scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Multimodal Depression Consultation Dataset of Speech and Text with HAMD-17 Assessments

Pengfei Cao[1,2], Yuanzhe Zhang[3], Chenxiang Zhang[1], Wei Chen[1], Yan Liu[1], Shuang Xu[1], Miao Xu[4,5], Wenqing Jin[4,5], Jinjie Xu[4,5], Dan Wang[4,5], Wei Wang[4,5], Xue Wang[4,5], Wen Wang[4,5], Yanping Ren[4,5], Jun Zhao[1,2] ✉, Rena Li[4,5] ✉ & Kang Liu[1,2] ✉

The global surge in depression rates, notably severe in China with over 95 million affected, underscores a dire public health issue. This is exacerbated by a critical shortfall in mental health professionals, highlighting an urgent call for innovative approaches. The advancement of Artificial Intelligence (AI), particularly Large Language Models, offers a promising solution by improving mental health diagnostics. However, there is a lack of real data for reliable training and accurate evaluation of AI models. To this end, this paper presents a high-quality multimodal depression consultation dataset, namely Parallel Data of Depression Consultation and Hamilton Depression Rating Scale (PDCH). The dataset is constructed based on clinical consultations from Beijing Anding Hospital, which provides audio recording and transcribed text, as well as corresponding HAMD-17 scales annotated by professionals. The dataset contains 100 consultations and the audio exceeds 2,937 minutes. Each of them is about 30-min long with more than 150 dialogue turns. It enables to fill the gap in mental health services and benefit the creation of more accurate AI models.

Background & Summary

According to global estimates from the World Health Organization, approximately 350 million people worldwide suffer from depression, highlighting its profound impact on public health[1]. In China alone, more than 95 million individuals are impacted, with the condition significantly altering their thoughts, behaviors, emotions, and overall quality of life[2]. Compounding this issue is the acute scarcity of mental health professionals in China, where the rate of depression recognition stands at a mere 21%, and the rate of diagnosis and treatment is even lower at 10%—substantially beneath the global average of 55.65%[3,4]. The diagnosis of depression typically relies on clinical interviews, where patients complete Patient Health Questionnaire (PHQ). Their responses are then evaluated using standardized depression rating scales to facilitate a quantitative and accurate diagnosis[5]. However, insufficient investment in mental health services and a shortage of well-trained professionals often hinder effective diagnostic practices[6]. Therefore, enhancing the diagnostic process for depression is crucial to ensuring more accessible and effective clinical care[7]. Ongoing advancements in Artificial Intelligence (AI), the integration of AI into mental health diagnostics offers a promising solution[8–11]. For example, given impressive performance on various tasks of Large Language Models (LLMs)[12,13], some researchers attempt to employ LLMs for depression detection[7,14,15], where LLMs can assist in predicting PHQ scores that can be used to provide an indication of depression. By automating pre-consultation tasks and expanding access to depression diagnosis, particularly in remote and underserved regions, AI can alleviate the burden on mental health professionals. This advancement creates new opportunities to tackle complex health challenges like depression more effectively. To achieve this goal, it is essential to collect real-world medical consultation data and annotate it meticulously, enabling the reliable training and evaluation of AI models.

[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. [2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. [3]National Science Library, Chinese Academy of Sciences, Beijing, China. [4]Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China. [5]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China. ✉e-mail: jzhao@nlpr.ia.ac.cn; renali@ccmu.edu.cn; kliu@nlpr.ia.ac.cn

In recent years, several datasets have been developed for AI research in depression detection. Based on their data sources, these datasets can be broadly categorized into two types: 1) *interview-based datasets*, which are constructed from participant interview data, and 2) *social media-based datasets*, which are derived from user-generated content on social media platforms. For example, for interview-based datasets, DAIC-WOZ[16] contains 189 clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The data of the corpus is from Wizard-of-Oz interviews, which is conducted between participants and an animated virtual interviewer controlled by a human interviewer in another room. The dataset includes audio, video and transcribed text, along with clinical annotations like PHQ-8 scores. The BlackDog Dataset[17] includes data from 60 subjects, split equally between severely depressed patients and healthy controls. The data was gathered through open-ended interviews where participants described emotionally significant life events, capturing nonverbal behaviors such as facial expressions and body language. Depression severity was measured using the QIDS-SR (Quick Inventory of Depressive Symptomatology-Self Report)[18]. The dataset is primarily used for binary classification tasks, distinguishing between severely depressed individuals and healthy controls. The MODMA (Multi-modal Open Dataset for Mental-disorder Analysis)[19] dataset is designed for the study of mental disorders, particularly depression. It includes both EEG (electroencephalogram) and audio data from clinically diagnosed patients and healthy controls. The 52 audio recordings are gathered during interviews, reading tasks, and picture descriptions. MODMA also includes the PHQ-9 total score. Audio-Visual Emotion Challenge (AVEC)[20,21] is a competitive event established with the objective of evaluating multimedia processing and machine learning techniques for auomated analysis of emotions in audio, visual, and audio-visual media. The dataset uses the Beck Depression Index (BDI)[22], a self-reported multiple choice inventory. For social media-based datasets, Shen *et al.*[23] construct two datasets of depression and non-depression users from Twitter. They collect the profile information of Twitter user and an anchor tweets to infer the mental state. For the depression dataset, the users are labeled as depressed if their anchor tweets satisfied the strict pattern. The dataset contains 1,402 depressed users and 292,564 tweets. For the non-depression dataset, the users are labeled as non-depressed if they had never posted any tweet containing the character string "depress". It contains more than 300 million active users and 10 billion tweets. The Weibo User Depression Detection Dataset (WU3D)[24] includes more than 10,000 depressed users and more than 20,000 normal users, each of which contains enriched information fields, including tweets, the posting time, posted pictures, the user gender, etc. The DepressionEmo dataset[25] is designed to detect 8 emotions associated with depression by 6,037 examples of long Reddit user posts.

Although these datasets have promoted the development of using AI technology for automatically detect depression to some degree, these datasets are not from real clinical consultation scenarios. They lack authenticity and are only simulations of depression diagnosis scenarios. The AI model trained using such datasets is difficult to ensure the reliability of diagnosis. Therefore, the real-world data is urgently needed to train models and accurately evaluate their effectiveness. In real diagnostic scenarios, doctors and patients usually have face-to-face communication and interaction. By asking some questions, doctors can determine the patient's detailed information on every aspect of depression, leading to the final diagnosis. Thus, in order to better promote automatic detection of depression, it is necessary to construct high-quality datasets according to real clinical consultation. By leveraging such data, researchers can tailor algorithms to better understand and predict the nuances of depression symptoms across varied populations, thereby enhancing the precision of initial screenings.

