From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models

ZACHARY ENGLHARDT* and CHENGQIAN MA*, University of Washington, USA
MARGARET E. MORRIS and CHUN-CHENG CHANG, University of Washington, USA
XUHAI "ORSON" XU, Massachusetts Institute of Technology, USA
LIANHUI QIN, University of California, San Diego, USA
DANIEL MCDUFF and XIN LIU, University of Washington, USA
SHWETAK PATEL and VIKRAM IYER, University of Washington, USA

Passively collected behavioral health data from ubiquitous sensors could provide mental health professionals valuable insights into patient's daily lives, but such efforts are impeded by disparate metrics, lack of interoperability, and unclear correlations between the measured signals and an individual's mental health. To address these challenges, we pioneer the exploration of large language models (LLMs) to synthesize clinically relevant insights from multi-sensor data. We develop chain-of-thought prompting methods to generate LLM reasoning on how data pertaining to activity, sleep and social interaction relate to conditions such as depression and anxiety. We then prompt the LLM to perform binary classification, achieving accuracies of 61.1%, exceeding the state of the art. We find models like GPT-4 correctly reference numerical data 75% of the time.

While we began our investigation by developing methods to use LLMs to output binary classifications for conditions like depression, we find instead that their greatest potential value to clinicians lies not in diagnostic classification, but rather in rigorous analysis of diverse self-tracking data to generate natural language summaries that synthesize multiple data streams and identify potential concerns. Clinicians envisioned using these insights in a variety of ways, principally for fostering collaborative investigation with patients to strengthen the therapeutic alliance and guide treatment. We describe this collaborative engagement, additional envisioned uses, and associated concerns that must be addressed before adoption in real-world contexts.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Applied computing → Life and medical sciences**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: Passive sensing, large-language-models, clinical insights, mental health

ACM Reference Format:
Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 56 (June 2024), 25 pages. https://doi.org/10.1145/3659604

*Both authors contributed equally to this research.

Authors' Contact Information: Zachary Englhardt, zacharye@cs.washington.edu; Chengqian Ma, University of Washington, Seattle, Washington, USA; Margaret E. Morris; Chun-Cheng Chang, University of Washington, Seattle, Washington, USA; Xuhai "Orson" Xu, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; Lianhui Qin, University of California, San Diego, La Jolla, California, USA; Daniel McDuff; Xin Liu, University of Washington, Seattle, Washington, USA; Shwetak Patel; Vikram Iyer, vsiyer@uw.edu, University of Washington, Seattle, Washington, USA.

ACM 2474-9567/2024/6-ART56
https://doi.org/10.1145/3659604

1 INTRODUCTION

Mobile and wearable sensors that collect health and fitness data have seen explosive growth over the past five years [23, 27]. The sensing capabilities of products such as Fitbit and Apple Watch have dramatically advanced beyond simple step counts to include optical heart rate sensors and two lead electrocardiogram (ECG) measurements that can provide clinicians valuable information about a patient's symptoms outside their practice. Beyond simply tracking a run, or checking for high heart rate, researchers have shown the potential of leveraging such passive sensing data to model high-level, complex behaviors and mental health. [55, 64, 67, 74].

Despite their potential, mobile and wearable sensor data use in clinical mental health practice faces four key challenges. First, there is a lack of trust and clinically validated data; for instance, there is high uncertainty about the specific relationships between mobile health signals (such as activity levels) and mental states like depression [53]. Second, clinicians struggle with the additional burden of incorporating mobile health data during treatment and often have difficulty interpreting non-standard signals generated by wearable devices and mobile phones [28, 62]. Third, conventional machine learning (ML) approaches for interpreting sensor data perform poorly on abstract relations. For example, recent work evaluating the ability of 19 different ML models to predict depression using self-tracking data revealed that many achieved accuracies below 50% [74]. Finally, the qualitative aspects of mental health further limit the utility of binary data classifications since diagnoses on their own contain no context about the patient's lifestyle and history, which are critical to designing treatment plans [14].

In this context, we contribute the following key insights and findings:

(1) *We perform the first exploration of LLM use to process multi-sensor ubiquitous wearable data.* We develop a series of new prompting and model fine tuning strategies that enable LLMs to perform zero-shot depression classification using raw, multi-modal wearable sensor data. We explore the effects of data input formats and numerical accuracy on state-of-the-art LLMs like GPT-4 [47], PaLM 2 [16] and GPT-3.5 [49]. We demonstrate that these strategies outperform classical ML methods on state-of-the-art depression classification and highlight performance variations across models.

(2) As an intermediate step to classification, *we observe that LLMs can generate text reasoning about multi-sensor data,* correctly describing trends and anomalies in the data and making connections between multiple input signals and relevant mental health scenarios.

(3) To evaluate the feasibility of this approach, *we undertake an interactive interview study with mental health professionals to critically assess the practicality and limitations of interpreting mobile health data using LLMs,* gain insight into the clinical relevance of LLM-generated reasoning, and to consider the prospective role of LLMs in mental health contexts. We find that clinicians express a strong desire to have access to LLM-based tools for collaborative investigation of mobile health data with patients, and *we outline a series of scenarios for clinician-patient-AI collaborative therapy.*

2 RELATED-WORK

2.1 Multi-sensor Passive Sensing for Health and Well-being

Smartphones and wearable devices, now ubiquitous in our lives, function as passive sensors, seamlessly capturing a vast range of data. Their near-constant presence enables unobtrusive and continuous monitoring of behavior, activity, and physiological signals. Over the last decade, significant progress has been made in passive sensing and behavioral modeling, impacting areas such as physiological health condition detection [5, 42, 75], monitoring mental health status [65, 72], measuring job performance [41, 43], tracking education outcomes [66, 80], and tracing social justice [56]. Researchers employ various methods, including statistical analysis and conventional ML models, to explore these areas.

In the mental health context, initial research established statistical correlations between mental health conditions and mobile sensing data. For instance, Saeb et al. [54] identified significant correlations between depression scores and smartphone usage patterns, and Ben-Zeev et al. [7] identified links between changes in depression severity levels and features related to sleep duration, speech duration, and mobility. More recent efforts have focused on leveraging these results to build ML models for mental health disorder diagnosis and detection [15, 44, 63, 67]. To further growth in this area, Xu et al. [74] collected and released a multi-year passive sensing dataset and platform that covers a wide range of physical health, mental health, and social well-being measurements. However, most research in this domain relies on conventional statistical and ML methods, and the recent improvements in the performance of large foundation models [11] present an opportunity to explore new techniques for analyzing passively collected sensor data.

2.2 LLMs for Health Applications

The success of transformer-based language models, such as BERT [19] and GPT [50], has led to the development of larger and more powerful language models (e.g., GPT-3 [10] and T5 [51]). Instruction fine-tuning by including instructions (i.e., prompts) from a range of datasets and task domains during both the training and generation phases has led to the development of single models that can perform a wide range of tasks [69]. These instruction-fine-tuned LLMs, such as GPT-4 [47], PaLM [17], FLAN-T5 [18], LLaMA [61], and Alpaca [60], contain tens to hundreds of billions of parameters and achieve a promising level of performance on a variety of tasks, such as question answering [46, 52], logic reasoning [70, 82], machine translation [8, 22], and more.

In the health sector, these LLMs have been applied in several studies [30, 36, 38, 45, 59, 71]. For example, Singhal et al. [59] utilized a fine-tuned version of PaLM-2 to score as high as 86.5% on the MedQA dataset. Similarly, Wu et al. [71] fine-tuned LLaMA on a corpus of academic medical papers and textbooks, yielding promising results on multiple biomedical QA datasets. Jiang et al. [30] trained a medical language model on unstructured clinical notes from the electronic health record and fine-tuned for performance across a wide range of clinical and operational predictive tasks. These examples underscore the versatility and potential effectiveness of LLMs in the medical space.

In the mental health domain, LLMs have been explored for applications such as sentiment analysis and emotional reasoning [32, 49, 81]. Lamichhane [34], Amin et al. [2]. Yang et al. [78] tested the performance of ChatGPT on multiple classification tasks (stress, depression, and suicide risk) and found that it shows initial potential for these mental health applications, but it has room for significant improvement.

Despite this past work, scant research focuses specifically on *integration with mobile and wearable health data*, with most of the existing literature exploring text data rather than multi-sensor streams. Closer to our work, Liu et al. [39] demonstrated that with only few-shot tuning, a LLM can ground various physiological and behavioral time-series data and make meaningful inferences on numerous health tasks (e.g., heart rate measurement, atrial fibrillation detection, and mood score prediction). However, their work is based on self-curated toy datasets consisting of well-described physiological signals and behaviors.

3 MATERIALS AND METHODS

In this section we outline the procedures used in our experiments. We first explore the use of LLMs for depression classification. In these experiments, we notice that LLMs produce reasoning about the input mobile health data before providing a classification response. We grade this reasoning for numerical accuracy. Next, we conduct an interactive evaluation with mental health professionals to gauge the quality of LLM-generated reasoning as it relates to mental health and gain insights into the potential applications for LLM-based tools in therapeutic contexts.

