Exploring Large-Scale Language Models to Evaluate EEG-Based Multimodal Data for Mental Health

Yongquan Hu yongquan.hu@unsw.edu.au University of New South Wales Sydney, NSW, Australia

Hong Jia h.jia.cam@gmail.com University of Melbourne Melbourne, VIC, Australia Shuning Zhang zsn23@mails.tsinghua.edu.cn Tsinghua University Beijing, China

Flora D. Salim flora.salim@unsw.edu.au University of New South Wales Sydney, NSW, Australia

> Aaron J. Quigley aquigley@acm.org CSIRO's Data61 Sydney, NSW, Australia

Ting Dang ting.dang@unimelb.edu.au University of Melbourne Melbourne, VIC, Australia

Wen Hu wen.hu@unsw.edu.au University of New South Wales Sydney, NSW, Australia

Abstract

Integrating physiological signals such as electroencephalogram (EEG), with other data such as interview audio, may offer valuable multimodal insights into psychological states or neurological disorders. Recent advancements with Large Language Models (LLMs) position them as prospective "health agents" for mental health assessment. However, current research predominantly focus on single data modalities, presenting an opportunity to advance understanding through multimodal data. Our study aims to advance this approach by investigating multimodal data using LLMs for mental health assessment, specifically through zero-shot and few-shot prompting. Three datasets are adopted for depression and emotion classifications incorporating EEG, facial expressions, and audio (text). The results indicate that multimodal information confers substantial advantages over single modality approaches in mental health assessment. Notably, integrating EEG alongside commonly used LLM modalities such as audio and images demonstrates promising potential. Moreover, our findings reveal that 1-shot learning offers greater benefits compared to zero-shot learning methods.

CCS Concepts

Human-centered computing → Ubiquitous and mobile computing;
Applied computing → Life and medical sciences.

Keywords

Mental Health, EEG, Large Language Model, Prompt Engineering.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp Companion '24, October 5–9, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-1058-2/24/10

https://doi.org/10.1145/3675094.3678494

ACM Reference Format:

Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D. Salim, Wen Hu, and Aaron J. Quigley. 2024. Exploring Large-Scale Language Models to Evaluate EEG-Based Multimodal Data for Mental Health. In Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '24), October 5–9, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3675094.3678494

1 Introduction

Mental health, as defined by the World Health Organization (WHO), is a state of well-being where individuals can realise their potential, handle normal life stresses, work productively, and contribute to their communities [31]. Mental health issues are increasingly impacting the global economy [11], with conditions such as depression and anxiety estimated to cost trillions of dollars in lost productivity annually [8].

The accurate measurement and classification of such health conditions requires psychological evaluation which can include the recording of various indicators. Commonly, many physiological signals, such as Electroencephalogram (EEG) [12], Heart Rate Variability (HRV) [15], and Electrodermal Activity (EDA) [14], are integral for mental health assessments due to their reliability and difficulty to mask, ensuring more accurate identification [18, 40]. These signals are readily captured by widely available sensors [2, 10, 35].

In addition to capturing data, advancements in Artificial Intelligence (AI) technology have led researchers to develop various algorithms (e.g., machine learning) for the timely and accurately detection [1], modeling [43] and inference [29] of health conditions based on physiological signals. Recently, the capabilities of Large-scale Language Models (LLMs) have introduced a new paradigm for prediction and assessment in mental health [24, 42, 44]. LLMs offer several advantages, including enhanced multimodal data processing for improved assessment accuracy [25], interactive communication methods like human-in-the-loop to create more configurable health agents [4], and the potential for fine-tuning domain-specific purpose based on general models to reduce costs [39, 44]. However, most work using LLMs to detect mental health focuses on tasks of

single modality data such as Mental-LLM [44] and EEG-GPT [21], and the exploration of LLMs in evaluating multimodal sensing data for mental health remains limited. Moreover, existing multimodal LLMs have been developed primarily using audio and video modalities. They may lack the capabilities in handling other types of data, such as EEG and other physiological signals which play a crucial role [12] in mental health assessment. Among various physiological signals, EEG is particularly valuable, providing high-frequency data that accurately assesses conditions such as depression, mood, and stress levels [5]. Therefore, understanding how these LLMs process EEG data and how to effectively combine EEG with existing modalities remains an open question.

