Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Subject-level depression detection from Chinese clinical interview texts requires dataset-specific aggregation

Shengjie Xia

xiashj21@lzu.edu.cn

Lanzhou University

Article

Keywords:

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9258754/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Subject-level depression detection from Chinese clinical interview texts requires dataset-specific aggregation

Shengjie Xia[1,2,*,+]

[1]School of Information Science and Engineering, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, 730000, Gansu, P.R. China
[2]Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou, 730000, Gansu, P.R. China
[*]xiashj21@lzu.edu.cn
[+]Present address: Department of Data and Systems Engineering, The University of Hong Kong, Pok Fu Lam, Hong Kong Island, Hong Kong S.A.R.

ABSTRACT

Depression detection from Chinese clinical interviews remains underexplored due to scarce annotated datasets and the inherent complexity of interviews, which consist of multiple question-response pairs per subject. To address this gap, the DepInter-CN dataset was constructed, containing 50 participants (450 interviews) with subject-level depression labels from M.I.N.I. and PHQ-9 assessments. The task is formulated as subject-level classification, and a lightweight, interpretable framework is proposed featuring question-level encoding, cross-question aggregation, and subject-level prediction. Three aggregation strategies are compared: question-level probability averaging, answer-only encoding with mean pooling, and joint question-answer encoding with mean pooling. On DepInter-CN, answer-only encoding with mean pooling achieves the best performance (accuracy $\approx$ 0.86, F1 $\approx$ 0.84, AUC $\approx$ 0.93). On the public Chinese Multimodal Depression Corpus (CMDC) dataset, question-level probability averaging performs best (accuracy $\approx$ 0.98, F1 $\approx$ 0.96, AUC $\approx$ 0.97). Question-type analysis reveals that symptom-oriented and situational inquiries are more informative than picture-description tasks. Interpretability analysis indicates that the model relies on clinically relevant symptom-related language patterns rather than superficial cues. These findings underscore the importance of subject-level modeling for interview-based depression detection and indicate that the optimal aggregation strategy depends on dataset structure.

Introduction

Depression is one of the most prevalent and burdensome mental disorders worldwide. Its high global prevalence and substantial social/economic costs, combined with low treatment rates, make efficient and objective screening especially important[1,2]. In routine clinical practice, assessment still relies heavily on semi-structured interviews and psychometric questionnaires administered by trained clinicians. While such methods (e.g., structured clinical interviews and scales like PHQ-9) are effective gold standards for diagnosis, they are also time-consuming, costly, and inherently subjective[3]. This has motivated growing interest in automated, text-based depression detection approaches that can assist clinicians by analyzing patients' language for depressive cues[4].

Over the past decade, natural language processing (NLP) and machine learning techniques have increasingly been used to identify depression from spoken or written text. Early work focused on detecting signals of depression in social media and online forums using lexical cues and classical machine learning classifiers[5–8]. With the rise of deep learning, researchers began applying CNN and RNN models to users' posts, achieving improved accuracy in distinguishing depressed individuals from controls[9,10]. More recently, large pre-trained language models have driven significant progress in text-based mental health detection. For example, Devlin et al.[11] introduced BERT, which has since been fine-tuned for depression classification tasks, yielding state-of-the-art results in English and other languages. In the Chinese context, Wang et al.[12] demonstrated that fine-tuning BERT and RoBERTa on Weibo social media posts substantially outperformed earlier methods for depression risk prediction in Chinese, especially when further pre-trained on domain-specific data. These advances indicate the promise of language-based depression detection using deep NLP techniques, but also highlight new challenges.

One challenge is that powerful transformer models are resource-intensive and behave as "black boxes," making it hard to interpret their decisions. This is a serious concern in clinical applications, where transparency and trust are essential[13]. As a result, there is growing interest in lightweight and interpretable modeling approaches that can perform robustly on limited clinical text data while offering explainable results. In particular, healthcare NLP researchers have explored using simpler

architectures (e.g. hierarchical attention networks or even linear models) that incorporate clinically meaningful structure and are easier to interpret[14]. Such models can be augmented with modern explainability techniques – for example, using SHAP values to highlight important linguistic features in a patient's speech/text[15] or using dimensionality reduction like UMAP to visualize patient representations[16]. By combining efficient modeling with explainability, the author can better ensure that automated depression detectors are both accurate and clinically transparent.

