The Reports of My Capabilities Are Greatly Exaggerated – Small LLMs for Depression Inference from Mobile Sensing Data

Ilina Kirovska
Faculty of Computer and Information Science, University of Ljubljana
Ljubljana, Slovenia

Veljko Pejović
Faculty of Computer and Information Science, University of Ljubljana
Department of Computer Systems, Institute "Jožef Stefan"
Ljubljana, Slovenia

Abstract

Modern large language models (LLMs) trained on huge amounts of textual data excel in a number of tasks related to generation and analysis of textual and even multimodal input. Automated inferences related to mental health status are crucial as humanity faces an epidemic of mental illness. The potential of LLMs to tackle this problem given texts written by target participants has already been documented. However, the ability of LLMs to infer a user's health status from mobile sensing data alone has received only limited attention, while the practical feasibility of such LLMs running directly on end-user devices has not been addressed. In this paper, we conduct a preliminary analysis of the potential of a state-of-the-art mobile-ready LLM to infer a user's depression level from mobile sensing traces. Our investigation reveals that expectations based on the success of LLMs in other tasks are not justified in the case of inferring mental health status from sensor data. We discuss augmentations that could improve LLM-based inference in the future.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computer systems organization** → *Embedded and cyber-physical systems*; • **Computing methodologies** → **Machine learning**.

Keywords

Large Language Models, Mobile sensing, Depression inference, Ubiquitous computing

1 Introduction

The smartphone has been hailed as a tool poised to revolutionise our understanding of human psychology [9]. Swaths of multimodal sensor data opportunistically collected by smartphones has enabled researchers to investigate links between contextual factors and stress, depression, bipolar disorder, and numerous other aspects of mental health [2, 17]. Yet, uncovering subtle signals pertaining to a user's internal state from seemingly noisy, erratic, and unstructured sensor readings often requires rather complex data processing.

Large language models (LLMs), with an astonishing ability to generalise across domains, have recently arouse as the most promising means of analysing any *textual* data, including that that is related to mental health issues [21]. Nevertheless, focusing on texts, these approaches leave a large pool of mobile *sensing* data untouched.

In 2024 Penetrative AI – a concept where LLMs are used to directly process sensor data – was proposed [19]. Kim et al. [5] embrace this concept and demonstrate that, with proper fine-tuning of the model, sensing data from wearable devices can improve health state prediction accuracy. However, their work mostly focuses on models that are too large for edge-based deployment. Smaller LLMs, containing up to 8 billion parameters, have been ported to run directly on smartphones [20], thus enabling privacy-preserving on-device LLM inference. While different application domains have been recently addressed through on-device penetrative AI, mental health inference remains unexplored. Thus, in this work we tackle this gap. More specifically, we assess the ability of edge device-ready LLMs to infer a user's depression level from smartphone traces including a user's mobility, physical activity, and other features, pre-collected by two publicly available large-scale studies. We explore different LLM prompting strategies, different contextualisations of the prompt, and different means of defining the inference task. The thorough analysis we perform, however, points to significant obstacles to accurate depression inference from mobile sensing data with LLMs.

In summary, the main contribution of this work are:

- We, for the first time, consider the concept of on-device penetrative AI, i.e. LLM-based mobile sensing data processing, for mental health state inference.
- We perform thorough experimentation with different prompting techniques, different contextualisation techniques, different versions of the inference task, and two rich datasets comprising over 50 users.
- Based on our findings we sketch possible avenues for further exploration of penetrative AI for mental health.

Our experiments' code and the data are publicly available at our GitLab repository https://gitlab.fri.uni-lj.si/lrk/llm-mobile-sensing-depression-inference.

2 Related Work

The growing ubiquity of smartphones has allowed researchers to monitor individuals' behavioural patterns unobtrusively. Earlier research in this domain has demonstrated positive results in mental health assessment using mobile sensing, typically through classical machine learning algorithms. In particular, features such as physical activity, location, sleep, and phone usage have been used to successfully assess mental health conditions such as anxiety [8], depression [2, 17], and stress [13, 17].