This paper introduces MultiEEG-GPT, a method for assessing mental health using multimodalities, especially with EEG, i.e., EEG and facial expression or audio. The latest GPT-40 API¹ is adopted for processing multimodalities to recognize the health conditions. Unlike its predecessors such as GPT-4 and GPT-4v, which require separate interface calls, GPT-40 integrates multimodal data processing into a single interface, enhancing the development of this method [30]. This work aims to understand the capabilities of multimodal LLMs in categorising various mental health conditions. This work seeks to compare their ability to model different modalities and EEG and design optimal prompt engineering to facilitate reliable prediction.

The contributions of this paper include: i) the prompt engineering design using both zero-shot and few-shot approaches to examine the predictive capability of MultiEEG-GPT using multimodalities in recognizing different health conditions; ii) experiments across three different databases to validate the effectiveness of MultiEEG-GPT. iii) an in-depth analysis to understand how multimodalities enhance health condition predictions compared to single modalities. We aim to open up further developments, such as health-supportive social robots [4, 19, 23], within the context of ubiquitous computing, human-computer interaction, and affective computing.

2 Related Work

EEG-based physiological signal analysis has long been essential for monitoring mental health, evolving alongside AI advancements. Initially focused on traditional machine algorithms like k-nearest Neighbor (k-NN) and Support Vector Machine (SVM) for EEG data, Hou et al. demonstrated the potential of EEG for stress level recognition, with the accuracy of 67.07% [17]. Later, the field has shifted towards integrating deep learning and multimodal data. Zhongjie et al. developed a fusion algorithm levering deep neural networks that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks for emotion classification, markedly demonstrating the impressive accuracy in valence and arousal classifications to 93.20±2.55% and 93.18±2.71%, respectively [26].

Recently, the advent of general LLMs capable of processing multimodal data has further pivoted the focus towards using LLMs for evaluating mental health data, anticipating their role as future evaluation agents. For example, Xuhai et al. tested various LLMs, including GPT-3.5 and GPT-4, across multiple datasets using methods like zero-shot and few-shot prompting [44]. Jonathan et al.

introduced EEG-GPT, using GPT models to classify and interpret EEG data [21]. However, these studies still focus on single modality, such as text or EEG. Given various modalities can provide rich and complementary information to infer health conditions, it is proposed to consider different modalities in the automatic systems as well, especially with EEG in many mental health applications. However, research on LLMs for multimodal data with EEG is still limited for mental health prediction. Our proposed MultiEEG-GPT pioneers the work in examining multimodal data including EEG to infer health conditions, aiming to bridge this gap by enhancing the processing of multimodal signals, with a particular focus on EEG data.

3 Methodology

3.1 Dataset Selection

Various mental health dataset existed, of which numerous contained EEG modality. Applying the criteria that the dataset need to contain at least EEG modality, we selected 3 most commonly used datasets: (1) MODMA [5] was developed by Hanshu et al., and this multimodal dataset is designed for analyzing depression disorders and includes oral records (audio) of both patients and controls, and EEG data (convertible to images) from these groups. This dataset has binary labels of whether the participant was diagonsed with Major Depressive Disorder (MDD). (2) PME4 [7] is a multimodal emotion dataset featuring four modalities: audio, video (not publicly available), EEG, and electromyography (EMG) [7]. It was collected from 11 acting students (five female and six male) who provided informed consent. This dataset focuses on identifying seven emotions: anger, fear, disgust, sadness, happiness, surprise, and a neutral state; (3) LUMED-2 [9] was collected by Loughborough University and Hacettepe University, and it was designed to analyze facial expression, EEG, and galvanic skin response (GSR) data to recognize and classify three categories of human emotions (neutral, happy, sad) under various stimuli, advancing the understanding in affective computing.

For MODMA and PME4, we used audio and EEG modalities, while for LUMED-2, we used facial expression and EEG modalities. We chose audio and facial expression features because they were the among the most prevalent modalities in mental health analysis [28, 37]. Besides, the focus of this paper was to explore the possibility of GPT to analyze multimodal data, particularly with the important EEG modality [16]. Thus, we did not include the physiological modalities (e.g., GSR, Resp).