Another gap in the literature has been the relative lack of Chinese-language datasets and studies for depression detection. Most existing text-based depression corpora and models have focused on English – from Twitter and Reddit posts[9] to the DAIC-WOZ clinical interview dataset in English[17]. In contrast, Chinese resources remain limited. A few recent efforts have begun to fill this gap. Wang et al.[12] constructed a Chinese microblog (Weibo) dataset with depression annotations, and Zhang et al.[18] released an expert-validated Weibo depression dataset, illustrating the feasibility of NLP-based screening in Chinese social media. In the clinical realm, Shen et al.[19] introduced the EATD-Corpus, an Emotional Audio-Textual Depression dataset of 162 Chinese clinical interviews (with transcripts and audio, plus clinician-rated depression scores), and reported baseline results using a GRU/BiLSTM model. Similarly, Zou et al.[20] developed the CMDC – a Chinese Multimodal Depression Corpus – containing semi-structured clinical interviews in Chinese with synchronized text, audio, and video data for each participant, verified by professional clinicians. These datasets have enabled preliminary studies on Chinese-language depression detection. However, they are still relatively small in scale, and differences in interview structure (e.g. fixed script vs. varying dialogue) mean that techniques may not generalize directly across them. Overall, there is a clear need for more Chinese clinical datasets and for methods tailored to Chinese language and cultural context.

The work addresses these needs by constructing a new dataset of Chinese clinical interview texts for depression detection and by proposing a lightweight, explainable modeling framework. the author focuses on interview-based mental health detection, analyzing how different question types and patient responses contribute to identifying depression. Unlike prior Chinese studies that mostly used social media data[12,18] or multimodal cues[20], the author concentrates on transcripts from diagnostic interviews conducted in clinical settings. By formulating depression detection as a subject-level classification (where each subject's collection of answers produces one label) rather than a standard text classification, the author aligns with the way clinicians evaluate a patient across multiple questions. The author shows that this subject-level approach, combined with simple yet effective text encodings and aggregation strategies, yields strong performance even on the limited-size dataset. Furthermore, the author performs extensive explainability analyses – both globally, to visualize the landscape of depressive vs. healthy subjects, and locally, to attribute model predictions to specific questions and words. In summary, this paper presents a new Chinese clinical interview dataset for depression detection, analyzes the role of question types in eliciting depressive signals, and introduces a lightweight, explainable modeling approach that balances accuracy with interpretability for use in real clinical applications.

Related work

A growing body of research has studied how linguistic patterns in a person's writings or speech can indicate depression. Early studies leveraged social media corpora: for example, the work of Benton et al.[8] showed that multi-task learning on Twitter data could jointly predict several mental health conditions, including depression, by sharing text feature representations. Other researchers compiled specialized datasets such as the Reddit Self-reported Depression Diagnosis (RSDD) corpus, enabling supervised classification of depression based on users' forum posts[21]. Overall, these studies found that depressed individuals tend to exhibit distinguishable language usage – e.g. more first-person pronouns, more negative emotion words, and references to loneliness or hopelessness – which can be captured by NLP features[9]. Traditional approaches often used bag-of-words or lexicon-based features fed into SVM or logistic regression classifiers, yielding moderate success (with accuracies in the 0.70–0.80 range)[8,22].

With the advent of deep learning, more powerful text representations have improved depression detection performance. Orabi et al.[9] pioneered applying CNNs on user tweets to detect depression, reporting that deep models outperformed earlier n-gram classifiers. Subsequent work has used recurrent neural networks (LSTMs/GRUs) and attention mechanisms to model sequences of words indicative of depressive affect[9]. For instance, a study by Mallol et al.[23] introduced a hierarchical attention network that first encodes each utterance from a clinical interview and then aggregates across the interview, achieving promising results on the DAIC-WOZ interview transcript dataset. Similarly, Xezonaki et al.[24] proposed an affect-enhanced hierarchical attention model for transcribed interviews, incorporating an affective lexicon to bias the attention weights toward emotionally salient words. These deep models can automatically learn subtle linguistic cues (such as tone, syntax, and context) that correlate with depression, improving detection accuracy.