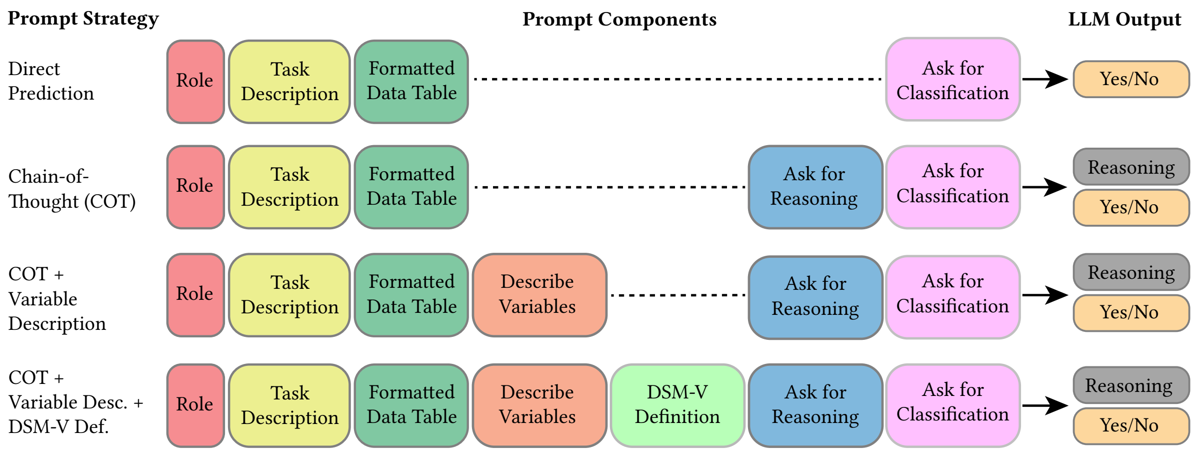


Fig. 1. The overall workflow of our prompting strategies, detailing the specific components included in each type of prompt. For examples of the text included in each block, see Appendix B.

3.1 Dataset

In this paper, we use the GLOBEM dataset [76], with its extensive collection of passive sensing data from smartphones and wearables and its wide range of well-being metrics. GLOBEM includes weekly Ecological Momentary Assessment (EMA) surveys to capture the recent status of participants' mental health evaluations. Within EMA measures, we use Patient Health Questionnaire 4 (PHQ-4) [33] as the ground truth in our depression classification task [25, 26, 29]. Paired with the survey data, GLOBEM contains 24×7 sensing data that measures user behavior, such as steps, GPS locations, phone calls, social activity proxies, and more. The dataset then extracts hundreds of features (time at home, time asleep, etc.) from these raw measurements.

3.1.1 Model Input Setup. To constrain the input token length, we select a subset of 16 diverse features, including Location, Phone Usage, Bluetooth, Calls, Physical Activity, and Sleep; we list all feature details in Appendix A. We set the time length of the data as 28 days, with the final day of each window coinciding with the weekly EMA assessment. Therefore, the size of mobile and wearable sensor data for each data sample is 28×16 (28 days × 16 features). Due to the data protection guidelines for the GLOBEM dataset, we are unable to directly open-source raw excerpts of sensor data or fine-tuned LLMs from which excerpts of training data may extracted[13]. To aid others in reproducing our results, we provide instructions and code to generate samples prepared from the GLOBEM dataset used in the following experiments in a GitHub repository[1].

3.1.2 Classification Label Setup. We focus on a binary classification task based on PHQ-4 scores, which attempt to quantify the severity of depression and anxiety from a range of 0 (normal) to 12 (severe) based on responses to a questionnaire. As an initial exploration, we avoid the borderline samples with PHQ-4 scores between 1-5, following PHQ-4 criteria [33] and supported by additional work finding frequent disagreement between self-reported and clinician-administered scores for borderline cases [3]. Thus, the classification task aims to distinguish samples with PHQ-4 below 1 or above 5. To create a balanced test set, we randomly sampled 30 data points from each year with an equal distribution of labels to equally represent a range of non-borderline cases. In total, the test data set contains 90 class-balanced samples from three years.

[1]https://github.com/ubicomplab/classification-to-clinical

3.2 Classification with LLMs

3.2.1 Depression Classification. We develop prompting strategies to enter the raw data into LLMs along with varying amounts of context and instructions to produce a depression classification result. For each strategy, we include prompt text for the model's role and task description, followed by the concatenated raw sensor data in varying formats, and we proceed to add additional context like variable descriptions and instructions for the task.For example, a prompt would begin with:
Role: You are a data analyst helping a psychiatrist understand human activity data.
Task: You will be shown data gathered from a smartphone and smart watch worn by an individual. Your goal is to analyze this data. You are presented with the following: 1. A table consisting of twenty-eight days of collected activity tracking data <formatted data> ...

We show a diagram of each of the utilized prompting formats in Figure 1 and a detailed description of the text formatting of each block in Appendix B. We focus on three different variants of prompting, all on a zero-shot setting detailed below:

(1) *Direct Prediction (DP):* we directly ask the LLM to perform depression classification with prompts that include only the basic role and task information and the formatted sensor data.
(2) *Chain-of-Thought (CoT) prompting:* building upon direct prediction, we induce models to perform step-by-step reasoning with carefully crafted instructions to hypothesize about the subject's overall mental health.
(3) *Reasoning with extra information:* based on CoT, we provide extra task-related domain information, such as more detailed explanations of input variables (Exp) and the depression criteria from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V)[1].

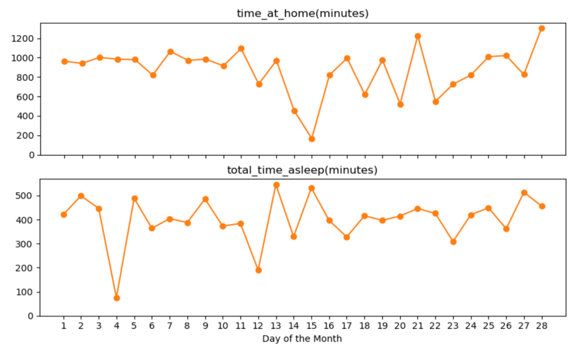
We test these prompting variants on top of three state-of-the-art LLM models: GPT-3.5, GPT-4 and PaLM 2. We evaluate the DP method with multiple data input formats to determine which is most suitable for the selected LLMs. Specifically, we tried four ways to format these raw data: comma-separated values (CSV), Tabular, Markdown and LaTeX, as shown in Appendix D. We found that Markdown performed best, so we utilize the Markdown table format in all other experiments.

Using these strategies, we observe that prior to outputting a classification decision, the models first produce intermediate text with reasoning and analysis of the input data. Fig 2 shows an example, with a subset of the raw time series input to the LLM plotted for visualization along with a subset of the analysis text produced by GPT-4. These outputs form the basis for our clinician study and envisioned use cases in which these generative LLM outputs can enable new capabilities that go beyond classification to aid clinicians in collaborative therapy. We discuss these extensively in Sections §3.4 and §4.6.

To contextualize the classification performance of LLMs, we also introduce a few baselines, including a classic ML method, Random Forest (RF) [9], as well as the state-of-the-art self-supervised learning method *Reorder*, proposed by [74] on this same dataset. Given that these comparison baseline ML models do not support a zero-shot setting and require pre-training, we curated a training dataset for comparative analysis. This dataset comprises the remaining data points that meet the PHQ-4 thresholds, which excludes the data points in the testing dataset, totaling 384 samples. These were formatted into two structures: the first is identical to the test set samples with a structure of 28 days $\times$ 16 features. The second is a set of 16 features computed by taking the average along the time dimension for the Random Forest. We conducted evaluations using Random Forest (RF) and Reorder algorithms, alongside the Reorder algorithm with the original implementation in [73] with 54 features (Reorder-54 Features). These evaluations were performed on the same test set as the LLMs.

3.2.2 Fine-Tuning. Evaluating the CoT reasoning produced by GPT-3.5 reveals that unlike GPT-4, smaller LLMs rarely incorporate analysis that relates specifically to the numerical values of the mobile health data, instead

Raw Data:



GPT-4 Reasoning:

1. Time Spent at Home:
- The time spent at home seems to be decreasing around the middle of the month, notably on 2019-05-14 and 2019-05-15 where it dips to 456 and 167 minutes respectively. This could indicate some unusual activity or a significant change in routine.

2. Sleep Patterns:
- The total time asleep varies across the month. On 2019-05-04, the individual slept for only 75 minutes. On the other hand, on 2019-05-13, they slept for 514 minutes. Erratic sleep patterns can be an indicator of various issues, including stress, anxiety, or other disorders.
- The total time spent awake while in bed also sees fluctuations, with some nights having prolonged awake times, suggesting possible insomnia or restless sleep.

Fig. 2. A plotted excerpt of raw mobile health data and the resulting analysis generated by GPT-4.

generating generic statements about mental health. Several studies[24, 58, 77] have shown that the step-by-step CoT reasoning capabilities of larger models can be distilled into smaller models through fine-tuning on the output of larger models. This raises an important question: can we use instruction fine-tuning to enable smaller LLMs to perform more effectively on such challenging wearable classification tasks?

To answer this question, we explore using the reasoning responses generated from the GPT-4 (see Appendix C) to fine-tune the GPT-3.5 model. For the fine-tuning experiment, the prompt design was kept consistent with the same methodology outlined above. Utilizing GPT-4, we generate a collection of reasoning responses based on data external to the test set. From this assortment, correctly classified reasoning responses free of numerical errors were selected to form a candidate training set. We make a balanced instruction training set, comprising 70 sample with high-quality reasoning, evenly distributed between positive and negative examples, and fine-tune GPT-3.5 on this dataset.