To alleviate the above limitations, this paper proposes a multimodal depression consultation dataset called **P**arallel Data of **D**epression **C**onsultation and **H**amilton Depression Rating Scale (PDCH). The dataset is constructed based on the real clinical consultation data. The valuable consultation data comes from Beijing Anding Hospital, a leading mental health hospital of China, thus the quality of consultations is guaranteed. The PDCH dataset encompasses paired data from patient consultations and their corresponding evaluations using the Hamilton Depression Rating Scale (HAMD-17)[26]. Collecting such detailed consultation data presents a significant challenge, necessitating capturing the nuanced interactions between patients and healthcare professionals. These interactions are then carefully transcribed into textual form, ensuring that the richness and nuance of the conversations are retained. During the consultation process, the patient's emotions are also annotated. Then, each consultation is assessed by doctors with HAMD-17. In this way, PDCH provides a direct alignment between consultation and HAMD-17. The dataset contains 100 consultations and the audio length exceeds 2,937 minutes. Each of them is about 30-min long with more than 150 dialogue turns. We analyze the characteristics of this dataset and validate the effectiveness of LLMs for depression detection on this dataset. By introducing this novel dataset into the research community, we aim to provide a foundational resource that supports the development of more nuanced and accurate AI models. This can lead to more personalized and effective treatment plans, bridging the gap in mental health services in China.

Methods

Raw Records Collection. Figure 1 illustrates the systematic process of dataset construction in this paper. The consultation data is collected from Beijing Anding Hospital, a leading mental health institution in China, which guarantees the professional quality of the clinical consultations. The research protocol has been formally approved by the Ethics Commission of Beijing Anding Hospital (Approval Number: 2021-research-102). Prior to data collection, written informed consent is obtained from all participants. To ensure the comprehensiveness and reliability of the dataset, we recruit a panel of 10 senior psychiatrists with diverse clinical expertise. These doctors conduct standardized face-to-face consultations with 100 inpatients diagnosed with depressive episodes in the Department of Psychiatry. The age range of the participating patients is from 18 to 65 years old, and there are no patients under 18 years old. During these clinical interactions, psychiatrists systematically evaluate patients' conditions based on their professional expertise and clinical experience. All consultations are audio-recorded with strict adherence to patient confidentiality protocols. The recorded data captures essential clinical interactions

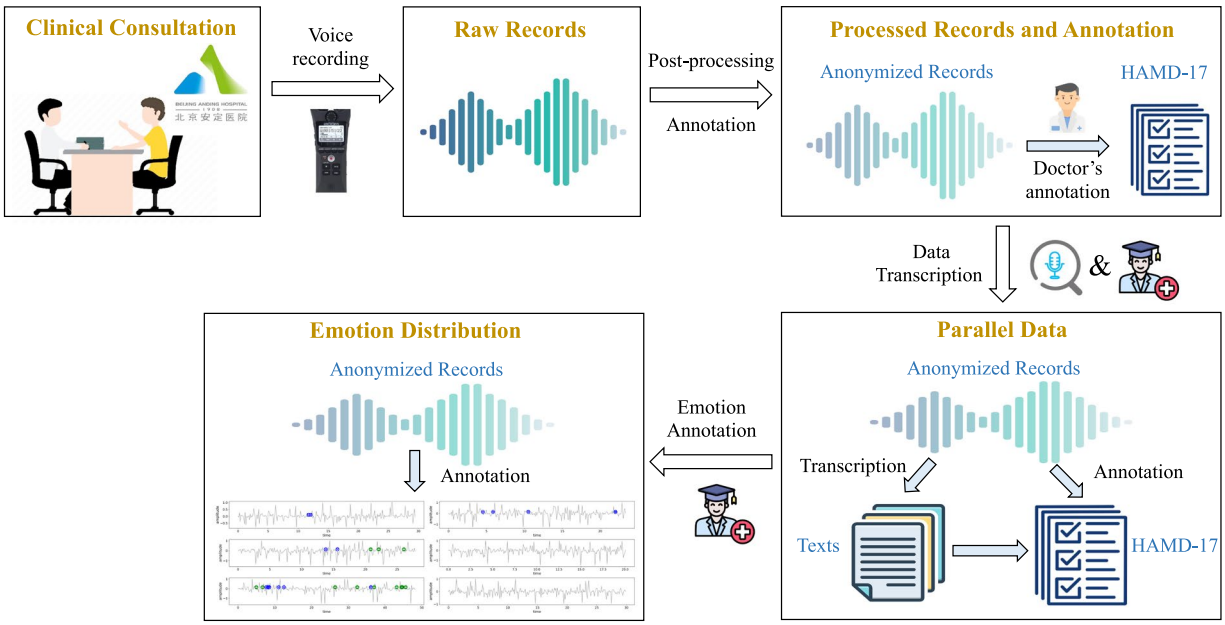

Fig. 1 The construction process of the proposed PDCH dataset.

Transcribed Text

Timestamps	Role	Dialogue Content
……		
01:34-01:40	Doctor	然后你就谈谈那个，当时的情绪状态。 Translation: Then you just talk about that, the emotional state at that time.
01:41-02:13	Patient	就是我很非常沮丧，然后，呃，我对以前感兴趣的事情都不感兴趣了，然后我也不太开心，主要是抑郁和悲伤，然后，呃，我更加悲伤一点。 Translation: It's just that I'm very, very depressed and then, uh, I'm not interested in anything that I used to be interested in, and then I'm not very happy, and I'm mainly depressed and sad and then, uh, I'm sad a little bit more.
02:13-02:25	Patient	然后我自己的行为也开始失控，拖延症越来越严重。 Translation: And then kind of my own behavior started to get out of control, and then the procrastination got worse and worse.
02:26-02:46	Patient	我喜欢熬夜，但可能是由失眠引起的。嗯，我玩手机直到感觉困了才睡觉，所以我逐渐从一两点变成三四点。 Translation: I like to stay up late, but it may be caused by insomnia. Well, I play with my mobile phone until feel sleepy before going to bed, so I will gradually change from one or two o'clock to three or four o'clock.
02:46-02:49	Doctor	是不是越来越晚了？ Translation: Is it getting late?
02:49-02:51	Patient	是的。 Translation: That's right.
……		

HAMD-17

Factor	Score
Depressed Mood	2
Feelings of Guilt	2
Suicide	1
Insomnia: Early in the Night	2
Insomnia: Middle of the Night	1
Insomnia: Early Hours of the Morning	2
Work and Activities	2
Retardation	1
Agitation	1
Anxiety Psychic	2
Anxiety Somatic	3
Loss of Weight	0
Genital Symptoms	2
……	

Fig. 2 An example of the transcribed text and its corresponding HAMD-17 assessment.

and therapeutic communications, providing valuable insights into the dynamics of psychiatric consultations and patient-clinician interactions.