In recent years, large pre-trained language models have become the state-of-the-art in many NLP tasks, including mental health detection. Models like BERT[11] and its variants provide contextual embeddings that capture rich semantic information, which can be fine-tuned on relatively small depression datasets. In the CLPsych 2019 shared task, for example, fine-tuned transformer models achieved top performance in classifying users' depression status and suicide risk from online text[25]. For

Chinese, Cui et al.[26] released a Chinese BERT model (BERT-wwm) that has been widely used to boost NLP tasks in Chinese. Wang et al.[12] applied BERT-wwm to Chinese Weibo posts and found it substantially improved depression risk prediction, raising F1 scores by 5–10 points over recurrent models. More recently, researchers have explored domain-specific pre-training: Zhang et al.[18] describe a MentalBERT for Chinese, obtained via further pre-training on mental health-related social media data, which achieved higher accuracy on depression detection than general BERT. Overall, leveraging pre-trained transformers has become a common theme, as they can extract nuanced features of depressive language (e.g., subtle changes in self-referential speech or emotional content) that may be missed by older models. However, these models typically have millions of parameters and require caution against overfitting when data is limited – a frequent scenario in clinical NLP.

Beyond social media, an important line of work focuses on clinical interviews – a more controlled and diagnostically relevant setting. The Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) is a seminal resource in this area[17]. DAIC-WOZ contains audio, video, and transcripts from 189 semi-structured interviews designed to diagnose psychological distress (depression, PTSD, etc.), conducted by a virtual agent. It has spurred numerous multimodal depression studies; for example, the AVEC 2016 challenge used DAIC-WOZ as a benchmark, and participants developed models to predict depression severity from participants' spoken answers, facial expressions, and vocal prosody[4]. In that challenge, the best approaches fused text and audio features, underscoring that linguistic cues in answers carry significant signal for depression even on their own. Subsequent works on DAIC-WOZ transcripts include the aforementioned hierarchical attention models[23,24], as well as graph-based models: Niu et al.[27] introduced HCAG, a hierarchical context-aware graph attention network that explicitly modeled the interactions between the interviewer's questions and the participant's responses. By encoding question-answer pairs and their context, HCAG improved depression classification accuracy, suggesting that how a patient responds to specific types of questions is diagnostically informative. Our work builds on this insight by analyzing question-type contributions in Chinese interviews.

In the Chinese language context, until recently there were no comparable interview corpora. Shen et al.[19] filled part of this gap with the EATD-Corpus, which contains 162 interviews of Chinese university volunteers, each with transcripts, audio recordings, and an associated depression score (from the Zung Self-Rating Depression Scale). They reported baseline classification results on EATD using a fused GRU/BiLSTM model (text and audio)[19]. Concurrently, Zou et al.[20] presented the CMDC (Chinese Multimodal Depression Corpus), composed of clinical interviews with 78 participants in China (26 depressed patients and 52 controls) recorded in text, audio, and video. Using early fusion of text features (from a Chinese BERT model) with acoustic (eGeMAPS) and facial features, they achieved an impressive F1-score of 0.94 on depression detection[20]. These efforts demonstrate the feasibility of automatic depression assessment from Chinese interviews, though each has limitations (e.g., EATD involves only students and uses self-report scales; CMDC includes multimodal data but with a somewhat fixed interview script). Notably, both corpora remain relatively small. As a result, researchers have emphasized techniques like data augmentation or transfer learning to bolster performance, and have kept models relatively simple to avoid overfitting. For instance, Shen et al.[19]'s baseline model was a two-layer RNN, and Zou et al.[20] relied on feature-based classifiers, whereas very deep models could not be reliably trained on these datasets.