One of the pioneering studies is the StudentLife [17] study, which used mobile sensing to assess behavioural trends, academic performance and mental health of college students. The findings showed

that passively gathered mobile sensor data can capture relevant trends regarding students' mental health and also highlighted the influence of educational workload on well-being. Canzian and Musolesi [2] decided to solely focus on location information, finding a significant correlation between movement patterns and depression symptoms. Furthermore, a predictive model was trained to forecast significant changes in the user's depressive state based entirely on their movement. Similarly, Servia-Rodriguez et al. [14] utilised both physical and software sensors on smartphones to identify users' routines and demonstrate their correlation with personality traits and well-being.

While these studies have proven that traditional machine learning models successfully capture the relation between behavioural patterns and mental health, they depend on hand-crafted features and task-specific pipelines. Recent advancements in deep learning have introduced LLMs as powerful models that embed large amounts of both general and domain knowledge, and that do not depend on feature engineering. The integration of LLMs with sensor-derived data is gaining increasing attention in health-related research, with recent studies exploring their potential for various health predictions using wearables and smartphone sensor data [5]. Nevertheless, both mobile sensing data and mental health state represent highly sensitive data, thus, preserving privacy through on-device processing is crucial for ensuring practical applicability of LLMs in this domain. Nepal et al. [14] have developed MindScape, a mobile application that provides personalisation in journaling through context-sensitive prompts delivered by LLMs. MindScape analyses the on-device sensor data to generate personalised, context-aware journaling prompts. Despite the rising trend, current studies that use LLMs and mobile sensor data are mostly focused on other domains or general health tasks, leaving space to further explore their capabilities in mental health-focused tasks.

3 Towards LLMs for Depression Inference

3.1 Problem Definition

In this work we explore the capabilities of a mobile-ready LLM for depression inference from passively collected mobile sensor data. Informed by the related work, we examine two common flavours of depression inference:

- Depression prediction based on 14-day behavioural history;
- Prediction of changes in depression levels over a finite horizon.

Both problems are formulated as binary classification tasks, using the Patient Health Questionnaire-4 (PHQ-4) [7] to define the target labels. PHQ-4 is a validated self-report 4-item questionnaire for anxiety and depression. Each question is rated on a scale from 0 to 3, resulting in a final score between 0 and 12.

3.1.1 14-day History Depression Prediction. According to the National Institute of Mental Health (NIMH), for a person to be diagnosed with depression, they must have symptoms most of the day, every day, for at least 2 weeks [11]. Motivated by this clinical criterion and supported by its common use in related studies [2, 22], we use a 14-day behavioural history to infer depression. For each day, the inference model receives behavioural patterns representing the previous two weeks, while the ground truth label is defined by the PHQ-4 score reported on that day, as it directly reflects the user's depressive symptoms over that same 14-day period. The total PHQ-4 score can be used to categorise depression levels into four categories: (i) Normal: 0-2 (ii) Mild: 3-5 (iii) Moderate: 6-8 (iv) Severe: 9-12. Yet, to reduce the inference task to a binary classification one, we assign the label **Not Depressed** to a day in which a user's response falls within *Normal*, and assign the label **Depressed** if the response falls within *Mild*, *Moderate* or *Severe* category.

3.1.2 Prediction of changes in depression levels. Here we embrace the problem definition presented by Canzian and Musolesi in [2], where the authors investigate whether changes in an individual's depressive state can be inferred from their mobility behaviour represented as a sequence of stops and moves.