The recording device used in this study is specifically designed for high-fidelity audio capture in clinical settings. It is engineered to account for the subtleties of human speech and the ambient acoustic environment of psychiatric consultations, ensuring the clarity and integrity of the recorded conversations. To safeguard patient privacy, the devices are offline and incapable of connecting to the internet. Additionally, all audio recordings undergo rigorous post-processing to anonymize sensitive information. Following each consultation, the participating doctors are asked to complete the Hamilton Depression Rating Scale (HAMD-17) for the corresponding patient. Each dimension in the scale (totally 17 dimensions) is scored based on the doctor's clinical judgment and observation of the patient's symptoms, resulting in a cumulative score that reflects the severity of the patient's depression. The HAMD-17 assessments are a critical component of the dataset, as they provide a standardized metric that can be correlated with the transcribed consultation data.

Datasets	Language	Modality	#Inst.	#Persons	Length	Scale	Data Source
AVEC2013[20]	German	video, audio	150	292	—	BDI-II	human-computer interaction
AVEC2014[21]	German	video, audio	300	84	274 min	BDI-II	human-computer interaction
DAIC-WoZ[16]	English	video, audio, text	189	193	2,756 min	PHQ-8	human-computer interaction
E-DAIC[35]	English	video, audio, text	275	351	4,282 min	PHQ-8	human-computer interaction
BlackDog[17]	English	video, audio, text	60	60	—	DSM-IV	answering open-ended questions
Mundt[36]	English	audio	35	35	—	HAMD-17, QIDS	automated telephone interface
MODMA[19]	Chinese	EEG, audio	53	53	431 min	PHQ-9	real-world clinical consultation
DepressionEmo[25]	English	text	6,037		—	—	Reddit posts
WU3D[24]	Chinese	text	—	30,000	—	—	Weibo posts
PDCH (Ours)	Chinese	audio, text	100	100	2,937 min	HAMD-17	real-world clinical consultation

Table 1. The comparison of existing depression detection datasets. "#Inst." and "#Persons" denote the number of instances and participants, respectively. "Length" represents the length of all audio records.

Data Transcription. The transcription of audio recordings into textual data constitutes a critical phase in our research methodology, employing a rigorous three-stage annotation protocol to ensure data accuracy, consistency, and compliance with ethical standards for anonymization. Our systematic approach is implemented as follows:

- **Pre-annotation Phase**: 1) The recordings are partitioned into coherent segments based on mute detection, each of which is less than 2 minutes long, 2) The open-source end-to-end speech recognition toolkit WeNet[27] is utilized to transcribe each audio segment into text, 3) The transcribed segments belonging to the same recorded interview are reassembled into complete pre-annotated transcripts.
- **Manual Verification Phase**: 1) Each recording and pre-annotated text is carefully proofread and transcribed by two trained medical students who are well-versed in the nuances of psychiatric consultations, marking speaker identities and correcting transcription errors, 2) Discrepancies between annotators are resolved through discussions, ensuring clinical accuracy and terminological consistency.
- **Temporal Alignment Phase**: 1) The two medical students annotate the start and end timestamps of the annotated text, ensuring complete alignment between the transcribed text and audio, 2) When they have different opinions, they discuss and reach a consensus.

This process is conducted on network-restricted devices, with strict prohibitions on the use of external storage media to further ensure data security. Additionally, all personally identifiable information was meticulously desensitized during transcription, thus it allows the dataset to be used broadly for research purposes while adhering to ethical standards and privacy regulations. After data transcription, we align the consultation text with the HAMD-17 assesment. An example is shown in Fig. 2. This alignment is essential for developing and refining AI models that aim to recognize and assess depressive symptoms through natural language processing. The dataset contains 100 consultations, and the total length of the audio exceeds 2,937 minutes. Each audio is about 30-min long with more than 150 dialogue turns. The comparison between our dataset with existing datasets are shown in Table 1. Compared with existing datasets, our dataset PDCH exhibits several distinctive advantages over existing resources:

- **Clinical Authenticity**: Collected from real-world psychiatric consultations in a tertiary hospital setting, providing real valid data.
- **Comprehensive Annotation**: Each consultation is annotated by certified psychiatrists using the standardized HAMD-17 scale, enabling fine-grained symptom analysis.
- **Extended Duration**: Longer consultation sessions compared to existing datasets, capturing more complete clinical interactions.
- **Expert Validation**: All transcripts are verified by medical professionals, ensuring clinical accuracy and terminological precision.

The integration of HAMD-17 assessments, recognized as the important standard for depression severity measurement, provides clinically validated ground truth for model development. These characteristics make PDCH particularly valuable for advancing research in AI-assisted depression detection and severity assessment through natural language analysis.

Emotion Annotation. Compared to textual transcripts, audio recordings preserve crucial paralinguistic features, including prosody, speech rate, and vocal intensity, which serve as important indicators of patients' emotional states and their dynamic changes throughout the consultation. To systematically investigate emotional expression patterns in clinical interactions, we conduct a fine-grained analysis of emotional states and their temporal dynamics using audio features and fluctuations of the recordings. Specifically, we randomly sample nearly half of the consultations from the 100 consultations, and employ the two medical students to annotate the nuanced emotional states. The categories for audio emotion labeling are based on GoEmotions[28] that provides a comprehensive taxonomy for fine-grained emotion classification. During the annotation process, these two