In high-stakes domains like mental health, model interpretability is nearly as important as predictive accuracy[13]. Clinicians are more likely to trust and adopt AI tools if they can understand the reasoning behind a prediction. Therefore, recent studies have integrated explainable AI (XAI) techniques into depression detection workflows. For example, researchers have used SHAP (Shapley Additive Explanations) to interpret text-based depression classifiers by identifying which words or behavioral features most contributed to a model's decision[15]. Mallol et al.[23] and Xezonaki et al.[24] both examined attention weight distributions in their models to see which parts of an interview the model focused on for depressed vs. non-depressed individuals. More broadly, Alghowinem et al.[28] provide a taxonomy of XAI methods and argue that transparent, human-interpretable models should be prioritized in healthcare AI applications. In line with this, some researchers intentionally choose lightweight models (with fewer parameters and straightforward mechanisms) for clinical text tasks, since these are easier to interpret and less prone to overfitting on small data[28]. For instance, logistic regression and decision trees have been revisited for mental health classification due to their inherent interpretability, sometimes augmented with topic models or lexicon features to provide explanations[29].

Another powerful tool for explainability is UMAP (Uniform Manifold Approximation and Projection), a nonlinear dimensionality reduction technique. UMAP can project high-dimensional text embeddings (for each subject or document) into a 2D space, enabling visual inspection of how depressed and healthy subjects cluster apart[16]. Such visualization can reveal whether the model's representations of patients align with clinical distinctions, and can uncover subgroups or outliers of interest. Prior work in health NLP hasn't used techniques like t-SNE or UMAP to verify that model embeddings capture meaningful structure (e.g., grouping patients by symptom severity). In our method, the author leverages UMAP to create a "global" view of the subject-level embedding space and confirm that depressed patients form a distinguishable cluster from controls. the author also applies SHAP at the "local" level to compute token-level importance scores for individual predictions, thus providing human-interpretable explanations for each subject's classification. By combining these approaches with a simpler neural

architecture, our aim is to produce a system that clinicians can both use and understand.

Methods

Dataset Construction and Characteristics

The paper constructed a Depression Interview Corpus in Chinese (DepInter-CN) by collecting and transcribing clinical interviews from a partner hospital in China. Each interview was a semi-structured session between a clinician and a patient (or healthy control) following a standardized protocol. The protocol included both medical inquiry questions (e.g., asking about symptoms, sleep patterns, appetite, mood) and picture-based elicitation tasks (where patients describe what they see in an image, used to probe thought content and emotional response). The author recruited 50 participants (25 diagnosed with major depressive disorder and 25 non-depressed individuals matched by age and gender). All participants gave informed consent, and the study was approved by the hospital's ethics board. Clinical diagnoses were established using the MINI International Neuropsychiatric Interview (MINI)[30] and corroborated by a PHQ-9[31] depression scale score for each patient. The paper uses these clinician-assessed labels as ground truth: label=1 for depressed (if the participant met DSM-5[32] criteria for MDD, typically corresponding to PHQ-9 $\geqslant$ 10) and label=0 for healthy controls. The interviews were audio-recorded and then transcribed into text by professional transcribers. The transcripts were manually reviewed to ensure accuracy and to remove any personally identifying information.

Our final DepInter-CN dataset contains 450 interview response units from 50 subjects. Each subject contributes on average 9 question-response pairs (since the interview script has 9 main questions/tasks, asked to every participant in the standardized protocol). The dataset is balanced with respect to class (25 depressed vs 25 control subjects). In terms of text length, most individual responses are fairly short: the majority are under 150 Chinese characters. This reflects the interview style – questions are often answered in a few sentences. Overall, DepInter-CN represents a clinically grounded, structured sample of Chinese therapeutic interviews, providing a new resource for depression detection research. To our knowledge, this is among the first openly available Chinese-language clinical interview datasets for mental health NLP.

In addition to DepInter-CN, the author utilizes the public CMDC dataset introduced by Zou et al.[20] for external validation. As noted in Related Work, CMDC is a Chinese multimodal depression corpus collected via semi-structured interviews in a clinical setting[20]. Importantly, the structure of CMDC interviews differs from DepInter-CN: CMDC interviews are less standardized, with each participant potentially asked a different number of questions depending on the clinical context (i.e. the interview length varies). This yields an average of 12 response units per subject in CMDC (range 7 to 12), compared to the fixed 9 per subject in DepInter-CN. Additionally, CMDC transcripts tend to be longer (many responses are paragraph-length). By evaluating on both DepInter-CN and CMDC, the paper can test the robustness of our aggregation strategies and see if the optimal approach changes when interview format changes.

