https://doi.org/10.1038/s44184-025-00175-1

Identifying psychiatric manifestations in outpatients with depression and anxiety: a large language model-based approach

Shihao Xu[1,2,3,8], Yiming Yan[1,4,8], Yanli Ding[1,4,8], Feng Li[2], Shu Zhang[2], Haoyun Tang[1,4], Chao Luo[1,4], Yan Li[1,4], Hao Liu[1], Yu Mei[1], Wenjie Gu[1], Hong Qiu[1], Yong Wang[1,4], Jianyin Qiu[1,4], Tao Yang[3], Zike Wang[2], Qing Zhang[1,4,5], Haiyang Geng[3], Yunyun Han[3], Jun Shao[2], Nils Opel[6,7], Lidong Bing[3], Min Zhao[1,4,5], Yifeng Xu[1,4,5], Xun Jiang[2,3] ✉ & Jianhua Chen[1,3,4,5] ✉

Accurate psychiatric diagnosis and assessment are crucial for effective treatment. However, current diagnostic approaches heavily rely on subjective observations constrained by time and clinical resources. This study investigates the potential of using Large Language Models (LLMs) to identify the symptoms in psychiatrist-patient dialogues and use them as intermediate features to predict the diagnostic labels. We collected audio recordings of 1160 outpatients with depressive disorder and anxiety disorder. LLMs were trained and utilized to identify clinical symptoms, rate assessment scales, and an ensemble learning pipeline was designed to classify diagnostic results and symptoms with 10-fold cross-validation. The system achieved 86.9% accuracy for identifying the appearance of clinical annotations and 74.7% (77.2%) accuracy for identifying symptoms of anxiety (depression). In addition, analysis of LLM-generated features shows that depression cases exhibited prominent markers of anhedonia and decreased volition, whereas anxiety disorders were characterized by tension and an inability to relax.

Depression and anxiety disorders represent two of the most prevalent mental health conditions globally. Globally, it is estimated that over 300 million people suffer from major depressive disorders, which is equivalent to 4.4% of the world's population. A similar number of people suffer from anxiety disorders, often with co-occurring depression[1]. The emerging field of digital phenotyping, which involves the nuanced quantification of human phenotypic expression at the individual level through digital device data, offers a quantitative approach to longitudinal observation[2].

The emerging field of digital phenotyping, characterized by continuous and nuanced quantification of human phenotypic expression at the individual level by leveraging digital device data, provides a quantitative approach for longitudinal observation[2]. Researchers have demonstrated that social signals (e.g., linguistics, speech, etc.) play a crucial role in the diagnosis and assessment of patients with depression and anxiety[3,4]. In particular, the content of a patient's speech provides rich information about their mental state, cognitive patterns, and emotional experiences[5,6]. The linguistic features, topic choices, and narrative structures employed by individuals can offer valuable insights into their psychological well-being[6].

Recent advances in NLP, particularly in LLMs such as GPT[7], Gemini[8], and Qwen[9], demonstrate diverse capabilities in clinical reasoning, social media analysis, and psychiatric education[10], which could potentially provide objective, data-driven insights in psychiatry. Moreover, LLMs are able to process, generate, and respond to natural language inputs, which fit naturally into the NIMH's Research Domain Criteria (RDoC) framework, which suggests new ways of classifying mental disorders based on dimensions of observable behaviors[11]. In recent psychiatric studies, these LLMs excel at understanding and generating complex linguistic patterns with human-like performance, making them widely explored for social media content analysis[12,13], treatment performance enhancement[14–16], chat counselor[17,18], and supporting clinical decision-making[19,20] from an evidence-based practice perspective. Although LLMs demonstrate linguistic understanding and generation, they remain relatively scarce in producing objective digital biomarkers in psychiatry[21]. Studies have shown that the speech of patients

[1]Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [2]Theta Health Inc., Redwood City, CA, USA. [3]Tianqiao and Chrissy Chen Institute, Shanghai, China. [4]Shanghai Clinical Research Center for Mental Health, Shanghai, China. [5]Shanghai Key Laboratory of Psychotic Disorder, Shanghai, China. [6]University Hospital Jena Department of Psychiatry and Psychotherapy, Jena, Germany. [7]German Centre for Mental Health (DZPG), Berlin, Germany. [8]These authors contributed equally: Shihao Xu, Yiming Yan, Yanli Ding. ✉e-mail: xun.jiang@thetahealth.ai; jianhua.chen@smhc.org.cn

with depression and anxiety contains distinctive quantitative verbal and nonverbal digital markers compared to healthy controls[4,6], but these characteristics often remain too subtle for humans to perceive actionable insights, making their practical application and improvement challenging[22]. LLM is able to generate diagnostic results and provide reasoning steps, benefiting from a large amount of pre-training data. However, the interpretation and alignment of answers or decisions generated by LLM remain challenging[23]. Moreover, most studies on depression and anxiety rely primarily on two data sources: social media and structured clinical reports, and are often constrained by limited data availability[3]. Distinguishing between depression and anxiety in clinical settings remains challenging due to the overlap of symptoms and the high comorbidity rate, with limited research on the discovery of objective biomarkers for both conditions[21]. In addition, during clinical interviews, psychiatrists translate patients' informal symptom descriptions into professional diagnostic terminology; however, there remains a lack of approaches to automatically and effectively bridge this "semantic gap" between patients and clinicians.

To address these gaps in existing research, we collected a comprehensive dataset of psychiatric interviews at the Shanghai Mental Health Center (SMHC) in China, comprising over 15,000 min of speech recordings from 1160 individual outpatients with 25 different diagnoses. These recordings, primarily featuring patients diagnosed with depression and anxiety disorders, were collected in unstructured real-world environments to ensure ecological validity. To mimic the characteristics of clinical diagnosis, we designed a corpus of clinical indicators that incorporates diagnostic criteria, main complaints, mental status evaluations, and components from assessment scales using the Electronic Medical Records (EMRs) in the SMHC and widely-used assessment scales. Subsequently, we employed the pre-trained LLM to indicate the appearance of a corpus of clinical-related symptoms, rate the components of several assessment scales, and further fine-tuned the LLM with clinical annotations from professional psychiatrists to enhance its understanding of clinical-related concepts. In parallel, we extracted linguistic usage patterns and acoustic features to broaden the spectrum of biomarkers. Through the fusion of these modalities, we constructed an ensemble machine-learning pipeline capable of predicting both outpatient diagnostic groups and symptoms with moderately high accuracies. Moreover, we conducted an in-depth analysis of salient patterns between different diagnostic groups to enhance clinical interpretability. Our results demonstrate that objective cues extracted by the LLM, combined with other behavioral markers, can serve as valuable features for differentiating diagnostic groups and identifying symptom disclosure, potentially enhancing both the efficiency and effectiveness of psychiatric diagnosis and assessment in clinical practice.

Methods

This study collected the audio recording of 1160 participants between August 2023 and January 2024, in collaboration with the SMHC. The overall pipeline is shown in Fig. 1. Data collection was conducted using the Scientific Speech Transcription Pen M1 (Iflytek Co., Ltd.) with a sampling rate of 44100 Hz. Firstly, the protocol involved three primary stages: preprocessing of audio samples, anonymization of personally identifiable information, and subsequent transcription via automated speech recognition systems followed by meticulous manual verification to ensure transcriptional accuracy. Secondly, we collaborated with professional psychiatrists to design a set of clinical entities and leveraged the LLM to identify these concepts using the transcripts as input, enhancing the LLM based on the psychiatrists' annotations through supervised fine-tuning (SFT). Linguistic and acoustic features were extracted from both the transcripts and the speech. Finally, we utilized different modalities to train an ensemble machine learning pipeline to differentiate diagnostic groups and the major symptoms.