Since each user has a different baseline of depressive symptoms or tendencies, the study introduces a personalised labelling technique. Each user is assigned a personalised threshold defined with the following formula:

$$\text{Threshold} = \mu_{PHQ} + \sigma_{PHQ} \tag{1}$$

where μ_{PHQ} is the mean PHQ score of the user calculated over all recorded days and σ_{PHQ} is the standard deviation of the user's PHQ scores. A day is labelled as 1, indicating **Elevated** depressive symptoms, if its PHQ score exceeded the user's threshold, and 0 otherwise, indicating **Stable** levels. Mathematically, we define the label assignment as:

$$y_i = \begin{cases} 1, \text{ if } PHQ_i > \mu_{PHQ} + \sigma_{PHQ} \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

where PHQ_i and y_i are the PHQ score and the label for i-th day.

The classification task was then framed as predicting whether the PHQ score *time horizon (T_{hor})* days in the future would be elevated or stable using mobility features aggregated over a *time history (T_{hist})*-day window preceding the prediction day. We set T_{hist} to 14 days to align with clinical diagnostic criteria and experimented with multiple values of T_{hor} (3, 4, and 5 days).

3.2 Datasets and Data Preprocessing

We use two multi-year passive mobile sensing datasets: College Experience Study [10] and GLOBEM [22] datasets. Both contain multimodal sensing and weekly survey data. Since our goal is depression inference, we focus exclusively on the PHQ-4 responses.

To ensure comparability across datasets, we apply a consistent feature selection procedure. We prioritise behavioural indicators, such as physical activity, sleep duration and screen usage, which have been associated with depression in clinical [3, 16, 18] and mobile sensing studies focused on mental health [17]. In addition, we perform Spearman correlation analysis between sensor features and PHQ-4 scores to confirm the selected features' relevance.

3.2.1 College Experience Study Dataset. The College Experience Study [10] is the most extensive longitudinal mobile sensing study to date, encompassing sensor data from over 200 Dartmouth College students spanning across five years (2017-2022). Participants are 69.35% female, 30.65% male, and racially diverse.

Data collection was done using the StudentLife application [17]. The sensor data is of types physical activity, mobility and semantic

locations, phone usage, audio plays and sleep. Certain sensor features were exclusive to either iOS or Android. Therefore, we split the dataset into two subsets. We only use features that represent full-day summaries.

To avoid introducing artificial user behaviour through imputation techniques, we discarded days with missing values for most features. An exception was made for location data: if at least one value was present, we replaced the missing values with zeros to indicate the likely absence of activity rather than sensor error.

3.2.2 GLOBEM Dataset. GLOBEM [22] is a multi-year passive dataset, with 700 user-years and 497 unique users' mobile and wearable sensor data. It has a diverse demographic: 58.9% female, 24.2% immigrants, 38.2% first-generation students, 9.1% disabled, and various racial identities.

Data was collected annually over a three-month period. We exclude the first year due to missing PHQ-4. We use only the 14-day history segment across features like location, screen, calls, Bluetooth, steps, sleep, and WiFi.

3.3 Model Choice

Our goal is to assess the potential of state-of-the-art LLMs for automated privacy-preserving depression inference from mobile sensing data. Open AI's GPT-4o mini represents one of the highest-scoring small LLMs on the Massive Multitask Language Understanding (MMLU) benchmark, achieving an accuracy of 82% [12]. Simultaneously, with only about 8 billion parameters, this model is comparable to models, such as Meta-Llama-3.1-8B-Instruct which have been successfully deployed on commodity smartphones [6]. Therefore, in our analysis we focus on GPT-4o mini, which we access through the Open AI Gym API.

3.4 Prompts Structure

The structure and phrasing of a prompt directly influence LLM's response and the quality of the inference. In our study, we follow established guidelines and best practices in prompt design to construct effective instructions for each task [1, 4].

For the 14-day History Depression Prediction task we developed eight different prompts. The differences range from minimal changes, such as rewording, rearranging sentences or tone shifts, to more significant changes such as adding or removing information[1]. One of the prompts used is shown in Figure 1. In some prompts, we included the elements of chain-of-thought (CoT) prompting, not by requesting intermediate reasoning steps, but by providing detailed behavioural interpretations which encourage the model to engage in implicit reasoning.