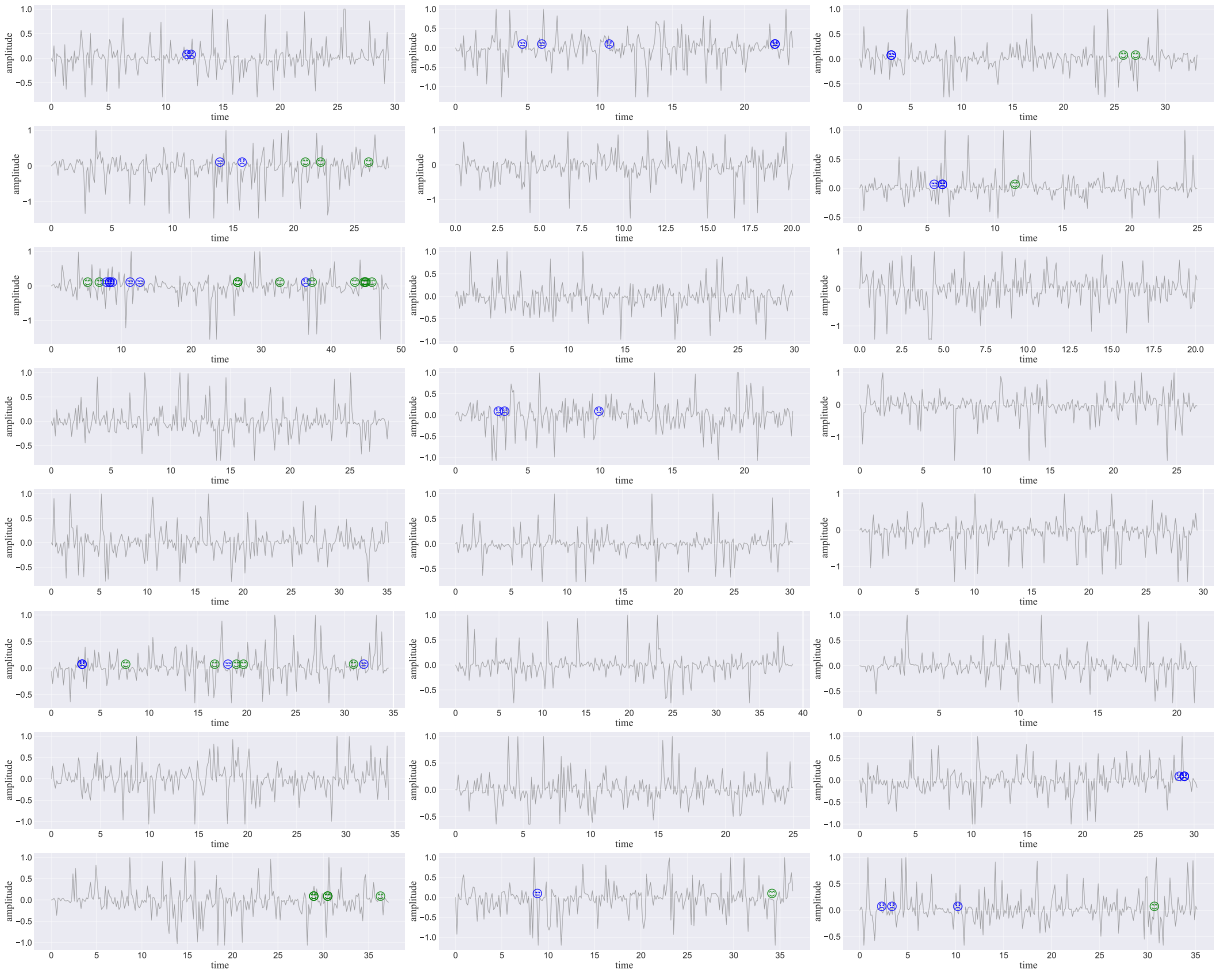

Fig. 3 The audio waves annotated with patient emotions. Green expressions correspond to the presence of positive emotions, while blue expressions reflect the occurrence of sadness or depressive tendencies.

medical students independently annotate the voice emotions of each speech segment. When they have different opinions, they discuss and reach a consensus. We show 24 annotated instances in Fig. 3, which reveals a diverse range of emotional characteristics. Most patients maintain relatively stable emotions throughout their interactions with doctors. With the doctor's active guidance and communication, some patients gradually experience emotional improvement, displaying positive emotional changes. However, some patients remain in a low emotional state even after communication, while others experience frequent emotional fluctuations between excitement and depression. In our annotation, the number of positive labels is 19 and the number of negative labels is 37. Such annotation is a new perspective on showing a patient's emotional changes during the consultation. It can enhance AI models for depression detection by providing fine-grained, clinically validated features that capture nuanced emotional states and dynamic changes in patient interactions.

Data Records

The PDCH dataset[29] is accessible via the Science Data Bank repository at https://doi.org/10.57760/sciencedb.27818. The dataset is organized into a directory called `depression_instances`, comprising one primary file and one main directory. The file, named `HAMD_annotation.xlsx`, contains detailed HAMD-17 assessment results annotated by clinical experts, including individual item scores and the corresponding total scores for each patient. The main directory, titled `wav_data`, encompasses 100 subdirectories, each representing a complete consultation session for a unique patient. Within each subdirectory, the data is organized into four distinct file types:

- `X.wav`: The anonymized audio recording of the consultation session that is tone converted through a speech processing tool to ensure patient confidentiality.
- `X.txt`: The pre-annotated transcript generated by an automated transcription model, where each conversational turn is prefixed with the speaker's role (e.g., doctor or patient).
- `X_correction.txt`: The manually corrected version of the pre-annotated transcript, verified and refined by two trained medical students to ensure accuracy and consistency.

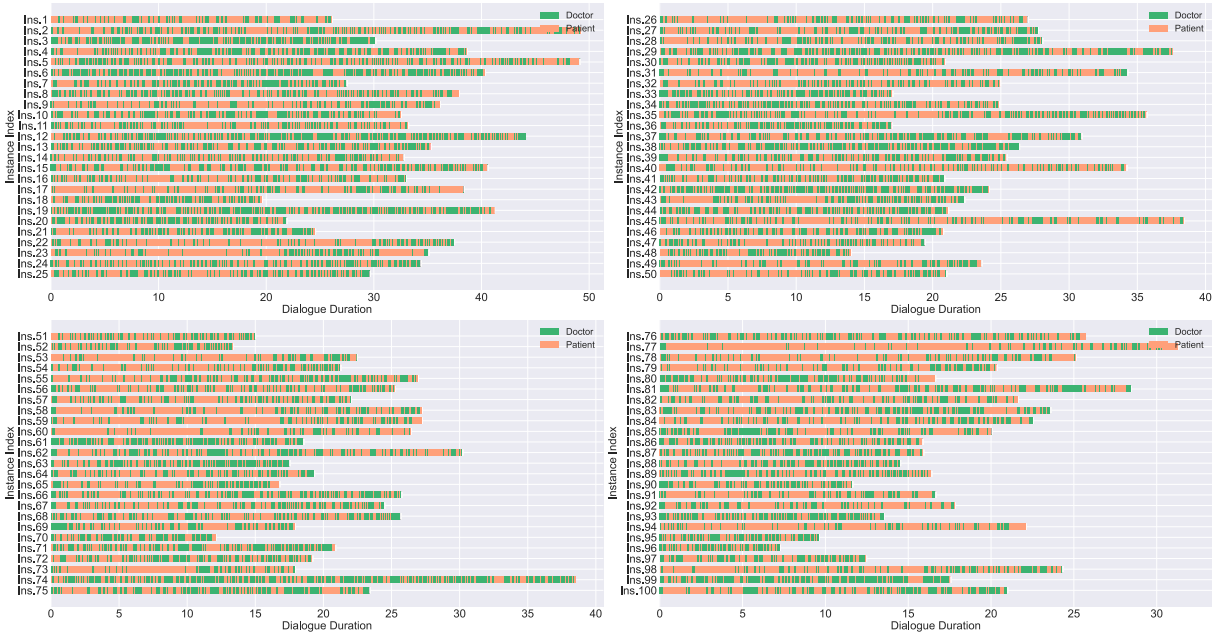

Fig. 4 The distribution of communication time between doctors and patients, with the green section representing the doctor's clinical consultation and the orange section representing the patient's feedback. "Ins.X" denotes the instance index.