Participants

The study sample comprised outpatients from the SMHC who attended clinical diagnostic interviews. Participants were aged 12 to 80 years and were fluent in Mandarin. Informed consent to participate in the study was obtained from all participants or their legal guardians, as appropriate. All diagnoses were established using the Chinese version of WHO International Classification of Diseases, Tenth Revision (ICD-10)[24]. The study protocol was approved by the Ethics Committee of the SMHC institutional review board (IRB) to ensure compliance with ethical research standards. Specifically, the recording setup consisted of a microphone placed between the psychiatrist and the participant, connected to a computer. At the beginning of each interview, participants were asked to read a standardized 30-second text passage, followed by the standard diagnostic procedure. All clinical information was documented in the EMR system by the psychiatrists. To protect the privacy of participants, all audio recordings and associated meta-information underwent a thorough manual de-identification process.

Feature extraction

We extracted a comprehensive clinical entity set to cover the intermediate features that assist psychiatrists in the diagnosis and assessment process: clinical observations and standardized assessment scales, which we designate as clinical-related and assessment-related feature sets. A clinical entity, in the context of our pipeline, is a structured representation of a psychiatric symptom or construct, developed from both clinical observations and standardized assessment scales. It encompasses key terms, expressions, and severity indicators related to specific diagnostic features, and serves as a unified unit for symptom detection and classification in our system. As compensation, we measured the linguistic usage and acoustic characteristics and form as individual feature sets. In the following paragraphs, we will introduce how we build and extract these feature sets in detail.

The clinical-related feature set encompasses essential depression and anxiety indicators extracted from EMRs with comprehensive descriptions (shown in Supplementary Table 1). This feature set was developed through a collaborative approach involving both psychiatrists and LLM analysis. Firstly, the process began with extracting 218 clinical entities from three sections in the EMR system: chief complaint, personal medical history, and psychiatric examination. These entities represent predefined features within the documentation framework of the SMHC EMR system based on psychiatric diagnostic systems, textbooks, and experts' opinions. Then, we included a supplementary of 44 additional symptoms identified through clinical expertise and diagnostic criteria (e.g., DSM-5 and ICD-10) suggested by psychiatrists. We then utilized the Gemini 1.5 Pro[8] to generate descriptions for all clinical entities, using the Chinese version of the DSM-5 guidance[25] as a reference, leveraging the model's strong extended context window capability. Through iterative psychiatric review, redundant and irrelevant items specific to depression and anxiety were eliminated, resulting in a refined set of 138 validated clinical-related features.

After rigorously defining the clinical-related features, we leveraged large language models to extract symptom information from diagnostic conversations. We employed Qwen2-72B-Instruct[26] as our foundational model due to its advanced Chinese language processing capabilities and suitability for offline deployment within secure hospital environments. To enhance domain-specific performance, we implemented SFT using psychiatrists' annotations from electronic medical records EMRs. This approach adapted the base model to better recognize specialized medical terminology and clinical reasoning patterns specific to psychiatric assessment contexts. The fine-tuning methodology treated symptom identification as an autoregressive task, where the model learns to predict token probabilities based on previous context, ultimately generating binary judgments ("yes" or "no") regarding specific symptom presence. The training data comprised individual samples where dialogue content, patient demographics (age, gender), and symptom categories were incorporated into prompt templates alongside corresponding symptom occurrence labels extracted from EMRs. For each clinical conversation, we systematically extracted all documented symptoms to create comprehensive training instances. The fine-tuning implementation utilized LLaMA-Factory (https://github.com/hiyouga/LLaMA-Factory), while inference processes were facilitated through vLLM (https://github.com/vllm-project/vllm). All

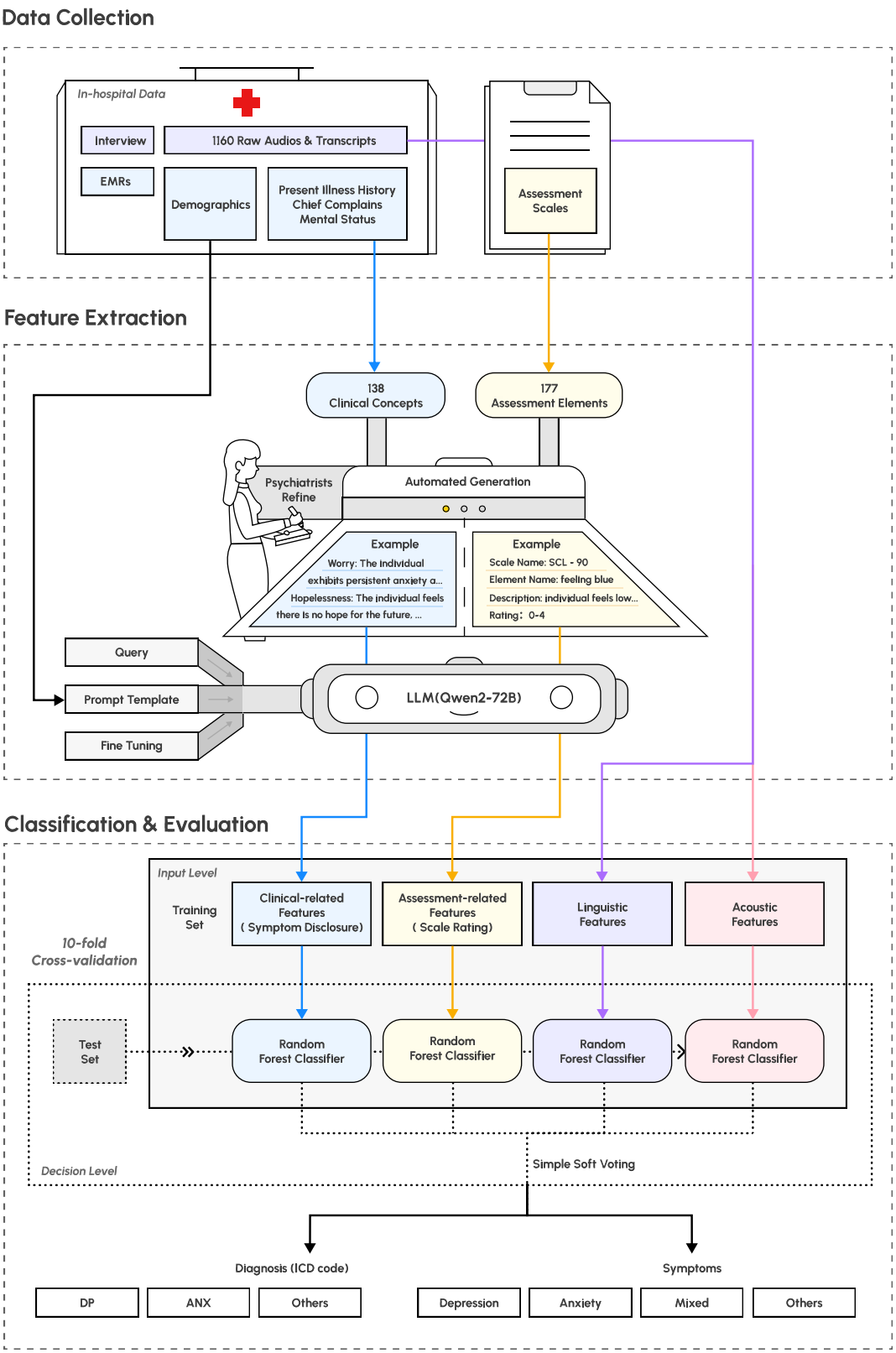

Fig. 1 | Diagram of the analysis pipeline. The audio recordings were collected during the diagnosis interview for outpatients. We extracted four types of feature sets from the recordings, two of which utilized LLM. These feature sets were utilized to classify different groups of participants and predict the appearance of depression and anxiety symptoms.

computational procedures were executed on a high-performance computing infrastructure consisting of four A100 GPUs. Table 1 presents the prompt architecture used both for clinical feature generation and model fine-tuning (Supplementary Table 4 for the Chinese version), demonstrating our structured approach to symptom extraction within extended clinical dialogues.