For the Prediction of Changes in Depression Levels task, we define two prompts, with two flavours of each: one including the user's personalised threshold and one without it. The main goal is to explore whether providing the model with the user's threshold impacts the ability to correctly predict changes in the user's depression levels. The difference in the two base prompts is in the phrasing and instruction style – the first prompt uses a direct question format, while the other is a declarative prediction prompt.

[1]Complete prompts are excluded due to space constraints, yet full prompt texts can be found in the project repository https://gitlab.fri.uni-lj.si/lrk/llm-mobile-sensing-depression-inference

```
You are a mental health assistant. Your task is to determine whether
the user is experiencing depression based on their behavioral data
from the past two weeks.
Classification Criteria (Based on PHQ-4 Categories):
- The PHQ-4 score ranges from 0 to 12 and is divided into four categories:
    - Normal (0-2) → Not Depressed
    - Mild (3-5) → Depressed
    - Moderate (6-8) → Depressed
    - Severe (9-12) → Depressed
- A user is classified as 'Not Depressed' if their predicted PHQ-4
score falls in the Normal (0-2) category.
- A user is classified as 'Depressed' if their predicted PHQ-4
score falls in the Mild (3-5), Moderate (6-8), or Severe (9-12) categories.
Behavioral Data Information:
- Feature Units: The unit of each feature (e.g., seconds, minutes)
is written in parentheses after the feature name.
Expected Response:
- Respond only with 'Yes' if the user is depressed (PHQ-4 score: 3-12).
- Respond only with 'No' if the user is not depressed (PHQ-4 score: 0-2).
```

Figure 1: Prompt example for the 14-day History Prediction.

Having two prompts with two versions enabled us to separate the impact of threshold inclusion and prompt structure on the model's inference capabilities. These prompts can be found in Table 1.

Prompt 1	
Without threshold	You are a mental health assistant. Given the user's past 14-day behavioral metrics, will the user's PHQ score be significantly higher than their usual baseline (mean + standard deviation), {t_horizon} days from now? Respond only with 1 if yes, and 0 if no.
With threshold	You are a mental health assistant. Given the user's past 14-day behavioral metrics, will the user's PHQ score be significantly higher than their usual baseline (mean + standard deviation) equal to {threshold}, {t_horizon} days from now? Respond only with 1 if yes, and 0 if no.
Prompt 2	
Without threshold	You are a mental health assistant. Predict whether the user's PHQ score {t_horizon} days from today will be higher than their baseline, which is defined as their average PHQ score + one standard deviation. Use the provided 14-day feature data. Respond only with 1 for elevated PHQ score and 0 for normal PHQ score.
With threshold	You are a mental health assistant. Predict whether the user's PHQ score {t_horizon} days from today will be higher than their baseline of {threshold}, which is defined as their average PHQ score + one standard deviation. Use the provided 14-day feature data. Respond only with 1 for elevated PHQ score and 0 for normal PHQ score.

Table 1: Prompts used for the depression change inference.

3.5 Prompts Contextualisation

We initially experimented with zero-shot and one-shot prompting, but they yielded suboptimal results, presumably because of the complexity of the task. Consequently, we focused on few-shot

prompting, which provides the model with several examples of the behavioural feature data together with their ground-truth labels before requesting a prediction.

Our few-shot prompting setup followed a structured approach where the LLM is first given a system prompt, followed by multiple input-output pairs. The system prompt provided the model with the task explanation and other important information, such as the features unit and interpretation of the PHQ-4 questionnaire. After the main prompt, we included several examples in the form of user-assistant interactions. Each interaction consisted of a user message containing the sensor features and their values, followed by an assistant message containing the corresponding ground truth label. The examples were sampled randomly while ensuring a balanced class distribution or matching the user's original class distribution, depending on the configuration.