Dataset statistics

As shown in Figure 1 and 2, DepInter-CN follows a relatively fixed interview structure, where each participant is associated with nine question-response units. In contrast, CMDC contains a variable number of available interview units per participant. This distinction is central to the later comparison because the structural regularity of the dataset may affect the relative advantage of representation-level aggregation and decision-level averaging.

As further illustrated in the current manuscript, the text-length distributions of the two datasets are also substantially different. In DepInter-CN, most texts are relatively short and concentrated within a compact range, whereas CMDC contains much longer texts and a broader distribution. This difference directly motivates the use of different maximum sequence lengths in later experiments.

Task formulation

Unlike conventional sample-level text classification, this study formulates depression detection as a subject-level binary classification problem. For each participant s, the input is a set of interview responses:

$$\mathscr{X}_s = \{x_{s,1}, x_{s,2}, \ldots, x_{s,n_s}\}, \tag{1}$$

where n_s denotes the number of available question-response units for participant s. The prediction target is a binary label $y_s \in \{0,1\}$, where 0 denotes healthy control and 1 denotes depression.

This formulation is more consistent with the clinical setting because labels are assigned at the participant level rather than at the individual response level. It also provides a cleaner basis for evaluation by reducing the risk of subject leakage and by explicitly modeling multiple interview responses as joint evidence for participant-level prediction.

Notation and overall framework

Let s denote a subject, and let

$$\{(q_{s,j}, a_{s,j})\}_{j=1}^{n_s}$$

denote the set of available interview units for subject s, where $q_{s,j}$ and $a_{s,j}$ are the j-th interview question and answer, respectively, and n_s is the number of question-response units associated with subject s. The prediction target is a binary subject label

$$y_s \in \{0, 1\},$$

where 0 denotes healthy control and 1 denotes depression.

The proposed framework has three stages: question-level encoding, cross-question aggregation, and subject-level prediction. This general formulation is shared by all compared methods, while the main difference lies in how the question-level information is aggregated into a subject-level prediction.

Question-level encoding

Two input formulations are considered. In answer-only encoding, the j-th input of subject s is

$$x_{s,j}^{(A)} = a_{s,j}. \tag{2}$$

In question-answer joint encoding, the corresponding input is

$$x_{s,j}^{(QA)} = [q_{s,j}; a_{s,j}], \tag{3}$$

where $[\cdot;\cdot]$ denotes textual concatenation.

Each input sequence is fed into the same ERNIE (a Chinese pre-trained language model similar to BERT) encoder[33], and the final sentence embedding is obtained by mean pooling over the contextual token representations:

$$\mathbf{h}_{s,j}^{(\cdot)} = \frac{1}{L_{s,j}} \sum_{\ell=1}^{L_{s,j}} \mathbf{H}_{s,j,\ell}^{(\cdot)} \in \mathbb{R}^d, \tag{4}$$

where $\mathbf{H}_{s,j,\ell}^{(\cdot)}$ is the hidden representation of the ℓ-th token, $L_{s,j}$ is the token length of the encoded sequence, and d is the embedding dimension.

Thus, the model first converts each interview unit into a fixed-dimensional vector representation, which serves as the basic information unit for all downstream aggregation strategies.

Cross-question aggregation strategies

Three aggregation strategies are investigated.

Answer-only representation mean pooling

In the `answer_meanpool_lr` setting, each answer is first encoded into $\mathbf{h}_{s,j}^{(A)}$. The subject-level representation is then defined as the arithmetic mean of all available answer embeddings:

$$\mathbf{z}_s^{(A)} = \frac{1}{n_s} \sum_{j=1}^{n_s} \mathbf{h}_{s,j}^{(A)}. \tag{5}$$

This representation is finally mapped to a depression probability by logistic regression:

$$p_s^{(A)} = \sigma\left(\mathbf{w}_A^\top \mathbf{z}_s^{(A)} + b_A\right), \tag{6}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function.

This strategy is a representation-level aggregation method, because question-level embeddings are combined before the final classifier is applied.