We first began with structuring EMR data to create reliable training labels for the SFT. Since EMRs contain unstructured text fields where

Table 1 | Prompt template for clinical-related feature generation

Prompt Template for Clinical-Related Features Generation

You are an excellent AI assistant capable of helping psychiatrists determine whether a patient exhibits certain symptoms based on the psychiatrist-patient interview dialogue. Here is the content of the interview dialogue:

Interview Dialogue

{dialogue}

Patient: {age}-year-old {gender} **Task:** Please determine if the participant shows symptoms of {symptom}.
Symptom description: {symptom_desc}
Answer directly with 'Yes' or 'No' without any explanation.
Your judgment is:

The content within the curly braces is the demographic, symptom descriptions, and dialogue information that form the prompt.

psychiatrists document patient information, we employed the LLM to analyze these 1160 EMRs. For each EMR, we leveraged LLM to evaluate the presence of above mentioned 138 predefined clinical features, including similar expressions and synonyms, generating a boolean value list (yes/no) for each record. The prompt for querying the LLM to generate labels from EMRs is shown in Supplementary Table 3 (Supplementary Table 6 for the Chinese version).

Secondly, we implemented a rigorous filtering process for choosing high-quality data for SFT. We first leverage LLM to verify whether the information recorded in EMRs was adequately reflected in the interview dialogue transcripts, yielding 877 valid examples. Then, we collaborated with specialist psychiatrists to establish comprehensive evaluation criteria, encompassing five standards for psychiatric examination, one for chief complaints, and five for present illness history assessment. By using these criteria as the prompt (shown in Supplementary Table 3), we employed the LLM to evaluate each case and select the top 60% (527 examples) as high-quality cases based on the total score. From these high-quality cases, we allocated 477 cases for the SFT and 50 cases for the high-quality test set. The 50 high-quality test cases and 633 lower-quality cases are combined as a completed test set to evaluate the accuracy of clinical-related feature extraction.

Subsequently, we fine-tuned the Qwen2-72B-Instruct model with Low-Rank Adaptation (LoRA)[27]. The LLM SFT involves training a pre-trained model on datasets with explicit input-output pairs to optimize the model's performance on specific downstream tasks. LoRA is a parameter-efficient fine-tuning technique that adds small, trainable rank decomposition matrices to the LLM's existing weights, allowing for efficient model adaptation while keeping most of the original model parameters frozen. The model was trained using the following hyperparameters: LoRA rank of 8, LoRA alpha of 16, batch size of 8, and an initial learning rate of 1e-4 for 7000 steps. During inference using the vLLM framework, we restricted the model's output to a single token "Yes" or "No" as the binary output, while we also extracted the probability output for these two tokens from the whole vocabulary. After normalization of the probabilities, along with the binary outputs, we formed 276 features in the clinical-related feature set.

The assessment-related feature set incorporates data from six validated psychiatric rating scales, combining self-rating and peer-rating instruments. Self-rating scales include SCL-90[28], SDS[29], and SAS[30], while peer-rating scales comprise HAMD[31], HAMA[32], and MADRS[33], totaling 177 items in all. These scales were selected for their proven reliability in clinical practice and research, offering comprehensive symptom coverage.

We designed two meta-prompts to enable the LLM to mimic both psychiatrists and patients in rating assessment scales in a zero-shot manner, as illustrated in Supplementary Table 2 and (Supplementary Table 5 for the Chinese version). The scales' content and rating guidelines were integrated into the prompts for LLM to generate the features. For instance, when extracting features related to the first item of the HAMD, which measures depressed mood, we use the peer-rating meta-prompt to instruct the LLM to evaluate the severity of the patient's depressed mood on a 0–4 scale based on age, gender, and conversation content, where 0 indicates the absence of depression and 4 represents severe depression. When the conversation lacks sufficient information about the depressed mood, the LLM is prompted to return "NULL". Similar to the clinical-related feature extraction, we extracted and normalized the logits of related tokens from the last layer of LLM and served as the features for classification and prediction tasks, resulting in a total of 1199 features. We did not SFT the LLM for assessment-related feature extraction, since we do not have sufficient assessment scale labels.

In addition to the features generated by LLM, we extracted verbal features through two bag-of-words approaches: LIWC[34] and TF-IDF[35], both of which measure the frequency of word occurrence within a document. The LIWC tool is specifically designed to provide rich insights into psychological states, including emotions, thinking styles, and social concerns. Notably, since our transcripts are in Mandarin, we used the Simplified Chinese version of LIWC[36]. It comprises word counts for 63 categories, including 52 categories related to linguistic counts (e.g., function words, common verbs, numbers, etc.), psychological processes (e.g., affect, sociality, cognition, perception, drive, etc.), and personal concern (e.g., work, home, religion, etc.), as well as 7 emotional categories (e.g., happy, sad, fear, etc.) and 4 general text metrics (e.g., the number of unique words, words in LIWC dictionary, etc.). We normalized the LIWC category counts by the total number of words.

The TF-IDF algorithm, which stands for Term Frequency-Inverse Document Frequency, is a popular technique used in text analysis to determine the importance of words within a document or collection of documents. Unlike simple word counting, TF-IDF considers both how often a word appears in a specific document and how common or rare that word is across all documents. This approach helps identify words that are particularly characteristic or important to specific documents. In this study, TF-IDF was applied alongside LIWC to provide a more comprehensive analysis of the verbal features in the documents, offering insights into both the frequency and relevance of words used by the subjects. We applied Jieba (https://github.com/fxsjy/jieba) for Chinese character segmentation, resulting in a total of 27,000 features.

In addition to examining the verbal aspects of participants' speech, we preprocessed the audio and extracted low-level acoustic and prosodic features using the OpenSMILE toolkits[37]. The audio recordings were manually edited to obscure names, addresses, and personally identifiable information before analysis. To reduce the impact of environmental noise and the varying distance from the microphone to the participant on recording quality, we used the pyAudacity toolkit (https://github.com/asweigart/pyaudacity) and the FFmpeg-normalized toolkit (https://github.com/slhck/

ffmpeg-normalize) to reduce the noise with a parameter of 12 dB and normalize the volume to −23 dB respectively. OpenSMILE is a versatile, customizable suite of acoustic features for signal processing and machine learning applications. We utilized OpenSMILE's emobase_live4 configuration to extract the following LLDs (Low-Level Descriptors): intensity, loudness, 12 MFCCs, pitch (F0), voicing probability (VoiceProb), F0 envelope (F0env), 8 line spectral frequencies (LSF), and Zero-Crossing Rate (ZCR). Next, we applied various functions to these LLDs and their delta coefficients, including minimum and maximum values with their relative positions (minPos and maxPos), range, mean, linear regression coefficients (linregc1-2), linear and quadratic error, standard deviation (STD), skewness, kurtosis, quartile values (quartile1-3), and interquartile ranges (iqr1-2, iqr2-3, iqr1-3). This process yielded 988 features to represent each speech utterance. Before LLD computation, pauses and silences were eliminated from the speech to create a continuous signal. We then extracted 988 emotion-based prosodic features using a 100 ms sliding window over the entire speech sample. Lastly, we calculated these emotion-based features' maximum, minimum, mean, and standard deviation to compose the final set of OpenSMILE features, totaling 3952 features.

Classification method

As explained in previous sections, we extracted five feature sets using LLM and existing toolkits: clinical-related, assessment-related, LIWC, TF-IDF, and OpenSMILE features. Subsequently, we built a machine learning pipeline to fuse the outputs from multiple feature sets to predict the appearance of the symptom and classify diagnostic groups, which was implemented using Scikit-learn 1.2.0 in Python 3.10. Notably, as detailed in explaining the clinical feature extraction, we fine-tuned the LLM using 138 high-quality clinical annotations to improve its ability to identify clinical concepts. We excluded diagnostic labels from this process to prevent data leakage.