3.2 Prompt Design

For our MultiEEG-GPT method, we use prompt engineering strategies (including zero-shot prompting and few-shot prompting) for prediction tasks on multiple datasets. These prompts are model-agnostic, and we present the details of language models and settings employed for our experiment in the next section.

For the prompting strategies, we built upon the design in [44] and [45]. We have designed the prompt to account for different modalities and incorporated flexibility in altering the number of modalities for evaluation. Additionally, we have verified and compared few-shot and zero-shot prompts for evaluation.

 $^{^{1}} https://openai.com/index/hello-GPT-4o/,\ accessed\ on\ June\ 11,\ 2024$

Zero-shot prompting. As shown in Table 1, the zero-shot prompting strategy consists of a role-play prompt, a specific task description, and an additional rule to avoid unnecessary output and restricted models to focus on the current task. The role-play prompt aims to inform the LLMs of the general task, while the specific task description provides the information for different modalities. Such description also provides the flexibility in adding or deleting modalities. Therefore, the final prompt for the model consisted of: {role-play prompt} + {task specification} + {rules}.

Few-shot prompting. The few-shot prompt added the few-shot samples after the same zero-shot prompt template. Specifically, we include the task-specific prompt followed the zero-shot prompt, but providing the correct class labels instead of offering different candidate class labels for prediction, which is similar to Xuhai et al's setting [44].

Table 1: The zero-shot and few-shot prompting strategies. <MOD1>, <MOD2> and <MOD3> as placeholders denote three different modalities. XXX is the description of collection and visualization process. <SYM> as a placeholder denotes the symptom to be diagnosed. For example, for depression analysis <SYM> should be replaced with depression. The example is for mental health diagonsis with three classes. The label description "0 denotes XXX" of the classes could be added or removed to accommodate for more or less classes.

Role-play prompt	Imagine you are a mental health expert expert at analyzing the emotion and mental health status.
Task specification	The below is <mod1>, <mod2> and <mod3> data. <mod1> data is collected through XXX and visualized in XXX form. <mod2> data is collected through XXX and visualized in XXX form. <mod3> data is collected through XXX and visualized in XXX form. Analyze the <sym> status of the person. 0 denotes XXX, 1 denotes XXX, 2 denotes XXX.</sym></mod3></mod2></mod1></mod3></mod2></mod1>
Rules	[Rule]: Do not output other text.

4 Experiment

4.1 Settings

4.1.1 Dataset Settings. As all the datasets used standard 10-20 electrode layout, we set the electrodes following this layout. MNE library is used for processing EEG signal. We processed the datasets using the raw data instead of their pre-processed data (e.g., PME4) because the pre-processed data only contained features instead of the original signals, which were infeasible for plotting topology map. We used a bandpass filter (low-frequency cutoff 0.1Hz, high-frequency cutoff 45Hz, Hamming Window) [34] with firwin window design. Afterward, the filtered data were re-referenced to an average reference [34]. Since the elicitation presented with different length for different datasets, we chose 530s, 5s and 1.65-4.15s for LUMED-2, PME4 and MODMA datasets respectively, to account for randomly set elicitation time, with 10 equidistant sampled

timestamps to create topology maps. For the facial expression, we chose the middle frame of the video (e.g., if the video's length is 10s, we chose the frame at exactly the 5s timestamp) or the image. For the audio, because GPT-4o ² have not yet released the audio input support, we used both the audio features and the text as inputs. For the audio features, we used librosa library to extract the features from the audio and represent these features in text format (which is similar to EEG-GPT's representation [21]), which includes MFCCs, Mel Spectrogram, Chroma STFT, etc. For the text, we transcribed the audio using automatic speech recognition (ASR) systems. We chose the open-sourced vosk library ³ with vosk-model-cn-0.15 (Chinese version) or vosk-model-en-0.22 (English version) according to the need. These models were the largest and most advanced ASR systems in the vosk library, which ensured the accuracy of recognition and was used in health care tasks [13, 32].

4.1.2 Model Settings. For all datasets and all tasks, we transformed the tasks into multi-class classification problems as in previous work [21, 44]. For MODMA, the binary class labels are 'MDD' or 'healthy'. For PME4, we followed the labels in the original datasets to classify the emotion into seven classes: anger, fear, disgust, sadness, happiness, surprise and neutral. For LUMED-2, we set the 3-class labels as in the original paper, which included neutral, happy ans sadness.