- `X_correction_timestamp_emotion.txt`: The timestamp-annotated file, derived from the corrected transcript, which specifies the start and end times of each transcribed segment in the corresponding audio recording.

Due to the technical limitations of the recording device, which stores individual recordings at 25 minutes, each subdirectory may contain multiple instances of the aforementioned file types, sequentially labeled (e.g., `1.wav`, `2.wav`, etc.). This structured organization ensures both the accessibility and integrity of the dataset for research purposes.

Technical Validation

To comprehensively characterize the annotated dataset, we systematically analyze its unique properties across three key dimensions: audio recording features, textual characteristics, and fine-grained annotation. Furthermore, we empirically validate the utility of the dataset through extensive experiments, demonstrating its effectiveness in both training and evaluating AI models for depression detection and analysis.

Recording Feature Analysis. Figure 4 presents the distribution of the audio duration between doctors and patients across 100 consultations in the dataset, providing insights into the interaction dynamics captured in the audio recordings. From the results, we have the following observations:

1) In real scenarios, the duration of speech between doctors and patients is very unbalanced. In detail, for some cases, the patient spends most of the time speaking, and sometimes it is the other way around. For example in instance 3, the doctor's speaking time is significantly longer than that of the patient. The reason may be that the patient is relatively quiet and introverted, and the doctor is constantly guiding him or her. It also causes that the duration of this sample is also longer than other samples. Meanwhile, this imbalance phenomenon may also have an impact on the predictions of AI models. Based on the proposed dataset, AI models can be evaluated more reliably.
2) The majority of consultations are less than 40 minutes in duration. Experienced doctors effectively capture the psychological state and symptoms of patients with depression, then conduct detailed diagnoses and treatment processes. Both overt and subtle emotions and symptoms expressed by patients during these interactions are recorded. The frequent interactions between patients and doctors indicate a well-established communication process. Doctors interact with patients based on their individual situations and characteristics, thus the distribution of each sample is different, which also indicates that the proposed dataset is more realistic.

These findings underscore the dataset's capacity to support the development of AI models that can accommodate the inherent variability and complexity of real-world psychiatric consultations.

Text Feature Analysis. This section presents a quantitative analysis of the transcribed consultation texts, focusing on turn-level characteristics. Figure 5 illustrates the distribution of utterance lengths (measured in the

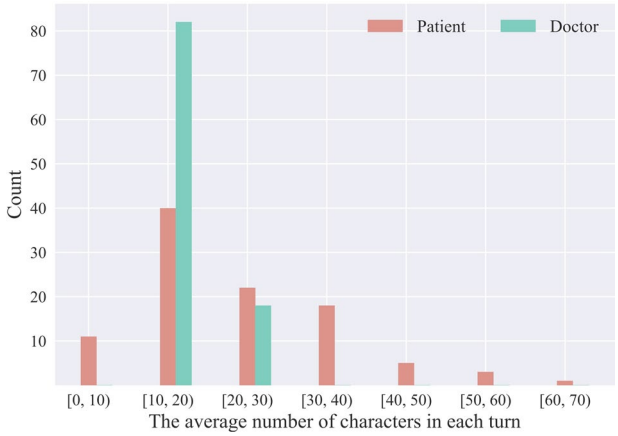

Fig. 5 The statistical result of the average length of transcribed texts per-round for each instance. The x-axis represents the length range, and the y-axis represents the corresponding quantity.

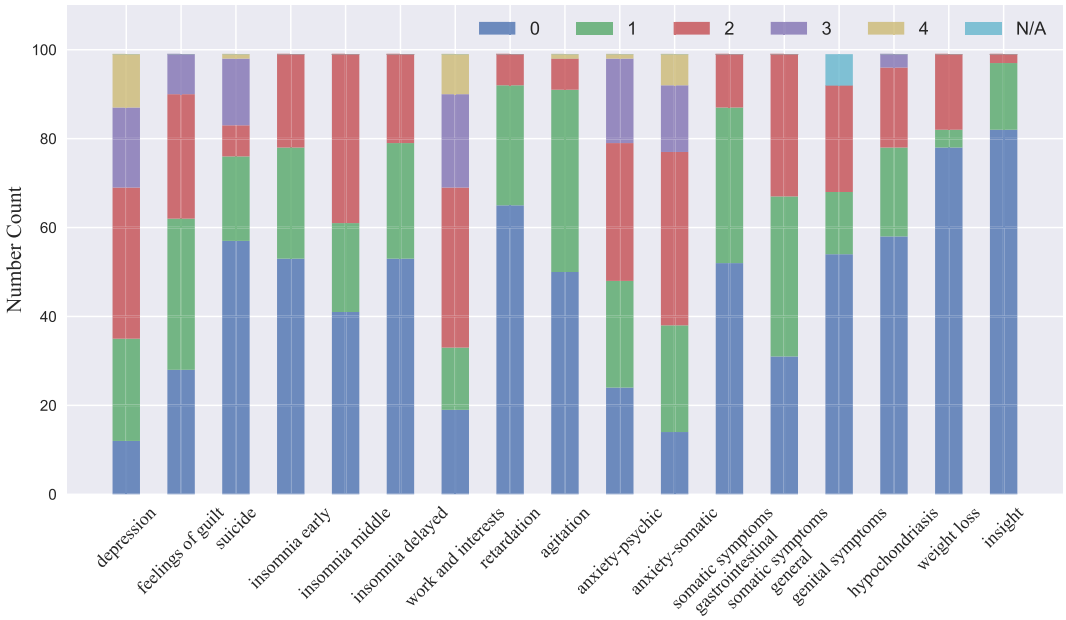

Fig. 6 The statistical information of the annotation results for each item on the HAMD-17 scale.

number of characters) for both doctors and patients across all consultations. From the figure, we can observe the following phenomena:
1) The transcribed texts of doctors are relatively brief and concentrated, typically ranging from 10 to 30 characters. This suggests that doctors primarily use brief questions and feedback to guide patients, empowering them to describe symptoms and express emotions more fully.
2) In contrast, the length of transcribed texts of patients is relatively varied, ranging from 0 to 70 characters. This suggests that different patients have different styles of dealing with doctor inquiries. From these statistical characteristics, we can further infer which patients are more willing to share their status and emotions, and which patients are more reticent. Compared to the results in Figs. 4, 5 can provide the extra statistical information about the imbalance in conversation length between patients and doctors.