To ensure robust validation, we employed 10-fold cross-validation (10-fold CV). This method involves dividing the data into 10 subsets, iteratively training the model on 9 subsets, and testing it on the remaining subset. The process is repeated 10 times, with each subset serving as the test set once, and the model's performance is averaged across all iterations. We implemented the random forest classifier for all feature sets as it constantly achieved better performance than other types of classifiers. To address the challenge of class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE)[38], which generates synthetic data for minority classes. Furthermore, we performed z-score standardization on all features, resulting in standardized features with a mean of 0 and a standard deviation of 1. This step ensures that all features are on a comparable scale, preventing any single feature from dominating the analysis due to its magnitude.

We also implemented probability calibration to standardize predictions from each feature set. This process involved an internal CV on the training set of the outer CV to obtain the probability distribution on training data, which were then used to calibrate test set predictions[6]. Moreover, based on the feature importance ranked by the classifier on training data, we filtered out features whose importance values fell below the mean to reduce unimportant features. For the final prediction, we employed a late fusion technique, a multi-modal machine learning approach that involved averaging the standardized prediction outputs from all feature sets to produce the final output. This method allows for the integration of diverse information sources while maintaining the integrity of each feature set's contribution to the final prediction.

Performance metrics

To evaluate the performance of the LLM in extracting clinical features from interview dialogues, we employed standard information extraction metrics: precision and recall. Precision measures the proportion of correctly identified symptoms among all symptoms extracted by the LLM, while recall measures the proportion of symptoms correctly extracted from the EMR annotations. Given that psychiatrists may not document every symptom mentioned during interviews in the EMRs, recall serves as a particularly valuable metric in our evaluation framework. Precision and recall are calculated as follows: $\text{Precision} = TP/(TP + FP)$; $\text{Recall} = TP/(TP + FN)$, where TP (True Positives) represents symptoms correctly identified by both the LLM, FP (False Positives) represents symptoms incorrectly extracted by the LLM, and FN (False Negatives) represents symptoms present in the EMR but missed by the LLM.

For classification and prediction tasks, we utilize a comprehensive set of standard metrics. Our analysis primarily focuses on balanced accuracy (BAC), which is particularly effective for imbalanced datasets by averaging sensitivity (SEN) and specificity (SPE). This metric provides a robust measure of overall performance, accounting for both true positive and true negative rates. In addition to BAC, we also employed AUPRC and weighted F1 score offering valuable insights into model performance across various classification thresholds, which are well-suited for machine learning tasks involving imbalanced data.

Understanding the key distinguishing features among various mental health conditions is crucial for improving diagnostic accuracy, developing targeted interventions, and enhancing our overall comprehension of these disorders. To address this critical need, we employed a comprehensive approach to identify the most important features distinguishing between different mental health conditions. We utilized various feature sets, including LLM-generated clinical and assessment-related features, LIWC categories, and TF-IDF terms, and applied the Mann-Whitney U test with FDR correction across all feature sets to calculate p-values and measure feature importance. Features are ranked by their p-values, with those below 0.05 indicating a statistically significant difference between the two groups.

Baseline experiment setup

Excepted the LIWC and TF-IDF, we incorporated both traditional transformer-based language models and LLM-based methods as the baselines for mental health dialogue classification. We established a comprehensive methodological framework encompassing three distinct classification paradigms: pre-trained language models, zero-shot LLM classification, SFT LLM classification, and our proposed pipeline.

We implemented two established pre-trained language models as baseline classifiers: BERT (Bidirectional Encoder Representations from Transformers)[39] and RoBERTa (Robustly Optimized BERT Pretraining)[40]. We utilized the "bert-base-chinese" model[39], which consists of 12 transformer layers with 768 hidden dimensions and 12 attention heads, totaling approximately 110 million parameters. BERT's bidirectional contextual representations enable effective capture of semantic nuances within clinical dialogues. For RoBERTa, we employed the "chinese-roberta-wwm-ext-large" variant[41], featuring 24 transformer layers, 1024 hidden dimensions, and 16 attention heads (approximately 325 million parameters). This model incorporates the whole word masking technique specifically optimized for Chinese language understanding. RoBERTa's enhanced training methodology and larger parameter space potentially offer improved representation capabilities for complex clinical narratives. Both BERT and RoBERTa were fine-tuned on the ANX vs. DP classification task using addition random initialed linear layer with softmax.

For the LLM-based baseline method, we implemented both zero-shot and SFT manner. (1) Zero-shot Classification: We implemented direct classification using Qwen2.5-72B-Instruct through carefully designed prompts (as shown in Supplementary Table 7) that incorporated dialogue content and diagnostic ground truth. The model was constrained to output binary classifications (depression/anxiety), with token probabilities extracted to calculate performance metrics. (2) SFT: We augmented the zero-shot approach through parameter-efficient fine-tuning using LoRA.

All methodologies underwent rigorous evaluation using consistent data partitioning and performance metrics. We implemented stratified sampling to allocate 60% of samples for training, 20% for validation, and 20% for testing across both depression and anxiety classes. Hyperparameter optimization was conducted using the validation set, while final performance evaluation utilized the held-out test set exclusively. Specifically, models exhibiting the lowest validation loss during the training process were

preserved and subsequently employed for final performance evaluation on the held-out test set. This ensured methodological consistency and facilitated direct comparative analysis of classification paradigms.

Results

Sample

The study included 1160 individuals, yielding about 15,000 minutes of speech data. All participants received diagnoses based on the ICD-10[24]. The sample comprised 553 participants diagnosed with "Depressive Episode" or "Depressive Disorder" (DP), 426 diagnosed with "Anxiety Disorder" or "Anxiety State" (ANX), and 181 classified as "Others" (patients not diagnosed with DP or ANX). Table 2 presents the demographic characteristics of the participants. Moreover, based on the clinical annotations of symptom episodes in the EMRs, we categorized the participants into four groups: patients who experienced/presented anxiety symptoms (A), participants who experienced/presented depressive symptoms (D), participants who experienced/presented mixed depressive and anxiety symptoms (M), and participants without experienced/presented depressive and anxiety symptoms (N).

LLM-generated clinical-related features evaluation

We evaluated the performance of LLM-generated clinical symptoms on the entire test samples and those with high-quality EMR, as shown in Table 3. Our evaluation of LLM-based clinical symptom extraction demonstrated a significant performance improvement after the SFT, with the accuracy increased from 81.2 to 86.9% on the test set and 83.7 to 89.1% on the high-quality test set ($p < 0.01$ in McNemar's test). The recall metric showed substantial improvements, increasing from 66.1 to 81.1% on the whole test set and from 74.0 to 86.1% on the high-quality test set, indicating enhanced capability in identifying symptoms documented by psychiatrists in the EMR. Meanwhile, precision improved from 81.2 to 87.4% on the test set and from 84.2 to 89.5% on the high-quality test set. This precision increase, coupled with recall improvement, suggests that the fine-tuned model became more comprehensive in detecting symptoms from clinical dialogues.

We present a comparative analysis of classification performance using clinical-related features extracted by the LLM in Fig. 2, comparing three feature sets: features extracted in a zero-shot manner, features extracted from the fine-tuned LLM, and psychiatrists' annotations derived from EMRs. Across all classification tasks, features from the fine-tuned LLM consistently demonstrate superior performance. For instance, in distinguishing between depression and anxiety diagnoses (A vs. D), the fine-tuned LLM achieves a BAC of 74.8%. In identifying depression (D vs. N) and anxiety symptoms (A vs. N), the BAC reaches 79.8% and 72.2% respectively. These results underscore the potential of fine-tuned LLMs for accurate and automated clinical manifestation extraction.

Classification of diagnostic groups

The results of automated classification tasks for distinguishing between ANX, DP, and Others groups (not diagnosed with ANX or DP) using various linguistic and LLM-generated features are shown in Table 4. For the binary classification task (ANX vs. DP), the model achieved a BAC of 75.5%, an F1 score of 0.762, and an AUPRC of 0.824, indicating good overall performance (permutation test $p < 0.01$, same for other tasks). In the three-way classification task (ANX vs. DP vs. Other), the model's performance was achieved with a BAC of 65.6% and an F1 score of 0.656, presenting a significant gain compared to the majority baseline (47.7%).