3.2.3 Anxiety Classification. In addition to depression, we also explore the generalization abilities of LLMs to other mental health classification tasks. Specifically, we use the same CoT prompt design as in the depression classification experiment described in Section 3.2.1, with the modification of asking for anxiety classification based on the PHQ-4 anxiety sub-score instead of the depression classification.

3.3 Evaluating Numerical Accuracy

We begin by investigating the accuracy of the LLMs when referencing the input data. Can the models identify real and specific trends in the data? Do they hallucinate numbers, or are the outputs truly reflective of the input data? To evaluate this, we utilize human graders to objectively evaluate the numbers and stated trends referenced in LLM responses against the input timeseries data according to a fixed rubric. We score a total of 480 responses, evenly split across four different models: PaLM 2, GPT-3.5, fine-tuned GPT-3.5, and GPT-4.

3.3.1 Producing Reasoning Samples. Instead of asking to hypothesize about the health of the patient, we tune this prompt to produce analysis on trends in the data. We show the specific prompt used in Appendix E. For input data, we select random samples from the test set with a PHQ-4 score of greater than 5, indicating likely moderate to severe depression. We select these samples since they are most likely to have a trends or anomalies on which we can evaluate the model. These excerpts consist of the same 16 features used in Section 3.2 formatted in markdown format. We produce 8 samples from each of the four models per set of input data.

3.3.2 Participants. To grade reasoning excerpts, we recruited 15 individuals over the age of 18 through fliers placed around a university campus and word of mouth. Participants were not required to have any domain knowledge of mental health to complete the assigned task. To aid in recruitment, compensation of a gift card ($20 USD) was offered to all participants included in the study. The study protocol was submitted to the IRB at the host institution for the study and deemed exempt from a full review. We follow all IRB procedures to avoid potential conflict of interest.

3.3.3 Prompt Grading Procedure. Each participant was sent an online form that contained a set of 32 randomly ordered responses (8 from each LLM). To reduce grader burden, all 32 responses were generated from the same input data. Graders were also provided the raw data input to the model in tabular form as well as timeseries plots of each data feature. For each reasoning excerpt, graders were asked to answer the following four questions:

(1) Does this response include numbers?
(2) Are these numbers consistent with the provided data?
(3) Does this response identify specific trends?
(4) Are these trends consistent with the provided data?

Graders were given specific instructions on how to evaluate each question as well as a series of graded example responses; we show the rubric provided to graders in Appendix F. We provide explicit instructions to evaluate solely the numbers and trends against the provided data table and plots, disregarding any conclusions the responses may make about how these numbers or trends might relate to mental health or other factors. An excerpt of plotted data and LLM-generated reasoning graded by participants is shown in Fig. 2.

3.4 Clinician Evaluation

As shown in Fig 2, the LLMs also generate analyses about the timeseries data. To evaluate the quality of this reasoning and understand how these clinical insights could be used in practice, we conducted a user study with clinician experts.

3.4.1 Participants. Eight mental health professionals completed this study: six with PhDs in Clinical Psychology and two with master's degrees. Participants described their approaches as Cognitive-Behavioral Therapy (2), Acceptance and Commitment Therapy (1), Dialectical Behavioral Therapy (1), Psychodynamic Therapy (1), Relational/Interpersonal Therapy (2), and Family Systems Therapy (1). They reported working in a variety of settings, including academic medicine, group private practice, individual private practice, and community mental health, across four states in the U.S. Participants were recruited through postings on group practice mailing lists, social media groups for practicing therapists, and word of mouth. To aid in recruitment, gift cards ($50 USD) were offered to all who participated in the study.

3.4.2 Ethics. This study protocol was submitted to the IRB at the host institution and was deemed exempt from formal IRB review. Participants were sent an information sheet outlining the study process and data management procedures prior to interviews. At the start of the interview but before initiating recording, researchers described the interview process, addressed questions, and obtained spoken consent.

3.4.3 Data Format. The data analyzed in each interview session consisted of a random 28-day sample from a participant in the GLOBEM dataset [76]. Only individuals with a PHQ-4 > 5 were selected (as in Section 3.2) to obtain a sample representative of individuals who might be likely to seek therapy. Time-series plots that included each variable were generated. The generative reasoning evaluation drew on the same 16 features and format used for the time series (listed in Appendix A). The chat thread was started with the following prompt, similar to

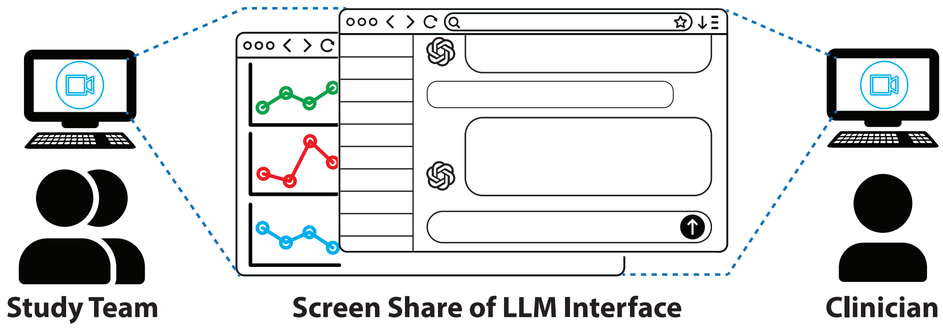


Fig. 3. Setup for interactive clinician evaluation. Clinicians interacted with live GPT-4 sessions via screen sharing over Zoom.

the prompt used in the previous section, but altered slightly for compatibility with the web-based chat interface of GPT-4: *Below is some data gathered from a fitness tracking smartwatch and a smartphone. Although it does not contain explicit information on mood, trends in physiological signals have been shown to correlate with mental health symptoms. Examine this data and point out any specific trends or data points that could spark fruitful conversation with a mental health professional. <formatted-data>.*

3.4.4 Interview and Interactive Exercise. Interviews were one hour in length and conducted over Zoom to enable recording and transcript generation. Participants were not shown the results of other experiments conducted in this paper. Interviews were conducted by two researchers, a clinical psychologist and a computer science graduate student.

Interviews began with a discussion of the clinician's practice, particularly as it includes patient self-monitoring. We then asked participants to imagine a scenario in which they had received a month of self-tracking data in advance of their first meeting with a patient. To build on this scenario, we shared examples of self-tracking data from the same publicly available data set used above [74] in two forms. First, we presented a set of 16 time series plots (in a shared Google doc), and clinicians were asked to provide feedback on anything they noticed in the data that they might use in a therapeutic context.

Next, we began a live interactive session with GPT-4 (using ChatGPT) through Zoom screen sharing, as shown in Figure 3. The mobile health data was then input into GPT-4 to produce a text response in real-time. We asked participants to read the response and talk about any reactions they had, including ways the data and GPT-4's observations about the data might shape their thinking about the patient. Participants were asked to type or say aloud follow-up queries that they had for GPT-4 (which were entered by the interviewer running the GPT-4 session). We also discussed their reaction to the responses that GPT-4 gave to these queries. We then asked participants to make up a hypothetical example in which they used this tool with a therapy patient, prompting them for what types of data or inputs should be available, what kinds of analyses they would like to see, and how they would envision using this tool with the patient. Finally, we asked general questions about treatment and how this tool might affect relationships with patients.

3.4.5 Post-Interview Survey. After the interview, participants completed a short online survey where they indicated their agreement or disagreement with several statements about GPT-4 on a seven-point scale.

4 RESULTS

In this section, we provide results of the depression classification, fine-tuning, and anxiety classification experiments as well as the numerical accuracy and clinician evaluations, all described in Section 3.

4.1 Depression Classification

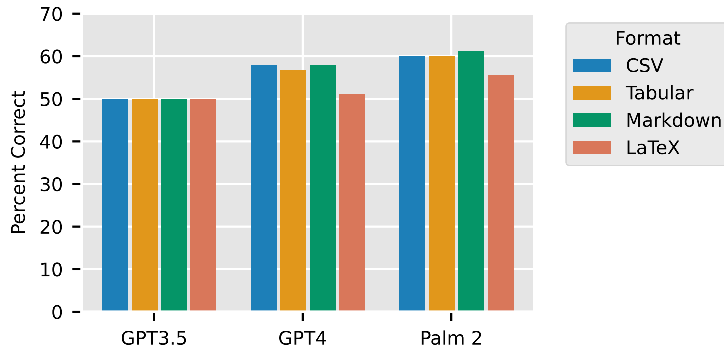


Fig. 4. Depression classification performance across a range of possible data input formats.

To evaluate multiple data input formats, we use the classification results of depression through Direct Prediction (DP) as a measure of these formats. From Figure 4, we observe that CSV, Tabular and Markdown formats exhibit comparable performance levels, and these results are consistent in both GPT-4 and PaLM 2. In contrast, the LaTeX format demonstrates a performance gap compared to the other three formats. These results align logically with expectations, considering the predominant sources of training data for LLMs. The vast majority of this data is sourced from the Internet, where formats like CSV and Markdown are far more prevalent than LaTeX. Given this disparity in data availability, it stands to reason that LLMs would exhibit higher accuracy in processing and interpreting CSV and Markdown inputs compared to LaTeX. Since Markdown shows the best overall performance, we chose Markdown as the data format for our subsequent experiments.