Annotation Results Analysis. To provide a detailed characterization of our dataset, we conduct a fine-grained statistical analysis of each item in the HAMD-17 scale, as visualized in Fig. 6. Male and female patients account for 48% and 52% respectively. Among these 100 participating patients, there were 13 normal patients, 27 mild depression patients, 13 moderate depression patients, 19 severe depression patients, and 37 very severe depression patients, respectively. Additionally, one patient was unable to evaluate HAMD-17 due to the short duration of symptoms. The annotation scheme follows the standard HAMD-17 scoring system: 1) Scores 0-4 represent increasing symptom severity (0: none, 1: mild, 2: moderate, 3: severe, 4: extremely severe), 2) Score 9 indicates either uncertainty or item inapplicability.

Models	Modality	Precision	Recall	F1
GPT4o-mini-audio-preview	audio	0.383	0.375	0.379
	text	0.405	0.400	0.403
	text+audio	**0.412**	**0.403**	**0.407**
Qwen2.5-Omni-7B	audio	0.383	0.385	0.384
	text	0.428	0.430	0.429
	text+audio	**0.431**	**0.433**	**0.432**
Qwen2-Audio-7B-Instruct	audio	0.114	0.107	0.111
	text	**0.149**	**0.137**	**0.143**
	text+audio	0.130	0.122	0.126

Table 2. The results of some representative LLMs on the proposed dataset PDCH. The bold denotes the best performance.

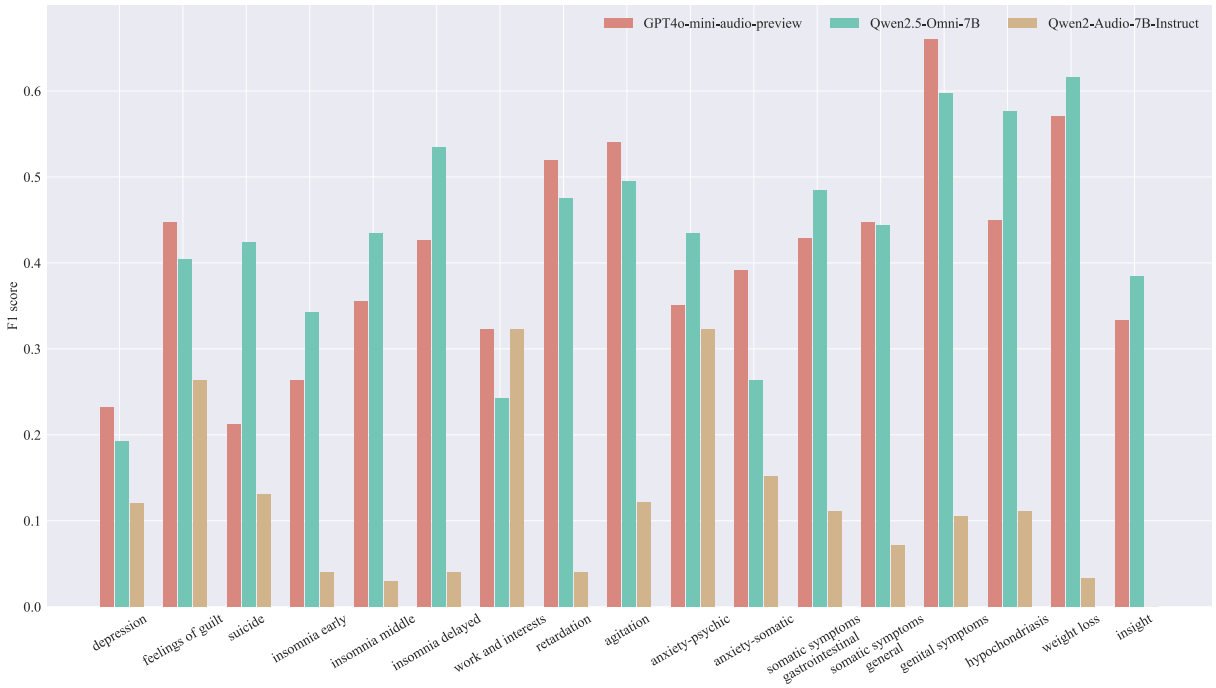

Fig. 7 The predictive results of LLMs for each item in the HAMD-17 scale.

From the figure, we can observe that: 1) Different HAMD-17 items exhibit distinct score distributions. For instance, the "depression" item spans all severity levels, while "insight" is limited to only three categories (none, mild, and moderate). 2) For "depression", "feelings of guilt", "work and interests", "anxiety-psychic" and "anxiety-somatic", the proportion of symptomatic patients in these items is high. However, for the "insight", there are more people without this symptom. Therefore, our data annotation can reflect the frequency of each item in the HAMD-17 scale.

Large Language Models for Scale Completion. Despite significant advancements in the application of AI to mental health diagnosis, there remains a critical gap in the authoritative evaluation of these technologies using real-world clinical data. To address this limitation, we develop the PDCH dataset, which enables a systematic assessment of LLMs' capabilities in predicting depression. The concordance between model's outputs and expert assessments provided an accurate measure of the LLM's efficacy, thereby validating its potential as a reliable tool in mental health diagnostics.

Experimental Settings. We evaluate three state-of-the-art multimodal large language models with text and speech processing capabilities: GPT4o-mini-audio-preview, Qwen2.5-Omni-7B[30], and Qwen2-Audio-7B-Instruct[31]. These models are specifically selected for their robust Chinese language support, a critical requirement for our study. We directly evaluate these three models on all instances without fine-tuning the parameters. It is important to design prompts for facilitating the HAMD-17 scale completion process. Prompts serve as instructions that guide LLMs in generating specific responses or performing targeted tasks. The designed and curated prompts for GPT4o-mini-audio-preview are as shown in Fig. 8. Given the discrete scoring system of

Prompt

System Instruction

你是一个尽职的助手，请依据医患访谈对话来分析出任务要求的目标因子分数。<HAMD17量表描述>。
Translation: You are a diligent assistant, please analyze the target factor scores based on the doctor-patient interview dialogue. <HAMD17 Scale Description>.

Task Requirements

请基于访谈对话片段，给出关注因子结果的分数，若对话中没有提到这个因子，则输出分数为“None”。
输出格式：id 因子名：分数为 (score)，以“;”分隔。
Translation: Please provide the scores for the focus factors based on the interview dialogue snippets. If the factor is not mentioned in the dialogue, the output score should be “None”.
Output format: id Factor name: score is (score), separated by “;”.

Examples

1 抑郁情绪因子：分数为(x)；2 有罪感因子：分数为 (y)；..... 16 体重减轻因子：分数为(None)；17 自知力因子：分数为 (z)。
Translation: 1 Depressed mood factor: score is (x); 2 Guilt factor: score is (y); ..... ; 16 Weight loss factor: score is (None); 17 Insight factor: score is (z).

Fig. 8 A prompt example used in the GPT4o-mini-audio-preview model for completing the HAMD-17 scale.

