxy S21	49.68	886	38.89
	LLM [‡]	Pixel	N/A	N/A	N/A
		Galaxy S21	N/A	N/A	N/A

Table 5: Smartphone overhead on scenarios of *FedTherapist*: CPU denotes average per-core utilization post 5-min task execution. [†]: restricted to measure individual Voice Activity Detector model execution from its API during duty cycling; [‡]: N/A measurements, as the model is not loaded on tested smartphones due to memory constraints.

- Duty Cycling: The speech-to-text collection module operates periodically to minimize device battery consumption.
- Full Execution: The speech-to-text collection module operates continuously, but only when the device is idle and charging.

Conclusion

- FedTherapist is a mobile system designed to monitor mental health by analyzing continuous speech and keyboard inputs. It ensures user privacy through the implementation of Federated Learning (FL).
- Context-Aware Language Learning (CALL): This methodology enables FedTherapist to effectively detect mental health signals within large and noisy text data generated on smartphones.
- Evaluation Results: In a study involving 46 participants, FedTherapist outperformed models that relied solely on non-linguistic features in predicting depression, stress, anxiety, and mood.
- Significance: This research demonstrates the feasibility of on-device Natural Language Processing (NLP) on smartphones, allowing the utilization of user-generated linguistic data without compromising privacy.

Limitations

- Data Collection Period: While previous studies (Wang et al., 2014, 2018; Li and Sano, 2020) collected data over a span of 10 weeks for mobile mental health monitoring, our study limited data collection to a 10-day period.
- Participant Demographics: The 46 English-speaking participants in our study may not adequately represent regional and cultural linguistic diversity. Therefore, the reported evaluation results should be considered exploratory.
- Participant Characteristics: Our research included participants aged 20 to 60 from the United States and Canada, offering a broader age range compared to prior studies that primarily focused on college students.
- The accuracy of speech data collection in uncontrolled environments was not investigated in this study.
- GPU memory constraints (24GB, NVIDIA RTX 3090) limited the use of larger batch sizes (e.g., 128) during BERT and LLM model training.

AttentionScore

$$\textit{Score}(Q, K) = \left(\frac{QK^T}{\sqrt{d_k}} \right)$$

- The attention score quantifies the relevance between a query and a set of key-value pairs, guiding the model to focus on specific parts of the input sequence.
- Query (Q): Represents the current input seeking context.
- Key (K): Represents the reference inputs.
- Value (V): Contains the actual information corresponding to each key.
- Calculate the dot product between the query and each key to assess similarity

Attention Score Graph



- In the attention heatmap, the "Age" feature on the far left is represented by the brightest color (yellow), corresponding to a value of 0.32. In contrast, the other features have values ranging from 0.1 to 0.13, depicted in a purplish hue, indicating lower attention scores. The attention scores among other features range from 0.1 to 0.13, suggesting that the model perceives the relationships among those features as less important.