We tested all prompting strategies with three state-of-the-art LLM models, GPT-3.5, GPT-4 and PaLM 2 in Markdown format. Figure 5 reveals that while GPT-3.5 attains an accuracy rate of 50%, it does not effectively address the question posed. Instead, it consistently defaults to a response of 'No', which results in an inflated 50% accuracy due to our balanced dataset. Figure 5 shows that CoT improves the accuracy of both GPT-4 and PaLM 2 compared to DP, which aligns with the results found in many related studies [68, 70]. Using the CoT + Exp. strategy, PaLM 2 achieved the highest accuracy of 61.11%. However, adding information, even if accurate and pertinent to the topic, does not always increase performance. Both PaLM 2 and GPT-4 perform their worst, at 48.89% and 51.11%, respectively, when provided with the DSM-V description of depression as part of the prompt. In these cases, we see that this results in a significant increase in the percentage of samples classified as depressed, with GPT-4 classifying as high as 98.89% of samples as positive.

The performance metrics reveal that the accuracies of Random Forest, Reorder, and Reorder-54 Features are 53.33%, 51.11%, and 58.89%, respectively. Notably, among the two baseline models utilizing the same set of 16 features provided to the LLMs, neither could surpass the zero-shot Chain of Thought (CoT) results achieved using the GPT-4 and PaLM 2 models.

4.2 Fine-Tuning

After fine-tuning GPT-3.5 with 2 epochs using this balanced instruction-tuning set, we see an improvement in its performance. Though it initially fails to properly perform classification, our fine-tuned version of GPT-3.5 achieves an accuracy of 56.67% on the test set, which exceeds the baseline Random Forest model performance of 53.33% and is closer to the GPT-4 performance of 57.78% on the same dataset and prompt structure.

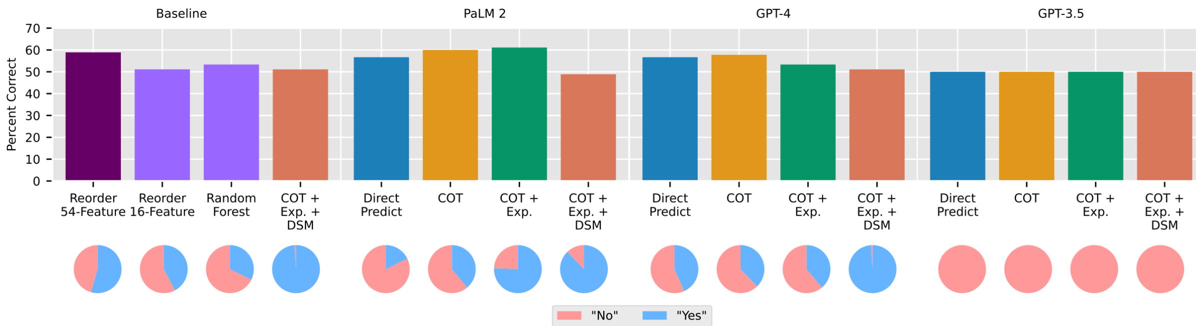


Fig. 5. Comparison of classification results of GPT-3.5, GPT-4, and PaLM 2 across four different prompting strategies, along with results from Reorder and Random Forest models trained on the same dataset. The performance of the Reorder model trained on more features in Xu et al. [74] is included for comparison. Observe how for LLMs, the percent of positive ("Yes") and negative ("No") classifications varies significantly based on the prompting strategy used.

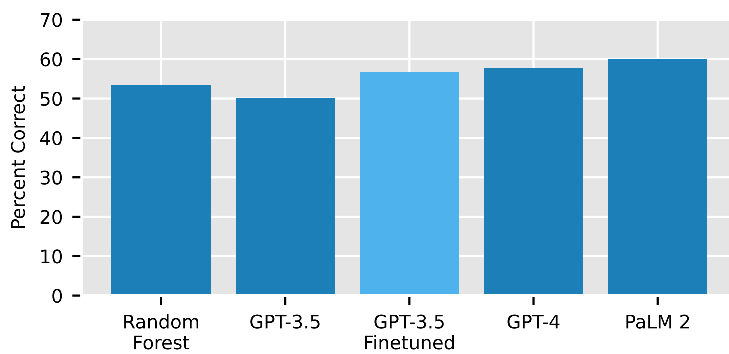


Fig. 6. Comparison of depression classification performance for GPT-3.5 fine-tuned for chain-of-thought (CoT) reasoning depression classification against CoT performance of GPT-3.5, GPT-4, and PaLM 2. The performance of the Random Forest model is included as a baseline.

4.3 Anxiety Classification

As in the depression classification tasks, we observe that GPT-3.5 still always defaults to a "No" response. GPT-4 can achieve an accuracy of 55.56%, while PaLM 2 achives a slightly higher accuracy of 56.67%. Notably, this result mirrors the trend observed in the depression classification results. For our fine-tuned GPT-3.5 model, although the model is fine-tuned for depression classification, it still shows some improvement in the anxiety detection task compared to the original GPT-3.5. This indicates the potential of fine-tuning to increase performance on classification tasks.

4.4 Evaluating Numerical Accuracy

Figure 8 shows the results of our study. Interestingly, we find that while PaLM 2 performs slightly higher on the classification tasks above, GPT-4 performs significantly better across our study rubrics. For example, while PaLM 2 also identifies trends at a high rate, GPT-4 is more likely to identify all trends correctly and include references to the numerical data, achieving scores exceeding 75% accuracy. We further note that the graders evaluated data as simple yes/no questions, meaning that all numbers and trends had to be correct. Although we

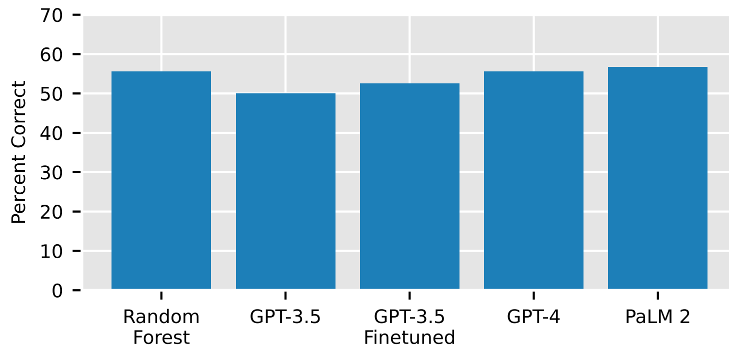


Fig. 7. Comparison of anxiety classification performance for GPT-3.5 fine-tuned for chain-of-thought (CoT) reasoning depression classification against CoT performance of GPT-3.5, GPT-4, and PaLM 2. The performance of the Random Forest model is included as a baseline.

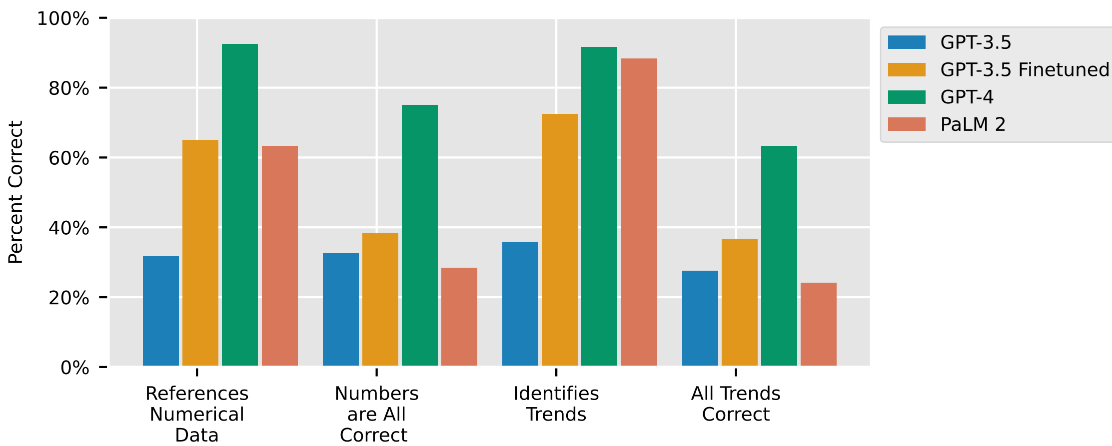


Fig. 8. Evaluation of the reasoning excerpts generated by GPT-3.5, Fine-tuned GPT-3.5, GPT-4, and PaLM 2.

observe some errors in the model outputs, even these responses often contain correct trends as well that may have utility in a collaborative human-AI approach. For example, if the LLM can observe an outlier or anomaly in the data, this is often easily visible to the user as well for confirmation.

4.5 Clinician Evaluation

We discuss findings from clinical interviews in two categories. First, we describe envisioned usages, that is how clinicians thought about incorporating the tool into their practice. Second, we describe concerns that came up as clinicians envisioned using the tool. We preface these findings by noting a general interest among clinicians in having access to LLM-based tools in their practice (see the post-study survey responses in Fig. 9.