Previous work showed that GPT-4 generally performed better than GPT-3.5 [44]. Given that GPT-40 is the most recent series of GPT-4 that naturally supports multimodal capabilities, we adopted GPT-40 as the tested LLMs. Specifically, we used "GPT-40-2024-05-13" as the targeted model through OpenAI Azure's API⁵. For the few-shot experiment, we tested the 1-shot learning scenario to examine the capability of multi-model LLMs with limited information provided. In each repeated trial, we randomly selected one sample from the corresponding dataset to act as the 1-shot sample. This strategy mitigates the bias of selecting samples. For all zero-shot and few-shot experiments, we tested across each dataset (for the few-shot experiment, we excluded that selected sample) for 5 times and reported the average accuracy and the standard deviation.

We use the image updating module of GPT-4o. However, we use no other any additional techniques (e.g., Chain-of-Thoughts [41]) to serve as a preliminary study in understanding how multimodal LLMs process multimodal information. This approach ensures the results reflect teh basic capability of the models, which was also consistent with previous work [21, 44].

4.2 Results and Discussions

4.2.1 Multimodal analysis. We showed two examples of zero-shot cases using LUMED-2 and PME4 dataset in Figure 1. The first person in the LUMED-2 video is in neutral mood. The MultiEEG-GPT aims to recognize the participant's mental state through the facial expression and the EEG topology map. As seen in Figure 1(a), MultiEEG-GPT first processed the image, and then analyzed the EEG topology map. It subsequently aggregated the results from

²https://community.openai.com/t/when-the-new-voice-model-for-chatgpt-4o-will-be-released/789928, accessed by Jun 16th, 2024

³https://alphacephei.com/vosk/, accessed by Jun 16th, 2024

⁴https://openai.com/index/hello-GPT-4o/, accessed by 11st June, 2024

⁵https://azure.microsoft.com/en-us/products/ai-services/openai-service, accessed by 11st June, 2024

Table 2: Ablation experiment on 3 different multimodal data (EEG image, facial expression, audio). The line with no EEG image, facial expression, audio was determined through majority voting. For few-shot prompting, we chose M=1, which meant we added one few-shot sample in the prompt.

Prediction Accuracy (%)								
Strategy	EEG	Facial Expression	Audio	MODMA	PME4	LUMED-2		
Zero-shot Prompting	×	×	×	$50.0_{\pm 0.00}$	14.28 _{±0.00}	33.33 _{±0.00}		
	✓	×	×	$53.79_{\pm 2.46}$	$21.05_{\pm 1.71}$	$34.61_{\pm 1.28}$		
	×	\checkmark	×	_	_	$38.46_{\pm 1.54}$		
	×	×	\checkmark	$69.35_{\pm 2.53}$	$15.38_{\pm 1.42}$	_		
		✓	×	_	_	46.13 _{±2.42}		
	\checkmark	×	✓	$73.54_{\pm 2.03}$	$28.57_{\pm 2.41}$	-		
Few-shot Prompting (M=1)	×	×	×	50.0 _{±0.00}	14.28 _{±0.00}	33.33 _{±0.00}		
	\checkmark	×	×	$62.71_{\pm 3.23}$	$26.00_{\pm 1.78}$	$36.37_{\pm 1.62}$		
	×	\checkmark	×	_	_	$43.64_{\pm 1.85}$		
	×	×	\checkmark	$69.92_{\pm 1.53}$	$19.13_{\pm 1.29}$	-		
	√	✓	×	_	_	52.73 _{±2.16}		
	\checkmark	×	\checkmark	$79.00_{\pm 1.59}$	$37.00_{\pm 2.30}$	-		

image and EEG jointly, and predicted the participant's emotion state as neural.

For Figure 1 (b), the participant is in a sad mood. The MultiEEG-GPT first analyzed the person's audio features, and then analyzed the EEG features in the topology map through the color of the map. It finally combined different features and predicted that the participant is in a sad mood. These cases showed the capability of MultiEEG-GPT to (1) analyze each modality separately, (2) aggregated the outputs based on different modalities jointly. It is also evident that a single modality is not adequate to identify the mood correctly. For example, MultiEEG-GPT identified the status of Figure 1 (b) as "an emotional reaction". However, it did not accurately state that the participant is sad from the EEG features. By combining the audio features and the EEG features, MUltiEEG-GPT achieved the accurate prediction.