4 Experimental Results

4.1 14-day History Depression Prediction

For this task, the model is given a set of behavioural features representing a 14-day history period and is asked to infer whether the user is depressed or not.

4.1.1 College Experience Study Dataset. We use 500 examples per subset for the inference of this task. The Android subset has a majority class baseline (B) of 68%, while the iOS's B is 61%. As an additional comparison, we train a Random Forest (RF) classifier with a 70-30 train-test split and used grid search for hyperparameter tuning. Our RF achieves an accuracy of 76% and 59% for the Android and iOS subsets, respectively. For LLM-based inference, we performed few-shot prompting with 10 balanced examples and results are shown in Tables 2 and 3.

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	44%	33%	**68%**	44%
Prompt 2	46%	33%	64%	44%
Prompt 3	56%	34%	38%	35%
Prompt 4	57%	36%	42%	39%
Prompt 5	56%	37%	49%	42%
Prompt 6	51%	34%	57%	43%
Prompt 7	58%	39%	53%	45%
Prompt 8	**60%**	**41%**	53%	**46%**

Table 2: Few-shot prompting results for the Android subset of the College Experience Study dataset. [B 68%, RF 76% accuracy]

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	55%	44%	**56%**	**49%**
Prompt 2	54%	42%	51%	46%
Prompt 3	**60%**	**45%**	22%	30%
Prompt 4	56%	40%	27%	33%
Prompt 5	58%	**45%**	36%	40%
Prompt 6	57%	44%	43%	44%
Prompt 7	57%	44%	37%	40%
Prompt 8	58%	44%	34%	38%

Table 3: Few-shot prompting results for the iOS subset of the College Experience Study dataset. [B 61%, RF 59% accuracy]

4.1.2 GLOBEM Dataset. The GLOBEM dataset contains 2,706 examples, with a nearly balanced class distribution: 51% of the samples are labelled as not depressed, and 49% as depressed, meaning a B would yield around 51% accuracy, while RF achieved 78% accuracy with hyperparameter tuning. To evaluate the performance of the LLM, we performed few-shot prompting with 10 examples (5 of each class). The obtained results are shown in Table 4.

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	52%	51%	**62%**	**56%**
Prompt 2	53%	53%	40%	46%
Prompt 3	53%	53%	35%	42%
Prompt 4	52%	52%	24%	33%
Prompt 5	53%	53%	29%	38%
Prompt 6	53%	53%	36%	43%
Prompt 7	**54%**	53%	49%	51%
Prompt 8	53%	**54%**	34%	42%

Table 4: Few-shot prompting results for the GLOBEM dataset. [B 51%, RF 78% accuracy]

4.2 Prediction of changes in depression levels

This task explores the model's capabilities to detect upcoming changes in the user's depressive state. Once again, the model was given a set of features representing a 14-day history period and is prompted to predict whether in a given number of days(T_{hor}), the user's PHQ-4 score would significantly increase relative to the personal threshold. The approach, proposed in [2], trained personalised classifiers for each user. This personalisation aspect is preserved in our approach by using prompting techniques that include only examples from the same user currently being inferred. We performed few-shot prompting, where the given examples were drawn to mimic the class distribution in each user's data, rather than using a balanced set of examples as we did for the previous task. For the College Experience Study Dataset, we selected 10 examples, while for GLOBEM, we used only 4 because of the small amount of data available.

4.2.1 College Experience Study Dataset. Because of the personalised nature of the task, we only included participants for which substantial amount of data was collected, leading to 23 users and 6684 examples in the Android and 10 users and 2722 examples in the iOS subset. We assessed the model across 3 different T_{hor} values: 3, 4 and 5 days. Both subsets have the same B of 87% across all T_{hor} values. As a standard ML baseline, we trained separate Support Vector Machine (SVM) classifiers for each user. We provide the average accuracy of all personalised SVM models, which acts as the general baseline for this task. For the Android subset it is equal to 77%, 74% and 74% for $T_{hor} = 3$, $T_{hor} = 4$ and $T_{hor} = 5$, respectively. Similarly, the iOS subset obtained SVM accuracies of 79%, 78%, and 74% for the same values. The LLM's inference results are shown in Tables 5, 6 and 7 for Android and Tables 8, 9 and 10 for iOS.