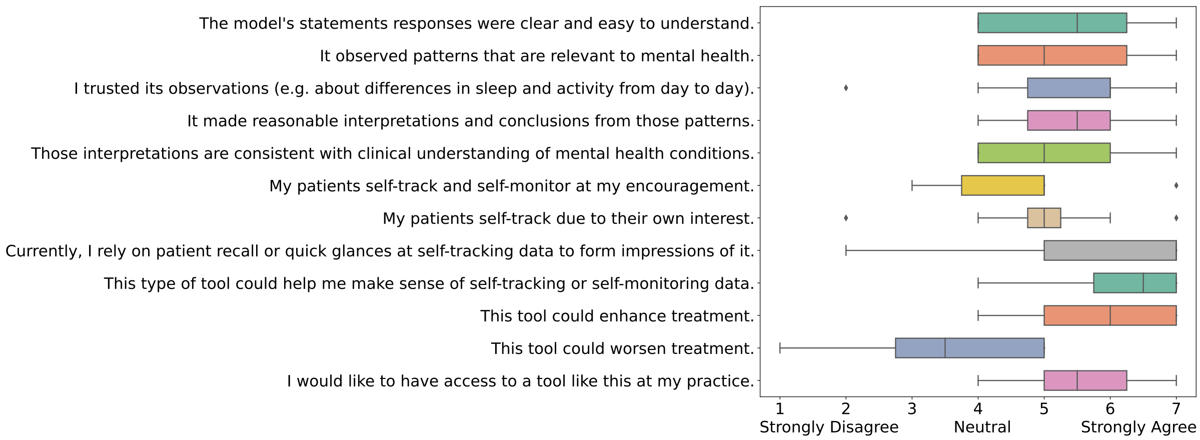


Fig. 9. Post-study survey responses from clinicians, indicating strong disagreement (1) to strong agreement (7) with statements about their experience interacting with GPT-4. We observe positive feedback and enthusiasm across most questions.

4.6 Envisioned Clinical Uses

Three major uses were envisioned by clinicians: collaborative investigation, identifying questions to explore, and documentation.

4.6.1 Collaborative In-Session Investigation. Clinicians generally saw the most value in this tool to aid collaboration with patients by using it as an *interactive data explorer during a therapy session.* For example, one clinician imagined querying the model to identify triggers for panic attacks and other anxiety symptoms. Another clinician outlined the high-level steps she would use if incorporating the model: *"First, set the goal... What's bringing them in? Then, agree on some metrics that are relevant."* (P4). Others imagined using it to *find evidence of change early in treatment* (e.g., indications of more energy or better sleep) as a patient struggled with whether to continue a particular medication or therapy. One therapist gave an example of how she might, through discussion with a patient, tie a concern such as relationship anxiety to the average duration of phone use periods and then use that metric to assess whether therapy was helping with the patient's anxiety. Several imagined using it not only for retrospective analysis but also *to forecast improvement.* Another envisioned use was *asking the model questions during sessions to boost creativity*, e.g., when challenging a patient's worries or negative thoughts.

Across these and other forms of collaborative use, *some clinicians wanted their patients to be able to query the model.* In addition to identifying patterns, joint use was envisioned as a way to build a more general feeling of collaboration in the therapy, something that is important for the therapeutic alliance and positive outcomes [12, 40].

4.6.2 Identifying Issues to Explore with the Patient. Clinicians appreciated the model's ability to list concerns with references to specific dates (e.g., days with little sleep or little movement). One clinician shared that it would be useful for GPT-4 to prioritize the questions. Clinicians suggested that they would share their observations with patients, e.g., about particular days with anomalies such as decreased sleep, as a way of opening up discussions and jogging the patient's memory.

When discussing potential concerns, clinicians wanted the model to raise questions and supply clear data. As one clinician said, the model "Could cause less harm if it provided questions for a therapist to ask instead of conclusions for a therapist to rely on" (P8). They did not want the model to apply diagnostic labels to individuals or their behaviors.

4.6.3 Generating Documentation. Clinicians varied in their thoughts on whether models would meaningfully aid in documentation. One participant appreciated the neutral boilerplate language the model used and easily imagined it as the basis for an intake summary. Another had already started using AI software to aid in documentation of sessions. Others bristled at this idea, pointing out that the LLM might miss major insights and obviate the analysis that comes from writing notes.

4.6.4 Concerns. Clinicians raised additional concerns related primarily to privacy and quality of care.

Privacy. As clinicians considered inputs, such as patient mood logs, that would increase the model's relevance to mental health, they grappled with privacy and other ethical concerns. One clinician worried that it could be hard to obtain meaningful informed consent. Even if a model were secure and compliant with relevant patient privacy regulations, it could be difficult for therapists to explain how the system worked and what protections it afforded.

Quality of Care. Another concern pertained to relying on the model as a shortcut. Clinicians imagined the problems that could arise from overconfidence in the model's analysis. One clinician described the possibility of missing the insights that would come from pouring over data herself: *"If I was...doing this very quickly, like, before I see the client, I could totally see myself or anyone...just relying on on GPT-4. And just think … 'This is the answer. This is the knowledge,' ... as mental shortcuts as opposed to pouring over the data yourself. … And I would want to ask, 'Am I missing anything?'* (P6). Another clinician worried that not taking the time to write one's own notes and process each session could degrade the their memory of the session and ultimately shortchange the therapy. And contrasted her efforts to examine contextual factors, such as family conflict, with a model's narrower focus on symptoms. She noted that if relying solely on a model's observations, a clinician could overlook important factors in assessing and treating mental health struggles.

5 DISCUSSION

Our findings reveal a paradigm shift based on the value LLMs can bring to mental health care, switching from a focus of using computational tools for diagnostic classification to using them to generate clinical reasoning. Most prior ML research in mental health has focused on classification, or predicting the diagnostic category determined by a clinical interview or score on a self-report inventory such as the PHQ. While these coarse categorizations have a role in mental health care, they do not capture the specifics of a particular patient's struggles and are insufficient to meaningfully guide treatment. Therapeutic change relies on the patient becoming aware of their patterns and options for changing them: it is an active collaboration rather than a procedure performed on a patient based on a particular diagnosis. Further, therapy is hyper-personalized, grounded in each patient's specific concerns, goals, contexts, and dynamics.

Our findings indicate that LLMs have the potential to illuminate an individual's patterns by analyzing disparate sensor data that is otherwise unwieldy for clinicians to interpret. Unlike a diagnostic wizard used before treatment, LLMs can *bring value in-session* and *throughout treatment.* Instead of being used solely by the clinician or another professional in a care system, LLMs can be *used collaboratively by the patient and clinician.* In this vision, the LLM does not replace the clinician or replace dialogue between the clinician and patient. Instead, the LLM *facilitates their dialogue*, the alliance that supports change, and their investigation of the patient's patterns.

Below, we first discuss the challenges of using LLMs for classification followed by further steps needed to develop end-to-end tools for the proposed clinician-patient-AI collaborative therapy.

5.1 Challenges of Depression Classification

This study illuminated new approaches to and challenges of screening for depression based on mobile health data. Our best-performing method for classifying depression was incorporating chain-of-thought (COT) prompting and variable explanations with the PaLM 2 model, achieving an accuracy of 61.11% on our class-balanced dataset.

Although this represents a modest improvement over classical ML methods such as Random Forest and the state-of-the-art Reorder [74] method on this dataset, *this level of accuracy is still far too low to be useful as a clinical screening tool.* It is important to note that personalized depression forecasting approaches where model training data includes historical data from test set individuals report accuracy as high as 90.1% [35], but a reliance on extensive prior labeled data for each given individual significantly limits real-world utility.

In our experiments, we utilized a class-balanced dataset, as described in Section 3.1.2. This provides an indication of performance across an even distribution of individuals with potential PHQ-4 scores. With an unbalanced dataset constructed from all complete samples in the Globem dataset, the distribution would include 34.8% positive samples as opposed to 50%. With this distribution, we would expect to see a shift in the optimal prompting strategies, improving performance for direct prediction and COT prompting methods more likely to classify samples as not depressed and further decreasing performance of prompting strategies, including DSM-V depression criteria.

We also attempted experiments to predict severity through a variety of methods, such as asking the LLMs to select a PHQ-4 score between 0 and 12 or options ranging from "none" to "severe" with varying degrees of additional context. In these instances, though, we notice that all three LLMs always select the middle value of the provided range. This result aligns with the findings of other works, which show that GPT-3.5 and GPT-4 both require few-shot prompting with multiple labelled input examples to perform these types of regression tasks [31]. We forego few-shot prompting experiments for this work since they would require either the individual to track PHQ-4 scores as ground-truth, which is the variable we are trying to predict with the LLM in the first place, or use labelled data excerpts from other participants, which raises the challenging question of how to source representative samples from existing datasets to use for a new individual.