Question-answer representation mean pooling

In the `qa_meanpool_lr` setting, each question-answer pair is encoded into $\mathbf{h}_{s,j}^{(QA)}$, and the subject-level representation is obtained by

$$\mathbf{z}_s^{(QA)} = \frac{1}{n_s} \sum_{j=1}^{n_s} \mathbf{h}_{s,j}^{(QA)}. \tag{7}$$

The final probability is then

$$p_s^{(QA)} = \sigma\left(\mathbf{w}_{QA}^{\top} \mathbf{z}_s^{(QA)} + b_{QA}\right). \tag{8}$$

Compared with answer-only mean pooling, this method differs only in the definition of the question-level input. The aggregation mechanism itself remains unchanged.

Question-level prediction with subject-level probability averaging

In the `question_probavg_lr` setting, each answer embedding $\mathbf{h}_{s,j}^{(A)}$ is first passed through a question-level logistic regression classifier:

$$p_{s,j}^{(Q)} = \sigma\left(\mathbf{w}_Q^{\top} \mathbf{h}_{s,j}^{(A)} + b_Q\right). \tag{9}$$

The subject-level probability is then computed by averaging question-level probabilities:

$$p_s^{(Q)} = \frac{1}{n_s} \sum_{j=1}^{n_s} p_{s,j}^{(Q)}. \tag{10}$$

This method is therefore a decision-level aggregation strategy, because the aggregation is performed on predicted probabilities rather than on intermediate representations.

Subject-Level Prediction and Training Protocol

For all three methods, the final predicted subject label is obtained by thresholding the subject-level probability:

$$\hat{y}_s = \mathbb{I}(p_s \geq \tau), \qquad \tau = 0.5, \tag{11}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

To avoid subject leakage, training and evaluation are performed under subject-independent grouped cross-validation, where all interview units from the same subject remain in the same split. This protocol is consistent with the subject-level task formulation and ensures that the model is evaluated on previously unseen subjects rather than on unseen answers from previously observed subjects.

Explainability

Explainability analysis is developed for the best-performing private-dataset method, `answer_meanpool_lr`, and is organized into two complementary levels: global representation analysis and local attribution analysis. The first level examines whether the subject-level representation space is geometrically aligned with subject labels and predictions. The second level analyzes which questions, spans, and tokens contribute most strongly to the final subject-level decision.

Global representation analysis

Let $\mathbf{z}_s^{(A)}$ denote the subject-level answer embedding obtained by mean pooling. Before visualization, all subject representations are standardized:

$$\tilde{\mathbf{z}}_s = \mathbf{D}^{-1}\left(\mathbf{z}_s^{(A)} - \mu\right), \tag{12}$$

where μ and $\mathbf{D}$ are the empirical mean and diagonal scaling matrix estimated from the subject embedding table.

A two-dimensional manifold projection is then computed using UMAP:

$$\mathbf{u}_s = \mathscr{U}(\tilde{\mathbf{z}}_s) \in \mathbb{R}^2. \tag{13}$$

To ensure consistency with subject-level evaluation, out-of-fold probabilities are aggregated across repeated folds:

$$\bar{p}_s = \frac{1}{|\mathscr{R}_s|} \sum_{r \in \mathscr{R}_s} p_s^{(r)}, \tag{14}$$

where $\mathscr{R}_s$ is the set of out-of-fold predictions collected for subject s. The corresponding aggregated predicted label is

$$\bar{y}_s = \mathbb{I}(\bar{p}_s \geq 0.5). \tag{15}$$

The global visualization is then constructed from the triplet

$$(\mathbf{u}_s, y_s), \qquad (\mathbf{u}_s, \bar{y}_s), \qquad (\mathbf{u}_s, \bar{p}_s),$$

which respectively correspond to the ground-truth labels, aggregated predicted labels, and aggregated depression probabilities. This design allows the representation space, classifier outputs, and probability confidence to be examined in a unified geometric view.

Local attribution analysis

For `answer_meanpool_lr`, the subject-level logit is

$$f_s = \mathbf{w}_A^\top \mathbf{z}_s^{(A)} + b_A. \tag{16}$$

Let $\bar{\mathbf{z}}$ denote the empirical mean of the training subject embeddings. Since the subject representation is the average of question-level embeddings, the logit can be decomposed into question-level contributions:

$$f_s = b_A + \mathbf{w}_A^\top \bar{\mathbf{z}} + \sum_{j=1}^{n_s} c_{s,j}, \tag{17}$$

where the contribution of question j is

$$c_{s,j} = \frac{1}{n_s} \mathbf{w}_A^\top \left(\mathbf{h}_{s,j}^{(A)} - \bar{\mathbf{z}} \right). \tag{18}$$

This decomposition makes it possible to quantify how each answer contributes to the final subject-level decision under a linear classifier.