4.2.2 GLOBEM Dataset. Due to the personalised nature of the task, we only included participants with the most data, resulting in a subset of 160 examples from 10 users. Since this dataset is significantly smaller, we added a fourth T_{hor} value: 2 days. B was

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	**68%**	**18%**	41%	**25%**
Prompt 1 with threshold	67%	**18%**	**43%**	**25%**
Prompt 2	**68%**	17%	41%	24%
Prompt 2 with threshold	64%	16%	42%	23%

Table 5: Few-shot prompting results when the $T_{hor} = 3$ for the College Experience Study Android subset. [B 87%, SVM 77% accuracy]

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	**68%**	**18%**	40%	**25%**
Prompt 1 with threshold	67%	17%	**43%**	**25%**
Prompt 2	66%	17%	41%	24%
Prompt 2 with threshold	63%	16%	**43%**	23%

Table 6: Few-shot prompting results when the $T_{hor} = 4$ for the College Experience Study Android subset. [B 87%, SVM 74% accuracy]

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	**65%**	**18%**	49%	**26%**
Prompt 1 with threshold	**65%**	**18%**	49%	**26%**
Prompt 2	64%	17%	47%	25%
Prompt 2 with threshold	60%	16%	**50%**	24%

Table 7: Few-shot prompting results when the $T_{hor} = 5$ for the College Experience Study Android subset. [B 87%, SVM 74% accuracy]

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	69%	17%	36%	**24%**
Prompt 1 with threshold	69%	18%	**37%**	**24%**
Prompt 2	**71%**	**19%**	35%	**24%**
Prompt 2 with threshold	66%	16%	36%	22%

Table 8: Few-shot prompting results when the $T_{hor} = 3$ for the College Experience Study iOS subset. [B 87%, SVM 79% accuracy]

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	68%	**17%**	35%	**23%**
Prompt 1 with threshold	69%	16%	31%	21%
Prompt 2	**70%**	**17%**	32%	22%
Prompt 2 with threshold	67%	**17%**	**37%**	**23%**

Table 9: Few-shot prompting results when the $T_{hor} = 4$ for the College Experience Study iOS subset. [B 87%, SVM 78% accuracy]

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	72%	16%	26%	20%
Prompt 1 with threshold	72%	16%	25%	19%
Prompt 2	**74%**	**18%**	27%	**22%**
Prompt 2 with threshold	71%	17%	**30%**	21%

Table 10: Few-shot prompting results when the $T_{hor} = 5$ for the College Experience Study iOS subset. [B 87%, SVM 74% accuracy]

88% for $T_{hor} = 2$, 87% for $T_{hor} = 3$, and 86% for both $T_{hor} = 4$ and 5. The SVM accuracies are 71% for $T_{hor} = 2$, 44% for $T_{hor} = 3$, 50% for $T_{hor} = 4$, and 42% for $T_{hor} = 5$. Our few-shot prompting results are shown in Tables 11, 12, 13 and 14.

5 Discussion

For both inference tasks, our results show that when used with non-textual data and few-shot prompting, LLMs consistently underperform compared to traditional machine learning algorithms. Regarding the 14-day History Depression Prediction task, none of the College Experience Study dataset's results surpass the majority class baseline. In the iOS subset, only prompt 3 achieves the accuracy that slightly exceeds that of the random forest classifier, though it still falls short of the majority class baseline. On the other hand, on the GLOBEM dataset LLM exceeds the majority class baseline across all prompts, yet, it fails to improve over the random forest. These results suggest that, unlike conventional ML methods, the LLM does not recognise meaningful patterns in sensor data. For the prediction of changes in depression levels, the LLM's performance on the College Experience Study dataset does not surpass the majority class baseline. Only the iOS subset for $T_{hor} = 5$ obtains the accuracy equal to the SVM baseline. In contrast, for the GLOBEM dataset, the LLM outperforms the SVM baselines of all T_{hor} values, although still falling behind the majority class baselines. The fact that the SVM baselines for both datasets are lower than the majority class baseline is unusual and likely comes from the small amount of training data.