5.2 Managing Inaccuracy in a Clinical Context

Clinicians in our study expressed a strong preference for LLMs to generate observations and insights about data relating to a patient rather than apply diagnostic labels. Such usage of LLM-based tools for generative reasoning avoids the challenges of inaccurate classification, but inaccuracies still need to be managed. In particular, well-documented numerical inaccuracies may impact the quality of the generated reasoning. In our study, reasoning responses generated by GPT-4, based on a 448-element table (16 features x 28 days), contain at least one numerical error 25% of the time. The necessary accuracy and tolerance for inaccuracy may vary with applications of the LLM. For example, an inaccuracy in a question that is generated for the therapist to ask a patient may pose less risk than an error in a definitive statement that informs a clinician's impressions of a patient's mental state. These risks of inaccuracy may be mitigated by uses that prioritize the therapeutic relationship and the patient's perspective. The clinicians in this study anticipated that they would use LLM-based tools in collaboration with patients, drawing on the tool's observations as a starting point for discussion. They emphasized that their treatment decisions and diagnoses would be based on their direct patient interactions.

5.3 Supporting Clinician-Patient-AI Collaboration

Supporting the kind of collaboration we propose between clinicians, patients, and AI will require new approaches to LLM hosting and data protection. This is due in part to the sensitivity of the data required. To inform mental health treatment and illuminate the factors associated with a particular individual's struggles, models need to have data that are directly relevant to mental health and, ideally, personalized to individual patients. The clinicians we interviewed expected the model to have, at a minimum, daily mood and symptom tracking and ideally more in-depth data related to mood, social interactions, and behavioral routines. In addition, a clinician may want to input materials such as session notes or transcripts, manuals, or other documents describing relevant mental health issues and treatments. Much of this potentially personally identifiable data is sensitive and not appropriate as an input for existing LLM services such as ChatGPT, which may use this data to improve their models or for other commercial purposes. *Privately hosted models or robust user data protections are required for this purpose.*

The solution to the privacy challenges of using LLMs in therapy is not as straightforward as it is for compliant medical record systems. Such record systems primarily serve providers and other employees at medical institutions, limiting patients to read-only access to elements such as test results. In this study, we heard from clinicians that both patients and clinicians should be able to interact with the model. This joint interaction, clinicians anticipated, could build patients' curiosity about factors associated with their mental health and foster the collaborative alliance between patients and therapists that is associated with positive outcome [12, 40].

Such models therefore require a very different approach than that used for medical records; in this case, *the patient should own their data, but both parties (the patient and therapist) should be able to generate data and actively interact with it.* Meeting the computational requirements of hosting such services may be challenging for providers in smaller practices, while relying on third-party services raises questions about data ownership. Addressing these concrete privacy challenges will open the door for broader exploration. *AI systems intended not just to produce a single diagnosis but rather to engage users in collaborative investigation over time will require new modes of interaction.* This will pose research and design challenges in designing systems for use in mental health care and other contexts.

5.4 Limitations and Future Work

5.4.1 Dataset Generalization and Balancing. In our experiments, we utilize a class-balanced dataset confined to a subset of mobile and behavioral health data collected from undergraduate students at a single university over a three-year period. Future evaluations should seek to include both a broader range of data elements as well as data from more diverse population samples to better understand how well these approaches apply to any potential individual seeking mental health treatment.

5.4.2 Model and Prompting Bias. In addition to issues of population representation in the dataset, a growing body of work suggests that state-of-the-art LLMs are prone to producing output containing both implicit and explicit biases [20, 57]. This is further complicated by the sensitivity of classification performance to the specific information included in the prompt. For example, providing the DSM-V depression criteria results in GPT-4, which one may assume may improve depression classification, classifying 98.89% of data excerpts as depressed despite a class-balanced dataset. It is almost certain that prompt and model-specific biases exist in the generated reasoning about the provided mobile health data as well, but this is a significantly more complex problem to characterize. Benchmarking techniques are actively being explored in this area [48, 79], and further work to understand how these biases arise and can be mitigated are essential before deployment of LLMs mental health contexts.

5.4.3 Numerical Errors. Clinicians envisioned inputting both a wide range of additional features as well as a longer time period of data into LLMs for analysis. However, the maximum context window of these models is limited, and prior work has shown that the ability of LLMs to accurately recall detailed information from the entirety of the context window decreases as prompt length increases[37]. Pre-computing or summarizing portions of data could help address this technical challenge, but requiring this step could significantly reduce the ability of a tool to generalize across potential input data sources.

Some errors observed by clinicians involved GPT-4 generating erroneous responses when asked follow-up questions that could not be answered based on the provided mobile health data. This challenge of *grounding*, or constraining LLM output to refer only to a relevant set of input text or data, is not unique to this specific application. Current approaches to improve performance in these areas include model fine-tuning[4]; augmenting a base general-purpose LLM with a smaller model optimized for a specific task[6]; and utilizing LLMs capable of querying external databases and tools to request data and perform arithmetic operations[21]. We view this third option as especially promising since enabling a model to access data from a user-controlled database may simultaneously help address the privacy challenges identified in Section 5.3.

5.4.4 Evaluation in Clinical Settings. While we find clinicians were frequently able to identify errors generated by GPT-4 when they arose during our interviews, in an actual therapy session the clinician may not be able to balance the demands of critically evaluating model output while engaging with the patient. Additionally, our current work solely evaluates LLM use in therapy from the perspective of clinicians. Further evaluation in a more realistic therapeutic environment is needed to better understand how use of LLMs as part of a mental health treatment program could impact patients and their outcomes.

6 CONCLUSION

This paper examines LLMs in the context of mental health care, specifically psychotherapy. While we begin our investigation by developing methods to use LLMs to output binary classifications for conditions like depression, we find instead that their greatest potential value to clinicians lies not in diagnostic classification, but rather in rigorous analysis of diverse self-tracking data to generate natural language summaries synthesizing multiple data streams and identifying potential concerns. Clinicians envisioned using those insights in a variety of ways, principally for fostering collaborative investigation with patients. This collaboration was seen as potentially valuable for strengthening the therapeutic alliance and guiding treatment. We describe a human-AI collaborative model and its requirements for secure management of personal data. These findings highlight directions for impactful future research on human-AI collaborative tools in mental health care and other contexts.

ACKNOWLEDGMENTS

This research was partially supported by an Amazon Research Award, a Google Research Scholar Award, The Washington Research Foundation, and the Pastry-Powered T(o)uring Machine Endowed Fellowship.

REFERENCES

[1] 2013. *Diagnostic and statistical manual of mental disorders : DSM-5* (fifth edition. ed.). American Psychiatric Association, Arlington, VA.
[2] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT. http://arxiv.org/abs/2303.03186
[3] Chizobam Ani, Mohsen Bazargan, David Hindman, Douglas Bell, Muhammad A. Farooq, Lutful Akhanjee, Francis Yemofio, Richard Baker, and Michael Rodriguez. 2008. Depression symptomatology and diagnosis: discordance between patients and physicians in primary care settings. *BMC Family Practice* 9, 1 (Jan. 2008), 1. https://doi.org/10.1186/1471-2296-9-1
[4] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An Open Language Model For Mathematics. https://doi.org/10.48550/arXiv.2310.10631 arXiv:2310.10631 [cs].
[5] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 5 (Jun 2017), 36 pages. https://doi.org/10.1145/3090051
[6] Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. LLM Augmented LLMs: Expanding Capabilities through Composition. https://doi.org/10.48550/arXiv.2401.02412 arXiv:2401.02412 [cs].
[7] Dror Ben-Zeev, Emily A. Scherer, Rui Wang, Haiyi Xie, and Andrew T. Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal* 38, 3 (2015), 218.
[8] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation. (2007).
[9] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot

Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. http://arxiv.org/abs/2303.12712

[12] Sarah Kate Cameron, Jacqui Rodgers, and Dave Dagnan. 2018. The relationship between the therapeutic alliance and clinical outcomes in cognitive behaviour therapy for adults with depression: A meta-analytic review. *Clinical psychology & psychotherapy* 25, 3 (2018), 446–456.

[13] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[14] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine* 3, 1 (March 2020), 1–11. https://doi.org/10.1038/s41746-020-0233-7

[15] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel, and Anind K. Dey. 2021. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing. *ACM Transactions on Computer-Human Interaction* 28, 1 (Jan. 2021), 1–41. https://doi.org/10.1145/3422821

[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[17] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. http://arxiv.org/abs/2204.02311 arXiv:2204.02311 [cs].

[18] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. http://arxiv.org/abs/2210.11416 arXiv:2210.11416 [cs].

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). http://arxiv.org/abs/1810.04805

[20] Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. 2023. Probing Explicit and Implicit Gender Bias through LLM Conditional Text Generation. https://doi.org/10.48550/arXiv.2311.00306 arXiv:2311.00306 [cs].

[21] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided Language Models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 10764–10799. https://proceedings.mlr.press/v202/gao23f.html

[22] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language* 45 (2017), 137–148.

[23] Aodhán Hickey. 2021. The rise of wearables: From innovation to implementation. In *Digital health*. Elsevier, 357–365.

[24] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301* (2023).