4.2.2 Performance of MultiEEG-GPT. Table 2 presents the zeroshot and few-shot prompting performance for all three databases. The modalities used for MultiEEG-GPT depend on their availability in the datasets. For zero-shot prompting, our proposed model, utilizing both modalities—either EEG + facial expression or EEG + audio—achieved the best performance compared to other models using a single modality. The proposed model demonstrated relative improvements of 4.19%, 7.52%, 7.67% over the best single-modality performance for the three databases, respectively. This also highlighted the importance of including EEG data in addition to the commonly used modalities in LLMs, such as audio and video. It should be noted that the cases with all modalities removed (the first line) used majority voting similar to Xuhai et al.'s setting [44], serving as the baseline for model performance.

For the few-shot prompting, we observed a similar trend, with multimodal models outperforming single-modality models. Additionally, the 1-shot prompting achieved higher performance than zero-shot prompting, with relative improvements of 5.45%, 8.43%,

6.60% over zero-shot prompting for the MODMA, PME4 and LUMED-2 databases, respectively. This suggests that additional examples enhance recognition, consistent with previous findings [27, 38]. The extra example likely serves as a benchmark for feature comparison, allowing LLMs to assess the users' mental health status more effectively by comparing features of the few-shot and test samples. The results indicate the general benefit of an additional example, as no specific sample was intentionally selected in the 1-shot prompting setting. In summary, LLMs leveraging multimodalities including EEG could significantly benefit depression and emotion recognition.

5 Conclusion and Future Work

This paper proposes MultiEEG-GPT to explore multimodal data, specifically with EEG, for mental health recognition. We have designed zero-shot and few-shot prompting strategies to enhance prediction accuracy, leveraging the most recent GPT-40 as the LLM base model. Three datasets, including MODMA, PME4, and LUMED-2, were adopted for evaluation. Our study showed that predictions using multimodal data significantly outperform those using single-modal data. While the current prediction accuracy approaches that of traditional machine learning methods even without tuning the LLMs, there is significant potential for improvement with strategies such as instruction fine-tuning or multi-strategy hierarchical prediction in future research for mental health leveraging multimodal LLMs.

Moreover, the use of LLMs as health agents raises important ethical considerations. LLMs may exhibit value alignment problems, leading to racial and gender disparities [46] or producing outcomes misaligned with health assessment standards [20]. LLMs also pose privacy risks [3, 33] due to data memorization and extraction [6]. Fine-tuning with mental health data can lead to data leakage. These issues necessitate careful attention to ensure ethical compliance and accuracy. For example, input data should be anonymized beforehand

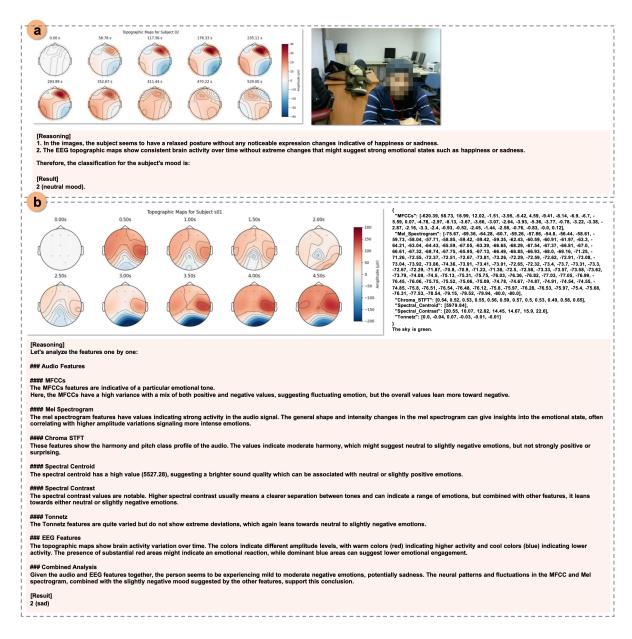


Figure 1: Case analysis for LUMED-2 and PME4 datasets (the person's face has been blurred for ethical reasons). Figure (a) illustrates one subject's input EEG topology map and his facial expression, as well as the prediction result and the text explanation from LUMED-2 dataset. Figure (b) illustrates one subject's input EEG topology map, audio features, input audio transcription "The sky is green.", as well as the prediction result and the explanation, from PME4 dataset. In both cases, the model makes the accurate predictions when processing both modalities.