Models	Input	Datasets	
		MODMA	PDCH (Ours)
GPT4o-mini-audio-preview	audio	**0.423**	0.303
Qwen2.5-Omni-7B	audio	**0.654**	0.364
Qwen2-Audio-7B-Instruct	audio	**0.346**	0.061

Table 3. The results of some representative LLMs on the MODMA dataset and our proposed dataset PDCH.

the HAMD-17 scale (ranging from 0 to 4 points per item), we employ standard classification metrics, including precision, recall, and F1 scores, to quantitatively evaluate model performance. This evaluation framework allows for direct comparison between the models' outputs and expert clinical assessments.

Results and Analysis. The experimental results, presented in Table 2 and Fig. 7, reveal several important findings regarding LLM performance in depression detection: 1) Our analysis demonstrates a significant discrepancy between LLM outputs and clinicians' assessments. Even the most advanced model in our study, GPT4o-mini-audio-preview, achieves only modest performance (F1=0.407) in multimodal scenarios combining text and audio inputs. This substantial performance gap underscores the inherent challenges in developing AI systems for accurate depression detection. Due to being collected from real medical consultation scenarios, this dataset can provide a reliable benchmark for AI model evaluation. 2) The performance of the model using text as input is better than using audio as input. For example, the GPT4o-mini-audio-preview model demonstrates a 6.3% improvement in F1 score (from 0.379 to 0.403) when using transcribed text versus raw audio input. This performance gap likely stems from current LLMs' more mature text processing capabilities compared to their audio analysis functionalities. The significant performance improvement further underscores the value of data transcription. Notably, while text inputs show superior performance to audio alone, the combination of both modalities achieves optimal results (The Qwen2-Audio-7B-Instruct model is excluded because it processes shorter audio lengths.). This multimodal advantage suggests that audio signals contain complementary paralinguistic information, including prosodic features, emotional tone, and speech patterns, that enhances the model's predictive capability beyond text-based analysis alone. 3) From Fig. 7, we can observe that the performance of LLMs varies greatly in each dimension of the HAMD-17 scale. Besides, the GPT4o-mini-audio-preview and Qwen2.5-Omni-7B achieve better predictive results in most dimensions of the scale.

The Performance Comparison of Large Language Models on Different Datasets. To further demonstrate the quality of the dataset, we compare the performance of LLMs on our dataset and another real-world clinical dataset MODMA. Since the MODMA is also a Chinese dataset and collected from real-world clinical consultation, we choose it as the comparison object.

Models	Imbalance	Training	Precision	Recall	F1
Qwen2.5-7B-Instruct	Large	No	0.427	0.430	0.429
	Large	Yes	**0.609**	**0.613**	**0.611**
	Small	No	0.387	0.388	0.387
	Small	Yes	**0.543**	**0.544**	**0.543**
LLaMA3.1-8B-Instruct	Large	No	0.448	0.451	0.449
	Large	Yes	**0.584**	**0.588**	**0.586**
	Small	No	0.396	0.397	0.397
	Small	Yes	**0.571**	**0.571**	**0.571**

Table 4. The performance of LLMs under imbalanced length of doctor-patient conversations. "Imbalance" represents the ratio of the length of the patient's transcribed text to that of the doctor's transcribed text. "Training" indicates whether to fine-tune the parameters of LLMs.

Models	Emotion	Precision	Recall	F1
Qwen2.5-7B-Instruct	wo. emotion	0.407	0.409	0.408
	w. emotion	**0.410**	**0.412**	**0.411**
LLaMA3.1-8B-Instruct	wo. emotion	0.422	0.424	0.423
	w. emotion	**0.425**	**0.427**	**0.426**

Table 5. The results of adding and ablating patient emotions for LLM prediction. "wo." and "w." denote "without" and "with", respectively.

Experimental Settings. The MODMA dataset does not contain transcribed text for audio, thus in order to make a fair comparison with our dataset, we only input the audio into LLMs for prediction. In addition, due to the use of different scales for these two datasets (PHQ-9 for MODMA, HAMD-17 for PDCH), we employ a doctor to map the results of HAMD-17 to those of PHQ-9.

Results and Analysis. The experimental results are shown in Table 3. From the results, we can observe that LLMs achieve better performance on the dataset MODMA than on our dataset PDCH. This indicates that our dataset is more challenging for depression detection of LLMs. The reason is that the average audio length of each sample in our dataset is relatively long. Specifically, the average audio length of each sample in our dataset is 29.37 minutes, while the length of the MODMA dataset is only 8.29 minutes. Therefore, our dataset can reflect more authentic medical consultations and provide more reliable data for training and evaluating AI models.

The Impact of Imbalance of Conversions Length. In clinical practice, the length of doctor-patient conversations is uneven. To systematically examine how this imbalance affects LLM performance, we conduct a dedicated investigation.

Experimental Settings. Based on the ratio of the length of the patient's transcribed text to that of the doctor's transcribed text, we divide the dataset into two subsets (Each subset contains 50 samples), namely *Small* and *Large*. For each subset, we divide the data into five parts and use the leave-one-out cross validation to evaluate the model. In this experiment, we adopt two settings: 1) *Training*, which fine-tunes the parameters of LLMs, and 2) *No_training*, which prompts LLMs to make predictions without fine-tuning. We conduct the experiments on two advanced LLMs, including Qwen2.5-7B-Instruct[32] and LLaMA3.1-8B-Instruct[33]. For the fine-tuning condition, the LoRA[34] is leveraged to update the parameters of LLMs, which is a compute-efficient technique that freezes the model weights and injects trainable rank decomposition matrices into each layer of the model:

$$\boldsymbol{h} = \boldsymbol{W}_0\boldsymbol{x} + \Delta\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}_0\boldsymbol{x} + \boldsymbol{B}\boldsymbol{A}\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{W}_0$ denote the pretrained weights that do not perform gradient updates. $\Delta\boldsymbol{W}$ is an additional low rank weight matrix. The $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ are trainable weights, and the rank $r \ll \min(d, k)$. $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{h} \in \mathbb{R}^d$ represent the input and output vectors, respectively.

Results and Analysis. Table 4 presents the experimental results. From the table, we can observe that: 1) Our experiments demonstrate that conversation length asymmetry substantially affects model predictions. In the zero-shot (No_Training) setting, Qwen2.5-7B-Instruct shows markedly different performance between conditions, achieving F1 scores of 0.429 (Large) versus 0.387 (Small). This performance discrepancy suggests that LLMs are particularly sensitive to the relative proportion of patient versus doctor speech. The superior performance in the Large condition (where patients speak more) likely stems from richer patient self-reports containing more clinically relevant information about depressive symptoms. 2) Parameter adaptation through LoRA fine-tuning effectively improves model performance across both subsets. For instance, LLaMA3.1-8B-Instruct

shows consistent improvements after fine-tuning, with 30.5% (Large) and 43.8% (Small) improvement in F1 score. These results highlight that utilizing our proposed dataset can help enhance the performance of depression diagnosis forLLMs, indicating the high-quality of the dataset.