While we aimed to explore the abilities of small LLMs with the goal of on-device depression inference, our findings indicate that these types of models are currently unable to match the accuracy of larger LLMs reported in [5]. The same work also reports that smaller general-purpose models of similar size(~10B) to GPT-4o mini underperform compared to bigger LLMs, unless they are domain-adapted through extensive fine-tuning, presented in the same work, which resonates with our results.

Several explanations exist for the observed underperformance of LLMs. GPT-4o mini, which is the LLM we used, is not designed for structured, non-textual time-series data. Instead, LLMs are trained on enormous text corpora and thus excel at natural language processing tasks. The difference between the training data and our inference input data likely contributed to the inability to capture important behavioural patterns. This is supported by recent research (e.g. [15]), which claims that while LLMs excel at certain time-series tasks such as anomaly detection, they frequently fail to outperform simpler models when it comes to prediction tasks.

Tables 2 and 3 show that prompt phrasing can significantly impact model accuracy, with up to 16% and 6% differences, respectively, thus highlighting the importance of prompt engineering. Prompts 3, 7, and 8 performed best for the 14-day History Prediction task, with prompts 7 and 8 incorporating behavioural feature thresholds derived from data analysis, acting as interpretable anchors resembling a form of Chain-of-Thought (CoT) reasoning. For the College Experience Study's depression change prediction, prompt variants without personalised thresholds generally outperformed those with thresholds, except in one case. These findings emphasize

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	77%	16%	19%	17%
Prompt 1 with threshold	**78%**	**17%**	19%	18%
Prompt 2	70%	16%	**31%**	**21%**
Prompt 2 with threshold	67%	14%	**31%**	19%

Table 11: Few-shot prompting results when the $T_{hor} = 2$ for the GLOBEM dataset.

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	75%	16%	20%	18%
Prompt 1 with threshold	**79%**	**24%**	27%	**25%**
Prompt 2	68%	8%	13%	10%
Prompt 2 with threshold	70%	17%	**33%**	23%

Table 12: Few-shot prompting results when the $T_{hor} = 3$ for the GLOBEM dataset.

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	74%	12%	15%	14%
Prompt 1 with threshold	**75%**	8%	8%	8%
Prompt 2	74%	**17%**	23%	19%
Prompt 2 with threshold	66%	**17%**	**38%**	**23%**

Table 13: Few-shot prompting results when the $T_{hor} = 4$ for the GLOBEM dataset.

Prompt	Accuracy	Precision	Recall	F1-score
Prompt 1	78%	11%	9%	10%
Prompt 1 with threshold	**84%**	**33%**	18%	24%
Prompt 2	70%	12%	18%	14%
Prompt 2 with threshold	72%	28%	**64%**	**39%**

Table 14: Few-shot prompting results when the $T_{hor} = 5$ for the GLOBEM dataset.

the sensitivity of the prompt design. Additionally, the reliability of predictions is influenced by the quality of ground truth labels, which in this study are based on the PHQ-4 – a screening, not a diagnostic, tool. As PHQ-4 relies on self-reported symptoms and does not confirm clinical diagnoses, it introduces label noise that may limit LLM performance.

6 Conclusion

Automated privacy-preserving inference of mental health state directly on a user's smartphone would be highly valuable in the light of rising cost and reduced availability of conventional screening methods. While LLMs have made tremendous advances on numerous data processing tasks, our examination of the performance of a state-of-the-art smaller LLM shows that depression inference from mobile sensing data remains out of the contemporary small LLMs' reach. We identified multiple factors that may have contributed to the model's poor performance, including the type of input data, prompt sensitivity, and data quality, and set guidelines for future efforts in the field of penetrative AI.