[36], and un-learning and alignment should be integrated to the training process to protect privacy and avoid harm [22].

References

- [1] Rohizah Abd Rahman, Khairuddin Omar, Shahrul Azman Mohd Noah, Mohd Shahrul Nizam Mohd Danuri, and Mohammed Ali Al-Garadi. 2020. Application of machine learning methods in mental health detection: a systematic review. *Ieee Access* 8 (2020), 183952–183964.
- [2] Usman Arshad, Cecilia Mascolo, and Marcus Mellor. 2003. Exploiting mobile computing in health-care. In Proceedings of demo session of the 3rd international workshop on smart appliances, ICDCS03. Citeseer.
- [3] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy?. In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 2280–2292.
- [4] Johana Cabrera, M Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In International Work-Conference on Bioinformatics and Biomedical Engineering.

- Springer, 313-326.
- [5] Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, et al. 2022. A multi-modal open dataset for mental-disorder analysis. Scientific Data 9, 1 (2022), 178.
- [6] 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.
- [7] Jin Chen, Tony Ro, and Zhigang Zhu. 2022. Emotion recognition with audio, video, EEG, and EMG: a dataset and baseline approaches. *IEEE Access* 10 (2022), 13229–13242.
- [8] Dan Chisholm, Kim Sweeny, Peter Sheehan, Bruce Rasmussen, Filip Smit, Pim Cuijpers, and Shekhar Saxena. 2016. Scaling-up treatment of depression and anxiety: a global return on investment analysis. *The Lancet Psychiatry* 3, 5 (2016), 415–424.
- [9] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. 2020. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access* 8 (2020), 168865–168878.
- [10] Ting Dang, Dimitris Spathis, Abhirup Ghosh, and Cecilia Mascolo. 2023. Human-centred artificial intelligence for mobile health sensing: challenges and opportunities. Royal Society Open Science 10, 11 (2023), 230806.
- [11] Nan Gao, Soundariya Ananthan, Chun Yu, Yuntao Wang, and Flora D Salim. 2023. Critiquing Self-report Practices for Human Mental and Wellbeing Computing at Ubicomp. arXiv preprint arXiv:2311.15496 (2023).
- [12] Ela Gore and Sheetal Rathi. 2019. Surveying machine learning algorithms on eeg signals data for mental health assessment. In 2019 IEEE Pune Section International Conference (PuneCon). IEEE, 1–6.
- [13] Lukas Grasse, Sylvain J Boutros, and Matthew S Tata. 2021. Speech interaction to control a hands-free delivery robot for high-risk health care scenarios. Frontiers in Robotics and AI 8 (2021), 612750.
- [14] Alberto Greco, Gaetano Valenza, and Enzo Pasquale Scilingo. 2016. Advances in Electrodermal activity processing with applications for mental health. Springer.
- [15] Unsoo Ha, Yongsu Lee, Hyunki Kim, Taehwan Roh, Joonsung Bae, Changhyeon Kim, and Hoi-Jun Yoo. 2015. A wearable EEG-HEG-HRV multimodal system with simultaneous monitoring of tES for mental health management. IEEE transactions on biomedical circuits and systems 9. 6 (2015). 758-766.
- [16] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. Sensors 21, 10 (2021), 3461.
- [17] Xiyuan Hou, Yisi Liu, Olga Sourina, Yun Rui Eileen Tan, Lipo Wang, and Wolfgang Mueller-Wittig. 2015. EEG based stress monitoring. In 2015 IEEE international conference on systems, man, and cybernetics. IEEE, 3110–3115.
- [18] Xiaozhu Hu, Yanwen Huang, Bo Liu, Ruolan Wu, Yongquan Hu, Aaron J Quigley, Mingming Fan, Chun Yu, and Yuanchun Shi. 2023. SmartRecorder: An IMU-based Video Tutorial Creation by Demonstration System for Smartphone Interaction Tasks. In Proceedings of the 28th International Conference on Intelligent User Interfaces. 278–293.
- [19] Yongquan Hu, Hui-Shyong Yeo, Mingyue Yuan, Haoran Fan, Don Samitha Elvitigala, Wen Hu, and Aaron Quigley. 2023. Microcam: Leveraging smartphone microscope camera for context-aware contact surface sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 7, 3 (2023), 1–28.
- [20] Inthrani Raja Indran, Priya Paranthaman, Neelima Gupta, and Nurulhuda Mustafa. 2024. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. Medical Teacher (2024), 1–6.
- [21] Jonathan W Kim, Ahmed Alaa, and Danilo Bernardo. 2024. EEG-GPT: Exploring Capabilities of Large Language Models for EEG Classification and Interpretation. arXiv preprint arXiv:2401.18006 (2024).
- [22] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* (2024), 1–10.
- [23] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with aibased large language models. arXiv preprint arXiv:2307.11991 (2023).
- [24] Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. arXiv preprint arXiv:2303.15727 (2023).
- [25] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–22.
- [26] Zhongjie Li, Gaoyan Zhang, Jianwu Dang, Longbiao Wang, and Jianguo Wei. 2021. Multi-modal emotion recognition based on deep learning of EEG and audio signals. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–6.
- [27] Liangliang Liu, Zhihong Liu, Jing Chang, and Xue Xu. 2024. A multi-modal extraction integrated model for neuropsychiatric disorders classification. *Pattern Recognition* (2024), 110646.