[25] Jeremy F Huckins, Alex W DaSilva, Elin L Hedlund, Eilis I Murphy, Courtney Rogers, Weichen Wang, Mikio Obuchi, Paul E Holtzheimer, Dylan D Wagner, and Andrew T Campbell. 2020. Causal Factors of Anxiety and Depression in College Students: Longitudinal Ecological Momentary Assessment and Causal Analysis Using Peter and Clark Momentary Conditional Independence. *JMIR Mental Health* 7, 6 (June 2020), e16684. https://doi.org/10.2196/16684

[26] Jeremy F Huckins, Alex W daSilva, Weichen Wang, Elin Hedlund, Courtney Rogers, Subigya K Nepal, Jialing Wu, Mikio Obuchi, Eilis I Murphy, Meghan L Meyer, Dylan D Wagner, Paul E Holtzheimer, and Andrew T Campbell. 2020. Mental Health and Behavior of College Students During the Early Phases of the COVID-19 Pandemic: Longitudinal Smartphone and Ecological Momentary Assessment Study. *Journal of Medical Internet Research* 22, 6 (June 2020), e20185. https://doi.org/10.2196/20185

[27] R Indrakumari, T Poongodi, P Suresh, and B Balamurugan. 2020. The growing role of Internet of Things in healthcare wearables. In *Emergence of Pharmaceutical Industry Growth with Industrial IoT Approach*. Elsevier, 163–194.

[28] Nino Isakadze and Seth S. Martin. 2020. How useful is the smartwatch ECG? *Trends in Cardiovascular Medicine* 30, 7 (Oct. 2020), 442–448. https://doi.org/10.1016/j.tcm.2019.10.010

[29] Nicholas C. Jacobson and Yeon Joo Chung. 2020. Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors* 20, 12 (June 2020), 3572. https://doi.org/10.3390/s20123572

[30] Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C. Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, Douglas Kondziolka, Alexander T. M. Cheung, Grace Yang, Ming Cao, Mona Flores, Anthony B. Costa, Yindalon Aphinyanaphongs, Kyunghyun Cho, and Eric Karl Oermann. 2023. Health system-scale language models are all-purpose prediction engines. *Nature* (June 2023). https://doi.org/10.1038/s41586-023-06160-y

[31] 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. https://arxiv.org/abs/2401.06866v1

[32] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* (2023), 101861.

[33] K. Kroenke, R. L. Spitzer, J. B.W. Williams, and B. Lowe. 2009. An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQ-4. *Psychosomatics* 50, 6 (Nov. 2009), 613–621. https://doi.org/10.1176/appi.psy.50.6.613

[34] Bishal Lamichhane. 2023. Evaluation of ChatGPT for NLP-based Mental Health Applications. http://arxiv.org/abs/2303.15727

[35] Heon-Jeong Lee, Chul-Hyun Cho, Taek Lee, Jaegwon Jeong, Ji Won Yeom, Sojeong Kim, Sehyun Jeon, Ju Yeon Seo, Eunsoo Moon, Ji Hyun Baek, Dong Yeon Park, Se Joo Kim, Tae Hyon Ha, Boseok Cha, Hee-Ju Kang, Yong-Min Ahn, Yujin Lee, Jung-Been Lee, and Leen Kim. 2023. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. *Psychological Medicine* 53, 12 (Sept. 2023), 5636–5644. https://doi.org/10.1017/S0033291722002847

[36] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. http://arxiv.org/abs/2303.14070 arXiv:2303.14070 [cs].

[37] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. https://doi.org/10.48550/arXiv.2307.03172 arXiv:2307.03172 [cs].

[38] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. In *arXiv*.

[39] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. *arXiv preprint arXiv:2305.15525* (2023).

[40] Daniel J Martin, John P Garske, and M Katherine Davis. 2000. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology* 68, 3 (2000), 438.

[41] Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, et al. 2019. The Tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[42] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. 2014. Toss "n" turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/2556288.2557220

[43] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino G. Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, et al. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2 (2019), 37:1–37:24. https://doi.org/10.1145/3328908

[44] Stefanie Nickels, Matthew D Edwards, Sarah F Poole, Dale Winter, Jessica Gronsbell, Bella Rozenkrants, David P Miller, Mathias Fleck, Alan McLean, Bret Peterson, et al. 2021. Toward a mobile platform for real-world digital measurement of depression: User-centered design, data quality, and behavioral and clinical modeling. *JMIR mental health* 8, 8 (2021), e27589.

[45] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. http://arxiv.org/abs/2303.13375 arXiv:2303.13375 [cs].

[46] Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466* (2023).

[47] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[48] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. https://doi.org/10.48550/arXiv.2110.08193 arXiv:2110.08193 [cs].

[49] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476* (2023).

[50] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

[51] 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. *Journal of Machine Learning Research* (2020).

[52] Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=yKbprarjc5B

[53] Darius A. Rohani, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2018. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. *JMIR mHealth and uHealth* 6, 8 (Aug. 2018), e9691. https://doi.org/10.2196/mhealth.9691

[54] Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research* 17, 7 (2015), 1–11. https://doi.org/10.2196/jmir.4273

[55] Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. 2018. A Weakly Supervised Learning Framework For Detecting Social Anxiety And Depression. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 26.

[56] Yasaman S. Sefidgar, Woosuk Seo, Kevin S. Kuehn, Tim Althoff, Anne Browning, Eve Riskin, Paula S. Nurius, Anind K. Dey, and Jennifer Mankoff. 2019. Passively-sensed behavioral correlates of discrimination events in college students. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 114 (Nov 2019), 29 pages. https://doi.org/10.1145/3359216

[57] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. https://doi.org/10.48550/arXiv.2212.08061 arXiv:2212.08061 [cs].

[58] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. 7059–7073.

[59] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. http://arxiv.org/abs/2305.09617 arXiv:2305.09617 [cs].

[60] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971 arXiv:2302.13971 [cs].

[62] Andreas Triantafyllidis, Haridimos Kondylakis, Konstantinos Votis, Dimitrios Tzovaras, Nicos Maglaveras, and Kazem Rahimi. 2019. Features, outcomes, and challenges in mobile health interventions for patients living with chronic diseases: A review of systematic reviews. *International Journal of Medical Informatics* 132 (Dec. 2019), 103984. https://doi.org/10.1016/j.ijmedinf.2019.103984

[63] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR mHealth and uHealth* 4, 3 (2016), e111. https://doi.org/10.2196/mhealth.5960

[64] Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev. 2016. CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016), 886–897. https://doi.org/10.1145/2971648.2971740

[65] 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 *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[66] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 295–306.

[67] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26. https://doi.org/10.1145/3191775

[68] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[69] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. http://arxiv.org/abs/2109.01652 arXiv:2109.01652 [cs].

[70] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. http://arxiv.org/abs/2201.11903 arXiv:2201.11903 [cs].

[71] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Further Finetuning LLaMA on Medical Papers. http://arxiv.org/abs/2304.14454 arXiv:2304.14454 [cs].

[72] Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–27. https://doi.org/10.1145/3448107

[73] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34.

[74] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34. https://doi.org/10.1145/3569485

[75] Xuhai Xu, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. Listen2Cough: Leveraging End-to-End Deep Learning Cough Detection Model to Enhance Lung Health Assessment Using Passively Sensed Audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–22. https://doi.org/10.1145/3448124

[76] 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. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 18.

[77] Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2023. Effective distillation of table-based reasoning ability from llms. *arXiv preprint arXiv:2309.13182* (2023).

[78] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the Evaluations of ChatGPT and Emotion-enhanced Prompting for Mental Health Analysis. http://arxiv.org/abs/2304.03347

[79] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating Interfaced LLM Bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, Jheng-Long Wu and Ming-Hsiang Su (Eds.). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei City, Taiwan, 292–299. https://aclanthology.org/2023.rocling-1.37

[80] Han Zhang, Margaret E. Morris, Paula S. Nurius, Kelly Mack, Jennifer Brown, Kevin S. Kuehn, Yasaman S. Sefidgar, Xuhai Xu, Eve A. Riskin, Anind K. Dey, and Jennifer Mankoff. 2022. Impact of Online Learning in the Context of COVID-19 on Undergraduates with Disabilities and Mental Health Concerns. *ACM Transactions on Accessible Computing* (July 2022), 3538514. https://doi.org/10.1145/3538514

[81] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198* (2023).

[82] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. http://arxiv.org/abs/2205.10625 arXiv:2205.10625 [cs].

A DATA ELEMENTS

GLOBEM Data Feature	Description
date	date
f_loc:phone_locations_doryab_totaldistance:allday	total distance traveled (meters)
f_loc:phone_locations_doryab_timeathome:allday	time spent at home (minutes)
f_loc:phone_locations_doryab_locationentropy:allday	location entropy
f_screen:phone_screen_rapids_sumdurationunlock:allday	phone screen time (minutes)
f_screen:phone_screen_rapids_avgdurationunlock:allday	average phone unlock duration (minutes)
f_call:phone_calls_rapids_incoming_sumduration:allday	phone call incoming duration (minutes)
f_call:phone_calls_rapids_outgoing_sumduration:allday	phone call outgoing duration (minutes)
f_blue:phone_bluetooth_doryab_uniquedevicesothers:allday	unique Bluetooth devices discovered nearby
f_steps:fitbit_steps_intraday_rapids_sumsteps:allday	step count
f_steps:fitbit_steps_intraday_rapids_countepisodesedentarybout:allday	number of sedentary episodes
f_steps:fitbit_steps_intraday_rapids_sumdurationsedentarybout:allday	total time spent sedentary (minutes)
f_steps:fitbit_steps_intraday_rapids_countepisodeactivebout:allday	number of activity episodes
f_steps:fitbit_steps_intraday_rapids_sumdurationactivebout:allday	total time spent active (minutes)
f_slp:fitbit_sleep_intraday_rapids_sumdurationasleepunifiedmain:allday	total time asleep (minutes)
f_slp:fitbit_sleep_intraday_rapids_sumdurationawakeunifiedmain:allday	total time spent awake while in bed (minutes)

Table 1. Data Fields and Descriptions

B PROMPT STRUCTURE

Each following subsection outlines the specific text used for each component of the prompt. To re-create our prompts for a given experiment, concatenate the relevant sections as shown in Fig. 1.