Prediction of depression and anxiety symptoms

In addition to identifying diagnostic results by ICD-10 code, we predicted whether participants exhibited symptoms of depression, anxiety, mixed depression/anxiety, or no symptoms at all, as shown in Table 5. In the anxiety vs. no anxiety (A vs. N) classification task, the model achieved a sensitivity of 0.683 and specificity of 0.810 for detecting anxiety, with an overall F1 score of 0.754 and BAC of 74.7%. For the depression vs. no depression (D vs. N) task, the model performed slightly better, with a sensitivity of 0.806 and specificity of 0.737 for detecting depression, resulting in an F1 score of 0.783 and a BAC of 77.2%. When distinguishing between anxiety, depression, mixed symptoms, and no depression and anxiety symptoms (A vs. D vs. M vs. N), we achieved an AUPRC of 0.606 and a BAC of 60.7%, which achieved a significant improvement of about 30% compared to the majority baseline.

Interpretability

The analysis revealed distinctive patterns across different mental health conditions and feature sets (Table 6). In differentiating ANX from DP, clinical-related features emphasized anxiety-specific symptoms such as "Unable to relax", "Uncontrollable restlessness", and "Anxiety", contrasting with depressive symptoms like "Sadness" and "Anhedonia". Assessment measures showed a mixed profile, with both anxiety indicators (HAMD_-Somatic anxiety) and depression markers (HAMD_Depressed mood). LIWC analysis revealed heightened use of anxiety and fear-related language, and TF-IDF identified anxiety-related terms. For depression detection, clinical-related features highlighted core depressive symptoms, with "Depressed mood", "Loss of interest", and "Anhedonia" emerging as primary indicators of depression. The assessment-related features showed

Table 2 | Demographics of all participants

	DP (N = 553)	ANX (N = 426)	Others (N = 181)
Age	29.2 ± 10.0	34.0 ± 12.0	27.5 ± 11.1
Gender			
Female	377 (68.2%)	288 (67.6%)	104 (57.5%)
Male	176 (31.8%)	138 (32.4%)	77 (42.5%)
Occupation			
Employed	266 (48.1%)	250 (58.7%)	74 (40.9%)
Student	201 (36.3%)	98 (23.0%)	81 (44.8%)
Unemployed	46 (8.3%)	44 (10.3%)	14 (7.7%)
Unknown	23 (4.2%)	15 (3.5%)	7 (3.9%)
Retired	9 (1.6%)	19 (4.5%)	4 (2.2%)
Dropped out	8 (1.4%)	NaN	1 (0.6%)
Personality			
Introvert	274 (49.5%)	198 (46.5%)	106 (58.6%)
Extrovert	148 (26.8%)	114 (26.8%)	34 (18.8%)
Gentle	46 (8.3%)	38 (8.9%)	13 (7.2%)
Sensitive	29 (5.2%)	22 (5.2%)	4 (2.2%)
Strong welling	12 (2.2%)	14 (3.3%)	8 (4.4%)
Others	44 (8.0%)	40 (9.4%)	16 (8.8%)

Table 3 | Performance comparison of LLM-generated clinical-related features between Zero-shot and SFT approaches

	Test set			High-quality test set		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
Zero-shot	81.2%	66.1%	81.2%	84.2%	74.0%	83.7%
the SFT	87.4%	81.1%	86.9%	89.5%	86.1%	89.1%

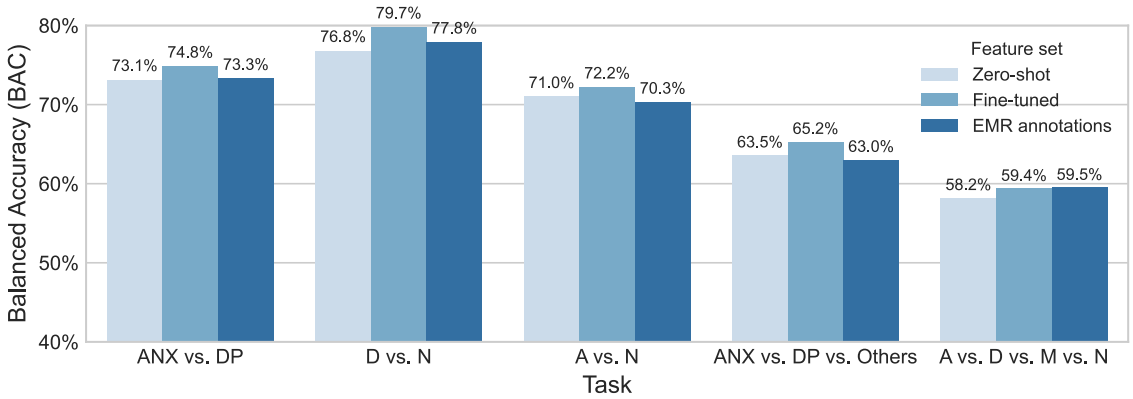

Fig. 2 | Comparative analysis of classification performance using the clinical-related features extracted by LLM in zero-shot, the SFT, and the annotations from EMRs across different classification tasks. "ANX" represents Anxiety Disorder, "DP" represents depression Disorder, "A" represents participants with anxiety symptoms, "D" represents participants with depressive symptoms, "M" represents participants with mixed anxiety and depressive symptoms, and "N" represents participants without anxiety and depressive symptoms.

Table 4 | Results for classification of ANX, DP, and Others groups

Task	Feature Set	Confusion Matrix				SEN	SPE	F1	AUPRC	BAC	MB
			Predicted Class								
			A	D							
ANX vs. DP	AssRel + CliRel + LIWC + TF-IDF	**A**	294	132		0.690	0.819	0.762	0.824	0.755	0.565
		D	100	453							
			A	**D**	**O**	**SEN**	**SPE**	**F1**	**AUPRC**	**BAC**	**MB**
ANX vs. DP vs. Others	AssRel + CliRel + LIWC + TF-IDF	**A**	271	91	64	0.636	0.838	0.656	0.715	0.656	0.477
		D	92	357	104						
		O	27	30	124						

AssRel Assessment-related, *CliRel* Clinical-related, *F1* F1-score, *SEN* Sensitivity, *SPE* Specificity, *AUPRC* Area under precision-recall curve, *BAC* Balanced Accuracy, *MB* Majority Baseline.

Table 5 | Results for classification of participants with depression (D), anxiety (A), mixed depression and anxiety (M), and no depression and anxiety symptoms (N)

Task	Features	Confusion Matrix					SEN	SPE	F1	AUPRC	BAC	MB
			Predicted Class									
			A		N							
A vs. N	AssRel + CliRel + LIWC + TF-IDF	**A**	285		135		0.683	0.810	0.754	0.813	0.747	0.565
		N	176		564							
			D		**N**		**SEN**	**SPE**	**F1**	**AUPRC**	**BAC**	**MB**
D vs. N	AssRel + CliRel + LIWC + TF-IDF	**D**	595		143		0.806	0.737	0.783	0.866	0.772	0.636
		N	111		311							
			A	**D**	**M**	**N**	**SEN**	**SPE**	**F1**	**AUPRC**	**BAC**	**MB**
A vs. D vs. M vs. N	AssRel + CliRel + LIWC + TF-IDF	**A**	114	2	17	12	0.786	0.865	0.586	0.606	0.607	0.399
		D	22	272	98	71						
		M	55	58	123	39						
		N	60	30	19	168						

AssRel Assessment-related; CliRel Clinical-related, *F1* F1-score, *SEN* Sensitivity, *SPE* Specificity, *AUPRC* Area under precision-recall curve, *BAC* Balanced Accuracy, *MB* Majority Baseline.

strong signals from SCL-90 scales, particularly in items related to feelings of sadness and loss of interest. LIWC analysis identified significant usage patterns in sadness-related words and negative emotions, while TF-IDF analysis captured depression-specific terms and notably, negation patterns (e.g., "Don't want", "No", etc.). For anxiety identification, clinical-related features strongly centered on anxiety manifestations, such as "Unable to relax", "Anxiety", and "Worry." The assessment-related features prominently featured inner tension and somatic anxiety, along with various SCL-90 anxiety-related items. Both LIWC and TF-IDF analyses consistently identified anxiety-specific language patterns, with LIWC showing "Anxiety" and "Fear" as top features, and TF-IDF highlighting terms related to physical symptoms (e.g., "Palpitations", "Heartbeat", etc.) and worry.