- [28] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope* investigative otolaryngology 5, 1 (2020), 96–116.
- [29] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and personalization of mobile sensingbased mood inference models: an analysis of college students in eight countries. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 6, 4 (2023), 1–32.
- [30] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/ [Accessed:June 2024].
- [31] World Health Organization et al. 2022. World mental health report: Transforming mental health for all. (2022).
- [32] Tiago F Pereira, Arthur Matta, Carlos M Mayea, Frederico Pereira, Nelson Monroy, João Jorge, Tiago Rosa, Carlos E Salgado, Ana Lima, Ricardo J Machado, et al. 2022. A web-based Voice Interaction framework proposal for enhancing Information Systems user experience. Procedia Computer Science 196 (2022), 235–244.
- [33] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the time of language models. In Proceedings of the sixteenth ACM international conference on web search and data mining. 1291–1292.
- [34] Kerstin Pieper, Robert P Spang, Pablo Prietz, Sebastian Möller, Erkki Paajanen, Markus Vaalgamaa, and Jan-Niklas Voigt-Antons. 2021. Working with environmental noise and noise-cancelation: a workload assessment with EEG and subjective measures. Frontiers in neuroscience 15 (2021), 771533.
- [35] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Passive mobile sensing and psychological traits for large scale mood prediction. In Proceedings of the 13th EAI international conference on pervasive computing technologies for healthcare. 272–281.
- [36] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Large Language Models are Anonymizers. In ICLR 2024 Workshop on Reliable and Responsible Foundation Models.
- [37] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry* 10, 1 (2020), 116.
- [38] Hao Sun, Jiaqing Liu, Shurong Chai, Zhaolin Qiu, Lanfen Lin, Xinyin Huang, and Yenwei Chen. 2021. Multi-Modal Adaptive Fusion Transformer Network for the Estimation of Depression Level. Sensors 21, 14 (2021). https://doi.org/10.3390/ s21144764
- [39] Teo Susnjak, Peter Hwang, Napoleon H Reyes, Andre LC Barczak, Timothy R McIntosh, and Surangika Ranathunga. 2024. Automating research synthesis with domain-specific large language model fine-tuning. arXiv preprint arXiv:2404.08680 (2024).
- [40] Zhiyuan Wang, Maria A Larrazabal, Mark Rucker, Emma R Toner, Katharine E Daniel, Shashwat Kumar, Mehdi Boukhechba, Bethany A Teachman, and Laura E Barnes. 2023. Detecting social contexts from mobile sensing indicators in virtual interactions with socially anxious individuals. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 7, 3 (2023), 1–26.
- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35 (2022), 24824–24837.
- [42] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, et al. 2024. MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–24.
- [43] 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.
- [44] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, 1 (2024), 1–32.
- [45] Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. IEEE Transactions on Knowledge and Data Engineering (2023).
- [46] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet Digital Health 6, 1 (2024), e12–e22.