B.1 Role

```
Role:
You are a data analyst helping a psychiatrist understand human activity data.
```

B.2 Task Description

The numerically listed items in this block should be added or removed according to the blocks included in the rest of the prompt.

```
Task:
You will be shown data gathered from a smartphone and smart watch worn by an individual. Your
    goal is to analyze this data.
You are presented with the following:
1. A table consisting of twenty-eight days of collected activity tracking data [Collected Data]
2. Explanation of the different types of data. [Description of Variables]
3. DSM-5 Criteria for depression. [Depression Criteria]
4. Instructions on how to analyze the data [Instructions]
```

B.3 Formatted Data Table

Begin with a header *Collected Data:* followed by the data. Each of the data formats used in this paper can be found in Appendix D.

B.4 Describe Variables

These descriptions may vary based on the specific dataset used.

```
Description of variables
For the duration of the data collection period, data was logged from the individual's
    smartphone and a fitness tracking smartwatch. An explanation of the different
    types of data logged is included below:
Date: the data in this row was recorded on that date.
total_Distance_traveled(meters): This represents the total distance traveled as measured
    by the GPS in the participant's smartphone. This includes walking, driving, and any
    other modes of transportation.
time_at_home(minutes): This is the total time that the participant was at home as
    determined by their smartphone GPS.
. . .
```

B.5 DSM-V Definition

```
Depression Criteria
Here is the DSM-5 Criteria for depression. The DSM-5 outlines the following criterion to
   make a diagnosis of depression. The individual must be experiencing five or more symptoms
    during the same 2-week period and at least one of the symptoms should be either
    (1) depressed mood or (2) loss of interest or pleasure. . .
```

B.6 Ask for Reasoning and/or Classification

Specify the type(s) of analysis required. The example below includes both reasoning and classification instructions.

```
Instructions
1. Based on the collected data, hypothesize about the overall mental health of the individual
    with detailed explanations.
2. Make your best guess if this person is currently experiencing depression. Please answer
    only "Yes" or "No" to this question.
```

C EXAMPLE GPT-4 RESPONSE

Below is a response generated by GPT-4 from a prompt consisting of the Chain of Thought, Variable Description, and DSM-V Definition elements:

To assess the mental health of the individual, we'll analyze the provided data and relate it to the DSM-5 criteria for depression. Hypothesis About Overall Mental Health:

- **Physical Activity**: The individual's step count fluctuates significantly, with days of very high physical activity (e.g., 55,755 steps on 2019-05-11) and others with much lower activity (e.g., 2,050 steps on 2019-05-20). The inconsistency in physical activity could indicate variability in energy levels or mood.
- **Time at Home**: There are significant variances in the amount of time spent at home. Some days show extended periods at home (e.g., 1,240 minutes on 2019-05-12), while others show much less (e.g., 5 minutes on 2019-05-19). Spending extended periods at home might suggest a lack of engagement in external activities, possibly indicating social withdrawal or lack of interest.

- **Phone and Screen Usage**: The phone screen time and average phone use unlock duration fluctuate, with some days showing high usage (e.g., 819 minutes on 2019-05-21) and others considerably lower (e.g., 138 minutes on 2019-05-19). This could indicate variations in the need for distraction or communication.
- **Sedentary and Active Episodes**: There's inconsistency in the number of sedentary and active episodes. Some days have high sedentary time and low activity (e.g., 1,402 sedentary minutes vs. 38 active minutes on 2019-05-20), which might suggest fatigue or lack of energy.
- **Sleep Patterns**: The total time asleep and awake while in bed varies. There are nights with less sleep (e.g., 249 minutes on 2019-05-10) and others with more (e.g., 679 minutes on 2019-05-12). Irregular sleep patterns could be indicative of disturbed sleep, which is a symptom of depression.
- **Location Entropy**: On most days, location entropy is 0, suggesting a lack of diversity in location visits. This could imply a routine life with limited exploration or lack of interest in diverse activities.
- **Phone Call Duration**: There is limited data on phone call durations, but on days with data, durations vary. This inconsistency might suggest fluctuating social interactions.

Best Guess if Experiencing Depression: Yes.

The data suggests variability in physical activity, time spent at home, phone usage, and sleep patterns. This inconsistency could reflect symptoms such as fatigue, loss of interest, and disturbed sleep, aligning with some depression criteria. However, it's important to note that this analysis is limited and cannot definitively diagnose depression. A comprehensive assessment by a mental health professional is necessary for an accurate diagnosis.

D DATA INPUT FORMAT

The following subsections include snippets detailing the input formats described in Section 3.2.

D.1 CSV

```
date,total_distance_traveled(meters),time_at_home(minutes),location_entropy,...
2019-04-29,49037.0,666.0,0.85,298.0,3.0,,,29.0,11430.0,40.0,1290.0,39.0,150.0,306.0,11.0
2019-04-30,69171.0,555.0,0.87,274.0,4.0,16.0,,13.0,11427.0,49.0,1264.0,48.0,176.0,,
. . .
```

D.2 Markdown

```
date|total_distance_traveled(meters)|time_at_home(minutes)|location_entropy|...
2019-04-29|49037|666|0.85|298|3|nan|nan|29|11430|40|1290|39|150|306|11|
2019-04-30|69171|555|0.87|274|4|16|nan|13|11427|49|1264|48|176|nan|nan|
. . .
```

D.3 Tabular

```
date total_distance_traveled(meters) time_at_home(minutes) location_entropy . . .
2019-04-29 49037.0 666.0 0.85 298.0 3.0 29.0 11430.0 . . .
2019-04-30 69171.0 555.0 0.87 274.0 4.0 16.0 13.0 11427.0 . . .
. . .
```

D.4 LaTeX

```
\begin{tabular}{lrrrrrrrrrrrrrr}
\toprule
date & total_distance_traveled(meters) & time_at_home(minutes) & location_entropy & . . . \\
```

```
\midrule
2019-04-29 & 49037.0 & 666.0 & 0.850 & 298.0 & 3.0 & NaN & . . . \\
2019-04-30 & 69171.0 & 555.0 & 0.870 & 274.0 & 4.0 & 16.0 & . . . \\
\bottomrule
\end{tabular}
```

E PRODUCING REASONING PROMPT

Below is the prompt used to generate the reasoning samples used in Section 3.3.1:

```
Role:
You are a data analyst helping a psychiatrist understand human activity data.

Task:
You will be shown data gathered from a smartphone and smart watch worn by an individual. Your
    goal is to analyze this data.
You are presented with the following:
1. A table consisting of twenty-eight days of collected activity tracking data [Collected Data]
2. Instructions on how to analyze the data [Instructions]

Collected Data
date|total_distance_traveled(meters)|time_at_home(minutes)|...
|2019-05-06|11996|1012|...
. . .

Instructions
Although the data does not contain explicit information on mood, trends in physiological
    signals have been shown to correlate with mental health symptoms. Examine this data
    and point out any specific trends or data points that could spark fruitful
    conversation with a mental health professional.
```

F REASONING GRADER INSTRUCTIONS

Thank you for taking the time to contribute to this study.

To start, please open this document that contains a table of data as well as plots of the data.

Link to document: [LINK HERE]

You will now be asked to grade a series of 32 different statements analyzing this data. Your goal is to check the accuracy of these statements to ensure that references to the data are correct

Here an explanation of the grading rubric. Please read this rubric carefully:

1. Does this response include numbers? (yes/no)

Yes – at least some part of the response lists or quotes specific numerical data or dates, regardless of correctness

No – the response does not include any specific numbers

Note – numbered lists don't count as numbers

2. Are these numbers consistent with the provided data? (yes/no)

Yes – all of the mentioned numbers or dates are included in the provided data

No – some or all of the numbers or dates are not consistent with the provided data, or there are no numbers (1 was answered "No")

For example:

- The text statement says the highest sleep time occurred on May 9, but based on the graph you can see it is actually on June 2
- The text statement lists the lowest distance travelled as 127 meters, but the lowest distance traveled listed in the table is 1270 meters

3. Does this response identify specific trends? (yes/no)

Yes – the response makes statements relating to concepts like minimum, maximum, averages, variability, upward or downward trends, etc. as they pertain to the data

No – There is no statement of specific trends that relate to the included data.

For example:

- “An increase in sleep might indicate a disturbance” or "the individual makes phone calls" would not be a specific trend relating to the provided data
- “The time spent asleep increased in the second half of the month” would be a specific trend relating to the data

4. Are these trends consistent with the provided data? (yes/no)

Yes – the listed trends are plausibly consistent with the provided data table and/or plots

No – some or all of the listed trends are contradicted by the provided data and/or plots or there are no specific trends (3 was answered “No”)

It is important to note that you should not evaluate further trends or reasoning as they may relate to, for example, mental health. For the purposes of grading these responses, it is only necessary to confirm if the response does or does not accurately describe the provided data.

We anticipate it will take 1.5-2 minutes to grade each statement.