References

[1] [n. d.]. Prompt Engineering Guide. https://www.promptingguide.ai/. Accessed: 2025-06-12.

[2] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *ACM UbiComp* (Osaka, Japan).

[3] Alexa Deyo, Josh Wallace, and Katherine M. Kidwell and. 2024. Screen time and mental health in college students: Time in nature as a protective factor. *Journal of American College Health* 72, 8 (2024), 3025–3032.

[4] P. A. Hill, L. K. Narine, and A. L. Miller. 2024. Prompt Engineering Principles for Generative AI Use in Extension. *The Journal of Extension* 62, 3 (2024), Article 20.

[5] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. arXiv:2401.06866 [cs.CL]

[6] Martin Korelič, Octavian Machidon, and Veljko Pejović. 2025. SELLMA: Semantic Location through On-Device LLMs and WiFi Sensing. In *ACM EdgeSys Workshop*. Rotterdam, Netherlands.

[7] Kurt Kroenke, Robert L. Spitzer, Janet B.W. Williams, and Bernd Löwe. 2009. An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQ–4. *Psychosomatics* 50, 6 (2009), 613–621.

[8] Dante L Mack, Alex W DaSilva, Courtney Rogers, Elin Hedlund, Eilis I Murphy, Vlado Vojdanovski, Jane Plomp, Weichen Wang, Subigya K Nepal, Paul E Holtzheimer, Dylan D Wagner, Nicholas C Jacobson, Meghan L Meyer, Andrew T Campbell, and Jeremy F Huckins. 2021. Mental health and behavior of college students during the COVID-19 pandemic: Longitudinal mobile smartphone and ecological momentary assessment study, part II. *J. Med. Internet Res.* 23, 6 (June 2021), e28892.

[9] Geoffrey Miller. 2012. The smartphone psychology manifesto. *Perspectives on psychological science* 7, 3 (2012), 221–237.

[10] Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. 2024. Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 38 (March 2024).

[11] NIMH [n. d.]. *National Institute of Mental Health - Depression.* https://www.nimh.nih.gov/health/publications/depression

[12] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[13] Akane Sano and Rosalind W. Picard. 2013. Stress Recognition Using Wearable Sensors and Mobile Phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 671–676.

[14] Sandra Servia-Rodríguez, Kiran K. Rachuri, Cecilia Mascolo, Peter J. Rentfrow, Neal Lathia, and Gillian M. Sandstrom. 2017. Mobile Sensing at the Service of Mental Well-being: a Large-scale Longitudinal Study. In *International Conference on World Wide Web (WWW)* (Perth, Australia).

[15] Francis Tang and Ying Ding. 2024. Are Large Language Models Useful for Time Series Data Analysis? arXiv:2412.12219 [cs.LG] https://arxiv.org/abs/2412.12219

[16] Norifumi Tsuno, Alain Besset, and Karen Ritchie. 2005. Sleep and Depression. *The Journal of Clinical Psychiatry* 66, 10 (2005), 19685.

[17] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *ACM UbiComp* (Seattle, Washington).

[18] Xiaoyan Wu, Shuman Tao, Yukun Zhang, Shichen Zhang, and Fangbiao Tao. 2015. Low Physical Activity and High Screen Time Can Increase the Risks of Mental Health Problems and Poor Sleep Quality among Chinese College Students. *PLOS ONE* 10, 3 (03 2015), 1–10. doi:10.1371/journal.pone.0119607

[19] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 1–7.

[20] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088* (2024).

[21] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 31 (March 2024), 32 pages.

[22] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve Riskin, Jennifer Mankoff, and Anind K. Dey. 2023. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. arXiv:2211.02733 [cs.LG]