Baseline experiment results

We conducted a systematic investigation of diagnostic efficacy in clinical dialogue classification utilizing multiple model architectures. The

Table 6 | Top ten salient features for each feature set in paired classification tasks

Task	Clinical-related	Assessment-related	LIWC	TF-IDF
ANX vs. DP	Sadness	**SCL-90_Feeling blue_2**	**Anxiety**	**Anxiety**
	Perturbed and uneasy	**SCL-90_Feeling no interest in things_2**	**Fear**	Ideas
	Unable to relax	HAMD_Depressed mood_4	**Biological Processes**	Emotion
	Uncontrollable	**HAMD_Somatic anxiety_3**	Death	**Palpitations**
	-restlessness	MADRS_Inner tension_NULL	Sad	**Anxious**
	Negativism	**SCL-90_Thoughts of ending your life_1**	**Health**	**Worried**
	Anxiety	**SCL-90_Feeling future hopeless_1**	**Body**	**Terrified**
	Anhedonia	**MADRS_Suicidal ideation_0**	**Motion**	**Heartbeat**
	Anxiety and unease	**HAMD_A sense of hopelessness_0**	**Perfect tense**	**Excited**
	Negative ideation	**SCL-90_Never feeling close to others_1**	Anger	**Behavior**
	Anhedonia			
A vs. N	**Unable to relax**	**MADRS_Inner tension_4**	**Anxiety**	**Anxiety**
	Anxiety	**HAMD_Somatic anxiety_3**	**Fear**	**Anxious**
	Uncontrollable	SCL-90_Feeling tense or keyed up_NULL	**Body**	**Palpitations**
	-restlessness	**HAMD_Psychiatric anxiety_3**	**Biological Processes**	**Worried**
	Anxiety and unease	**HAMA_Autonomic symptoms_2**	Good Social	Ideas
	Feeling of tension	SCL-90_Worrying too much about things_NULL	**Perfect tense**	**Heartbeat**
	Somatic anxiety	SCL-90_Nervousness or shakiness inside_NULL	**Health**	**Anxiety disorders**
	Excessive worrying	SAS_Anxiety_NULL	**Motion**	**Chest tightness**
	with anxious	**HAMA_Cardiovascular symptoms_2**	Death	**Behavior**
	Worry	SCL-90_Feeling restless_NULL		**Comfortable**
	Delusion of Guilt			
	Palpitations			
D vs. N	**Feeling Down**	SCL-90_Feeling blue_NULL	**Sad**	**No**
	Perturbed and Uneasy	SCL-90_Feeling no interest in things_NULL	**Affect**	**Depression**
	Sadness	SCL-90_Feeling hopeless about the future_NULL	Health	**Interests**
	Loss of interests	SCL-90_Feeling everything is an effort_NULL	**Death**	**Emotion**
	Anhedonia	**HAMD_Depressed mood_3**	**Humans**	**Yes**
	Negativism	**HAMD_Work and interests_3**	Biological processes	**Obstacle**
	Hypobulia	SCL-90_Feeling low in energy	**Anger**	**Don't**
	Low self-evaluation	-or slowed down_NULL	**Achievement**	**Contrary**
	Helplessness	SCL-90_Feelings of worthlessness_NULL	**Negative emotion**	**Ideas**
	Abulia	**SDS_Depression_4**	Anxiety	**Anger**
		MADRS_Apparent sadness_0		

All features in the table have $p < 0.01$.

For Assessment-related features, the feature nomenclature follows the format: Scale_Symptom-Name_Rating, where a 'NULL' rating indicates the absence of symptom in dialogue identified by LLM. The **bold** feature name represents that the underlined class has a higher mean value.

conventional transformer models demonstrated variable performance: BERT achieved modest results with 88.1% sensitivity but only 39.5% specificity and 64.3% F1 score, while RoBERTa exhibited superior sensitivity (96.3%) but similarly limited specificity (37.2%). When examining LLM-based approaches, Qwen2.5-72B-Instruct demonstrated substantial improvements in balanced performance, achieving 75.0% balanced accuracy in zero-shot configuration and 76.7% after SFT. Our proposed methodology, leveraging LLM-generated feature sets with ensemble random forest classifiers, outperformed all other approaches across most metrics, most notably achieving 79.1% balanced accuracy and 88.1% AUPRC, demonstrating the efficacy of feature extraction over direct classification when utilizing large language models for clinical diagnostic tasks.

Discussion

Inspired by promising early research on digital phenotypes for diagnosing and classifying symptoms in psychiatric patients, we investigated using signal processing and state-of-the-art LLM to capture symptom-related expression cues in outpatient conversations. We developed an ensemble classification pipeline to automatically differentiate between clinical diagnostic outcomes and the presence of symptoms.

Although recent studies have demonstrated promising capabilities of utilizing LLMs in medical diagnosis[42], applications in mental health have predominantly centered on developing conversational agents[43], while the potential of extracting precise symptoms from psychiatric conversations for evidence-based diagnosis has not been fully explored. In this study, we investigated the efficacy of LLM in detecting clinical and assessment-related symptoms. Our investigation revealed that without any additional training, the model achieved a recall rate of 77.3% on high-quality dialogue-case pairs, and increased to 86.1% by fine-tuning the LLM using clinical annotations. This aligns with recent observations regarding LLMs' strong zero-shot performance in healthcare domains and the fine-tuning could further boost LLM performance[23]. Furthermore, this enhanced base capability led to substantial improvements across all downstream classification and prediction tasks (e.g., the classification accuracy for ANX and DP increased from 72 to 75%). Current approaches to automated symptom detection predominantly rely on traditional natural language processing methods with predefined linguistic categories or rule-based systems[6,44], which often struggle to capture the complex presentation of psychiatric symptoms in natural conversation. Some researchers have explored the use of LLMs to assist in medical information retrieval[45]. We further investigated the information extraction capabilities in clinical dialogues and enhanced them through SFT.

Our study demonstrated moderate to high performance in anxiety symptom detection (BAC = 74.7%, AUPRC = 0.813), depression symptoms detection (BAC = 77.2%, AUPRC = 0.866), and a four-class classification of patients with anxiety/depression/mixed/none symptoms (BAC=60.7%, AUPRC = 0.606). As shown in the summarization of existing literature (Supplementary Table 8), while anxiety detection in social media text has demonstrated promising results with high accuracy[46], the performance of similar methods on spoken language data, such as interview transcripts and therapy dialogues, remains limited with accuracy rates below 65%. Recent advances combining LLM embedding with acoustic features have shown improved results, reaching 75% accuracy in a small cohort of 65 patients[4]; however, in our experiments, incorporating acoustic features did not yield improvements in overall classification performance as shown in Supplementary Fig. 1. This might be attributed to the noisy hospital environment and the limitations of our recording equipment, which resulted in suboptimal audio quality. While depression detection studies have reported wide-ranging accuracy rates (65−95%), some results should be interpreted with caution due to several methodological limitations: small sample sizes[44], reliance on PHQ screening tools rather than clinical diagnoses[47], data collection in structured experimental settings[48], and not studied the first-episode outpatients in real-world, unstructured clinical environments. Our study leverages clinical diagnoses from psychiatrists of first-episode outpatients in real-world clinical environments, achieving moderate to high accuracy despite the inherent challenges and variability of unstructured, naturalistic settings representing a significant advancement over controlled laboratory conditions. This success particularly highlights the potential of

Table 7 | Comparison of classification performance using Qwen2-72B-Instruct (Zero-shot and SFT) and Ours (LLM-generated feature sets with ensemble random forest classifiers) for DP/ANX classification

	SEN	SPE	F1	AUPRC	BAC	MB
BERT	0.881	0.395	0.643	0.735	63.8%	0.566
RoBERTa	**0.963**	0.372	0.669	0.708	66.8%	0.566
Qwen2.5-72B-Instruct (Zero-shot)	0.748	0.753	0.751	0.842	75.0%	0.566
Qwen2.5-72B-Instruct (SFT)	0.757	0.776	0.766	0.828	76.7%	0.566
Ours	0.788	**0.793**	**0.791**	**0.881**	**79.1%**	0.566

The bold numbers indicate the highest performance within each metric column.