The Impact of Patient Emotions on LLM Prediction. In clinical practice, patients' emotional expressions serve as crucial diagnostic indicators for mental health assessment. Recognizing this clinical value, we have systematically annotate emotional states in the original consultation recordings. These annotations provide complementary affective features that transcend the information contained in pure textual transcripts. Building upon this enriched dataset, we present a comprehensive investigation into how patient emotional states influence the predictive performance of LLMs in depression detection.

Experimental Settings. The ablation study is conducted as follows: 1) "*wo. emotion*", which denotes that the LLM only makes predictions based on transcribed text. 2) "*w. emotion*", which denotes that the LLM combines transcribed text and annotated emotional information to make diagnostic predictions. We also conduct the experiments on two competitive LLMs, i.e., Qwen2.5-7B-Instruct and LlaMA3.1-8B-Instruct. To isolate the effect of emotional features from other variables, all experiments are performed in a zero-shot prompting setting without parameter fine-tuning.

Results and Analysis. Table 5 shows the experimental results. From the table, we can observe that both evaluated models showed measurable improvements when augmented with emotional information. Specifically, Qwen2.5-7B-Instruct achieves an F1 score improvement of 0.7% (from 0.408 to 0.411), while LLaMA3.1-8B-Instruct also showes a 0.7% increase (from 0.423 to 0.426). These consistent gains across architectures suggest that emotional states provide complementary predictive signals beyond textual content alone. This finding aligns with clinical practice where affective states are known to be crucial diagnostic indicators for depression.

Code availability

In line with the philosophy of reproducible research, all codes used in this paper, including those for data preprocessing and technical validation, are accessible at https://github.com/Miraclemarvel55/PDCH. The usage instructions and parameter settings for all codes can be found at this link.

Received: 22 April 2025; Accepted: 14 August 2025;
Published online: 29 September 2025

References

1. Mossie, A., Kindu, D. & Negash, A. Prevalence and severity of depression and its association with substance use in jimma town, southwest ethiopia. *Depression research and treatment* **2016**, 3460462 (2016).
2. Caiping, L. *et al*. Number and characteristics of medical professionals working in chinese mental health facilities. *Shanghai Archives of Psychiatry* **25**, 277 (2013).
3. Hou, Z., Jiang, W., Yin, Y., Zhang, Z. & Yuan, Y. The current situation on major depressive disorder in china: research on mechanisms and clinical practice. *Neuroscience bulletin* **32**, 389–397 (2016).
4. Organization, W. H. *et al*. Depression and other common mental disorders: global health estimates. (2017).
5. Kroenke, K. *et al*. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders* **114**, 163–173 (2009).
6. Evans-Lacko, S. *et al*. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological medicine* **48**, 1560–1571 (2018).
7. Sadeghi, M. *et al*. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research* **3**, 66 (2024).
8. Omar Sr, M. *et al*. Applications of large language models in psychiatry: A systematic review. *medRxiv* 2024–03 (2024).
9. Lawrence, H. R. *et al*. The opportunities and risks of large language models in mental health. *JMIR Mental Health* **11**, e59479 (2024).
10. Olawade, D. B. *et al*. Enhancing mental health with artificial intelligence: Current trends and future prospects. *Journal of medicine, surgery, and public health* 100099 (2024).
11. Perlis, R. H., Goldberg, J. F., Ostacher, M. J. & Schneck, C. D. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology* **49**, 1412–1416 (2024).
12. Zhao, W. X. *et al*. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
13. Minaee, S. *et al*. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).
14. Omar, M. & Levkovich, I. Exploring the efficacy and potential of large language models for depression: A systematic review. *Journal of Affective Disorders* (2024).
15. Lan, X., Cheng, Y., Sheng, L., Gao, C. & Li, Y. Depression detection on social media with large language models. *arXiv preprint arXiv:2403.10750* (2024).
16. Gratch, J. *et al*. The distress analysis interview corpus of human and computer interviews. In *LREC*, 3123–3128 (Reykjavik, 2014).
17. Alghowinem, S. *et al*. From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS* (2012).
18. Rush, A. J. *et al*. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry* **54**, 573–583 (2003).
19. Cai, H. *et al*. A multi-modal open dataset for mental-disorder analysis. *Scientific Data* **9**, 178 (2022).
20. Valstar, M. *et al*. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 3–10 (2013).
21. Valstar, M. *et al*. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 3–10 (2014).
22. Beck, A. T., Steer, R. A., Ball, R. & Ranieri, W. F. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment* **67**, 588–597 (1996).
23. Shen, G. *et al*. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, 3838–3844 (2017).
24. Wang, Y., Wang, Z., Li, C., Zhang, Y. & Wang, H. A multitask deep learning approach for user depression detection on sina weibo. *arXiv preprint arXiv:2008.11708* (2020).

25. Rahman, A. B. S., Ta, H.-T., Najjar, L., Azadmanesh, A. & Gönul, A. S. Depressionemo: A novel dataset for multilabel classification of depression emotions. *Journal of Affective Disorders* **366**, 445–458 (2024).
26. Bagby, R. M., Ryder, A. G., Schuller, D. R. & Marshall, M. B. The hamilton depression rating scale: has the gold standard become a lead weight? *American Journal of Psychiatry* **161**, 2163–2177 (2004).
27. Yao, Z. *et al*. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547* (2021).
28. Demszky, D. *et al*. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054 (2020).
29. Cao, P. *et al*. A multimodal depression consultation dataset of speech and text with hamd-17 assessments. *Science Data Bank* https://doi.org/10.57760/sciencedb.27818 (2025).
30. Xu, J. *et al*. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215* (2025).
31. Chu, Y. *et al*. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759* (2024).
32. Yang, A. *et al*. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).
33. Grattafiori, A. *et al*. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
34. Hu, E. J. *et al*. Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).
35. DeVault, D. *et al*. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 1061–1068 (2014).
36. Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K. & Geralts, D. S. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics* **20**, 50–64 (2007).

Acknowledgements

This work is supported by the Key Research Program of the Chinese Academy of Sciences (Grant No. ZDBS-SSW-JSC006).

Author contributions

Pengfei Cao contributed to conceptualization, methodology design, and original draft writing; Yuanzhe Zhang was responsible for investigation, and draft writing; Chenxiang Zhang contributed to method development and experiment analysis; Wei Chen, Yan Liu, Shuang Xu, Miao Xu, Wenqing Jin, Jinjie Xu, Dan Wang, Wei Wang, Xue Wang, Wen Wang and Yanping Ren contributed to data collection, transcription, and annotation; Jun Zhao contributed to supervision and manuscript review; Rena Li oversaw supervision and manuscript review; Kang Liu contributed to funding acquisition, supervision, and manuscript review. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Z., R.L. or K.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.