LLMs in extracting and analyzing clinical symptoms for predicting anxiety and depression in outpatient populations, offering a more ecologically valid and scalable solution for mental health screening and monitoring.

DP and ANX present significant assessment challenges due to their high prevalence, frequent comorbidity, and overlapping symptomatology[49]. By leveraging LLM-generated features, our approach achieved robust performance in distinguishing these disorders, with a BAC of 75.5% and AUPRC of 0.824 for binary classification between DP and ANX, and the performance outperformed the directly using LLMs as classifiers (see Table 7). In the more challenging multi-class scenario (ANX vs. DP vs. Others), the model maintained reasonable performance with a BAC of 65.6% and AUPRC of 0.715. Prior approaches to differentiating depression and anxiety disorders, such as cognitive tasks[50] and structured questionnaires[51], have achieved accuracy rates of 70–80%. In addition, we tested the classification performance of each assessment scale as the feature set, where the results are presented in Supplementary Fig. 2. We observed that assessment-related features, particularly from scales like SCL-90, HAMD, and MADRS, showed strong discriminatory power across all comparisons, and early fusion and late fusion present similar classification performance. A potential reason is that these scales contain sufficient depression-related symptoms, which are key components for differentiating different groups. To our knowledge, no study has explored the objective diagnosis of DP and ANX using speech data from clinical interviews, potentially due to a lack of data and inherent subjectivity. Our study addresses a critical gap by analyzing the linguistic and symptom-related markers in various participant groups, providing objective cues to assist psychiatrists.

The feature analysis provides several key insights into the differential characteristics of different groups of participants, as shown in Table 6. We illustrate the distribution of clinical and assessment-related features for each group of participants in Supplementary Figs. 3 and 4. The clinical-related features demonstrate clear condition-specific patterns: features that show more importance in patients with depression cluster around mood (sadness and disappointment) and motivational disturbances (anhedonia, reduced volition), while anxiety features predominantly reflect an inability to relax and worry. The observation for depression is in line with previous studies which also observed that patients with depression presented blunted facial affect and increased sadness in language[4,13] and anhedonia is specific to depression[52]. For anxiety recognition, the consistency of findings across different feature sets strengthens the reliability of these discriminators. For instance, the prominence of somatic symptoms in anxiety, captured in both assessment-related features and TF-IDF terms, suggests this could be a robust marker. Similarly, the persistent appearance of mood-related terms in depression across multiple feature sets reinforces their diagnostic utility. It is worth noting that acoustic features extracted using OpenSmile were not included in our feature importance analysis, as they did not demonstrate statistical significance in discriminating between the participant groups.

Real-world implementation of this LLM pipeline demands careful consideration of practical and clinical factors. Our approach, leveraging LLMs on conversational data to derive symptom insights and classifications, underscores the need for stringent data privacy protocols and computationally capable infrastructure. Furthermore, the pipeline's interpretability, stemming from its focus on clinically relevant features, must be clearly presented within EMR workflows to foster clinician trust. Beyond these pipeline-specific needs, seamless integration and clinician training are critical for usability. Building trust also requires ongoing validation, performance monitoring to detect model drift, and transparent ethical protocols, including patient consent and equity audits, ensuring the tool responsibly supports clinical decision-making.

Our study has several limitations that should be addressed in future research. The absence of detailed symptom severity measures during the experiment limits our ability to correlate speech patterns with specific symptom intensities. Additionally, the study's focus on specific disorders and potential biases in data collection may affect the generalizability of the results. Future work should prioritize the inclusion of comprehensive symptom severity assessments and explore the application of this approach to a broader range of mental health conditions. Besides, in the future, we will collect more data to perform longitudinal analysis, as it could provide insights into how linguistic patterns evolve with symptom progression or treatment response. Furthermore, expanding the use of more advanced LLMs in this context could potentially enhance the extraction of nuanced clinical concepts and provide even more detailed, interpretable insights for clinicians. Validating the model's performance in diverse clinical settings and with larger, more diverse patient populations will be crucial to ensure its practical utility and generalizability. These advancements could significantly contribute to improving the efficiency and objectivity of consultations for depression, anxiety, and potentially other mental health disorders.

In summary, this study demonstrates the potential of using LLM to analyze digital biomarkers in speech for automatic assistance in psychosis diagnosis and assessment. Our model achieved promising accuracy in identifying individuals with anxiety and depression symptoms, as well as differentiating between DP and ANX groups. Using LLMs to extract clinically relevant features and rate assessment scales improved the interpretability of the results, offering a novel approach to bridging the gap between automated analysis and clinical practice. While further research is needed, our findings suggest that well-developed LLMs could potentially serve as valuable tools in standardizing psychiatric evaluation and decision-making.

Data availability

The research data cannot be publicly shared due to privacy concerns, but the code and instructions for requesting local access to the data are available at https://github.com/Shanda-Group-Ltd/SMHC_llm_psychiatry_study, with qualified researchers able to apply for on-site data access through this repository.

Received: 24 January 2025; Accepted: 10 November 2025;
Published online: 02 December 2025

References

1. Chodavadia, P., Teo, I., Poremski, D., Fung, D. S. S. & Finkelstein, E. A. Prevalence and economic burden of depression and anxiety symptoms among Singaporean adults: Results from a 2022 web panel. *BMC Psychiatry* **23**, 104 (2023).
2. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* **9**, 211–217 (2016).
3. Sharma, C. M., Damani, D. & Chariar, V. M. Review and content analysis of textual expressions as a marker for depressive and anxiety disorders (DAD) detection using machine learning. *Discov. Artif. Intell.* **3**, 38 (2023).
4. Jiang, Z. et al. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE J. Biomed. Health Inform.* **28**, 1680–1691 (2024).

5. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig. Otolaryngol.* **5**, 96–116 (2020).
6. Xu, S. et al. Identifying psychiatric manifestations in schizophrenia and depression from audio-visual behavioural indicators through a machine-learning approach. *Schizophrenia* **8**, 1–13 (2022).
7. Openai. ChatGPT. https://chatgpt.com/chat (2024).
8. Team, G. et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at https://doi.org/10.48550/arXiv.2403.05530 (2024).
9. Team, Q. Qwen2.5: a party of foundation models. https://qwenlm.github.io/blog/qwen2.5/ (2024).
10. Omar, M. et al. Applications of large language models in psychiatry: a systematic review. *Front. Psychiatry* **15**, 1422807 (2024).
11. Marzano, L. et al. The application of mHealth to mental health: opportunities and challenges. *Lancet Psychiatry* **2**, 942–948 (2015).
12. Lan, X., Cheng, Y., Sheng, L., Gao, C. & Li, Y. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track* (2025).
13. Wang, Y., Inkpen, D. & Kirinde Gamaarachchige, P. Explainable depression detection using large language models on social media data. In *Proc. 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 108-126 (Association for Computational Linguistics, 2024).
14. Wang, X., Liu, K. & Wang, C. Knowledge-enhanced Pre-training large language model for depression diagnosis and treatment. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, 532-536 (IEEE, 2023).
15. Agrawal, A. Illuminate: a novel approach for depression detection with explainable analysis and proactive therapy using prompt engineering. Preprint at https://doi.org/10.48550/arXiv.2402.05127 (2024).
16. Chen, Z., Lu, Y. & Wang, W. Y. Empowering psychotherapy with large language models: cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4295–4304 (Association for Computational Linguistics, 2023).
17. Liu, J. M. et al. ChatCounselor: a large language models for mental health support. Preprint at https://doi.org/10.48550/arXiv.2309.15461 (2023).
18. Li, J. et al. Agent hospital: a simulacrum of hospital with evolvable medical agents. Preprint at https://doi.org/10.48550/arXiv.2405.02957 (2024).
19. Xin, A. W. et al. Using large language models to detect outcomes in qualitative studies on adolescent depression. *J. Am. Med. Inform. Assoc.* ocae298, https://doi.org/10.1093/jamia/ocae298 (2024).
20. Elyoseph, Z., Levkovich, I. & Shinan-Altman, S. Assessing prognosis in depression: Comparing perspectives of AI models, mental health professionals and the general public. *Fam. Med. Community Health* **12**, e002583 (2024).
21. Barua, P. D. et al. Artificial intelligence assisted tools for the detection of anxiety and depression leading to suicidal ideation in adolescents: a review. *Cognit. Neurodyn.* **18**, 1–22 (2024).
22. Ferguson, S., Aoyagui, P. A., Rizvi, R., Kim, Y.-H. & Kuzminykh, A. The explanation that hits home: the characteristics of verbal explanations that affect human perception in subjective decision-making8, 517:1–517:37. https://doi.org/10.1145/3687056.
23. Lawrence, H. R. et al. The opportunities and risks of large language models in mental health. *JMIR Ment. Health* **11**, e59479 (2024).
24. ICD-10 Version:2016. https://icd.who.int/browse10/2016/en.
25. Duckworth, K. *Understanding Mental Disorders: Your Guide to DSM-5.* (American Psychiatric Association Publishing, Washington, D.C., 2015) https://doi.org/10.1176/appi.ajp.2015.15070879.
26. Yang, A. et al. Qwen2 technical report. Preprint at arXiv preprint at https://arxiv.org/abs/2407.10671 (2024).
27. Hu, E. J. et al. LoRA: Low-rank adaptation of large language models. Preprint at arXiv preprint at https://arxiv.org/abs/2106.09685 (2021).
28. Derogatis, L. R. & Unger, R. Symptom Checklist-90-Revised. In Corsini Encyclopedia of Psychology. https://doi.org/10.1002/9780470479216.CORPSY0970 (2010).
29. ZUNG, W. W. K. A self-rating depression scale. *Arch. Gen. Psychiatry* **12**, 63–70 (1965).
30. Zung, W. W. A rating instrument for anxiety disorders. *Psychosomatics* **12**, 371–379 (1971).
31. Hamilton, M. The Hamilton Rating Scale for Depression. In Sartorius, N. & Ban, T. A. (eds.) *Assessment of Depression*, 143−152. https://doi.org/10.1007/978-3-642-70486-4_14 (Springer Berlin Heidelberg, 1986).
32. Hamilton, M. The assessment of anxiety states by rating. *Br. J. Med. Psychol.* **32**, 50–55 (1959).
33. Williams, J. B. W. & Kobak, K. A. Development and reliability of a structured interview guide for the Montgomery Asberg Depression Rating Scale (SIGMA). *Br. J. Psychiatry.* **192**, 52–58 (2008).
34. Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. The development and psychometric properties of LIWC2015 https://repositories.lib.utexas.edu/items/705e81ca-940d-4c46-94ec-a52ffdc3b51f.
35. TF-IDF. In Sammut, C. & Webb, G. I. (eds.) *Encyclopedia of Machine Learning*, 986-987 (Springer US). https://doi.org/10.1007/978-0-387-30164-8_832.
36. Zeng, X., Yang, C., Tu, C., Liu, Z. & Sun, M. Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention. In *Proc. AAAI conference on artificial intelligence (AAAI Press)*, Vol. 32, https://doi.org/10.1609/aaai.v32i1.11982 (2018).
37. Eyben, F., Wöllmer, M. & Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. 18th ACM international conference on Multimedia*, MM '10, 1459-1462 (Association for Computing Machinery). https://doi.org/10.1145/1873951.1874246.
38. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
39. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805, http://arxiv.org/abs/1810.04805 (2018).
40. Liu, Y. et al. Roberta: a robustly optimized bert pretraining approach. Preprint at arXiv preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).
41. Cui, Y. et al. Revisiting pre-trained models for Chinese natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 657-668, https://www.aclweb.org/anthology/2020.findings-emnlp.58 (Association for Computational Linguistics, Online, 2020).
42. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**, e2440969 (2024).
43. Stade, E. C. et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. In: *npj Mental Health Research 3.1*, pp. 1-12. https://doi.org/10.1038/s44184-024-00056-z (2024).
44. Xu, S. et al. Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Schizophrenia and Depression. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 225-228 (IEEE, 2019).
45. Li, L., Zhang, X., Zhou, X. & Liu, Z. AutoMIR: Effective zero-shot medical information retrieval without relevance labels. http://arxiv.org/abs/2410.20050.

46. Sadariya, T. & Verma, S. Early prediction and detection of anxiety level using support vector machine. In Swaroop, A., Polkowski, Z., Correia, S. D. & Virdee, B. (eds.). In *Proc. Data Analytics and Management*, 279–291 (Springer Nature, 2023).
47. Harati, A. et al. Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7273-7277 (IEEE, 2021).
48. Dibeklioğlu, H., Hammal, Z., Yang, Y. & Cohn, J. F. Multimodal detection of depression in clinical interviews. *Proc. ACM Int. Conf. Multimodal Interact. ICMI* **2015**, 307–310 (2015).
49. Hettema, J. M., Aggen, S. H., Kubarych, T. S., Neale, M. C. & Kendler, K. S. Identification and validation of mixed anxiety–depression. *Psychol. Med.* **45**, 3075–3084 (2015).
50. Richter, T., Fishbain, B., Markus, A., Richter-Levin, G. & Okon-Singer, H. Using machine learning-based analysis for behavioral differentiation between anxiety and depression. *Sci. Rep.* **10**, 16381 (2020).
51. Liu, K., Doncheff, B. & Warren, S. L. Predictive utility of symptom measures in classifying anxiety and depression: a machine-learning approach. *Psychiatry Res.* **312**, 114534 (2022).
52. Clark, L. A. & Watson, D. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *J. Abnorm. Psychol.* **100**, 316–336 (1991).

Acknowledgements

This work was supported by Tianqiao and Chrissy Chen Institute (TCCI) with the Program of Chen Frontier Lab for AI and Mental Health - Shanghai Mental Health Center (2023-TX-018), the National Natural Science Foundation of China (82071500), the Natural Science Foundation of Shanghai, China (23ZR1454600), the Program of Shanghai Academic/Technology Research Leader (21XD1423300), Shanghai Shen-Kang Hospital Development Center (SHDC12025118, SHDC22025303) to J.C.; the National Social Science Foundation of China (25BKX030) to Q.Z. and the Integrated Innovation Team Project of Shanghai Mental Health Center. We deeply appreciate every participant involved in this study and all the efforts made by TCCI and SMHC colleagues who are not on the author list.

Author contributions

J.C. was the overall principal investigators for the study who conceived the study and obtained financial support, and was responsible for study design and supervised the entire study. S.X., Y.Y. and Y.D. participated in the study design. S.X. developed the large language model methodology, designed the machine learning pipeline, conducted the experiments, and wrote the original draft. Y.Y. and Y.D. performed clinical concept verification, analyzed the data and finalized the manuscript. F.L. and S.Z. developed the large language model methodology, performed data processing, and conducted the experiments. Z.W., T.Y. and H.G. provided technical support and data resources. J.S., X.J., Y.H., Q.Z., M.Z., Y.X. and J.C. supervised the project, acquired funding, and reviewed the final manuscript. N.O., L.B., reviewed the final manuscript. H.T., C.L., Y.L., H.L., Y.M., W.G., H.Q., Y.W. and J.Q. contributed to the participant recruitment and clinical data collection. All authors contributed to the manuscript revision and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44184-025-00175-1.

Correspondence and requests for materials should be addressed to Xun Jiang or Jianhua Chen.