To further trace subject-level evidence back to raw text, a centered contribution predictor is defined for each answer:

$$g_{s,j}(x) = \frac{1}{n_s} \mathbf{w}_A^\top \left(\phi(x) - \bar{\mathbf{z}} \right), \tag{19}$$

where $\phi(x)$ denotes the ERNIE sentence embedding of raw answer text x.

A SHAP text explainer is then applied to $g_{s,j}$, producing token-level attributions:

$$g_{s,j}(x) = \phi_0 + \sum_{m=1}^{M_{s,j}} \varphi_{s,j,m}, \tag{20}$$

where $\varphi_{s,j,m}$ is the SHAP value of the m-th token and $M_{s,j}$ is the number of attributed text units.

Adjacent tokens with the same attribution direction are merged into spans. For span k,

$$C_{s,j,k} = \sum_{m \in \mathscr{S}_{s,j,k}} \varphi_{s,j,m}, \tag{21}$$

where $\mathscr{S}_{s,j,k}$ is the set of token indices belonging to that span.

To identify stable explanation patterns across subjects, normalized tokens and spans are aggregated over samples. For a normalized unit u, let $S_{\text{abs}}(u)$ denote the cumulative absolute attribution over all occurrences and let $N(u)$ denote the number of distinct samples in which u appears. The aggregated impact score is defined as

$$\text{Impact}(u) = S_{\text{abs}}(u) \log\left(1 + N(u)\right). \tag{22}$$

Units are ranked separately according to attribution direction. Positive units indicate linguistic evidence that pushes the prediction toward the depression class, whereas negative units indicate evidence that shifts the prediction toward the healthy-control class. In addition, high-impact positive spans can be embedded again and clustered to reveal higher-level semantic motifs shared across subjects.

Overall, the explainability module provides a two-level interpretation of the proposed framework: the global level evaluates whether subject representations are geometrically consistent with predictions, while the local level identifies the linguistic units that contribute most strongly to the final subject-level decision.

Experimental Settings

As illustrated in Figure 3, the proposed framework performs depression detection at the subject level based on multiple question-answer units from Chinese clinical interviews. All experiments were conducted under repeated 5-fold subject-level stratified cross-validation with five random seeds. This protocol was adopted to reduce the instability of single-run evaluation in a small-sample setting and to provide a more reliable estimate of performance across different subject splits. To avoid subject leakage, all interview units from the same participant were kept in the same fold.

ERNIE was used as the encoder in all experiments. The maximum input length was set to 150 for DepInter-CN and 512 for CMDC, which is consistent with the substantial difference in text-length distributions between the two datasets. Performance was evaluated using Accuracy, Precision, Recall, F1-score, and AUC. Since the task is formulated at the participant level, all reported metrics are computed from subject-level predictions rather than from individual interview units.

Results and analysis

To evaluate the proposed subject-level framework, the author focuses on three lightweight subject-level strategies. In addition, several earlier sample-level methods are retained only as historical comparison references rather than as the methodological focus of the present study.

Subject-level lightweight methods

The first group of compared methods corresponds to the three subject-level strategies investigated in this study.

Question-level prediction with subject-level probability averaging

Each question-response unit is encoded independently, and a question-level classifier predicts a depression probability for each unit. The final subject-level prediction is obtained by averaging the probabilities of all available units from the same participant. This method serves as a decision-level aggregation baseline.

Answer-only encoding with representation mean pooling and subject-level logistic regression

Each response is encoded independently using ERNIE, and the resulting answer embeddings are mean pooled to form a subject-level representation. A logistic regression classifier is then applied to the aggregated representation. This method serves as the primary representation-level aggregation baseline.

Question-answer joint encoding with representation mean pooling and subject-level logistic regression

In this variant, the interview question and its corresponding response are concatenated before encoding. The resulting embeddings are then mean pooled and classified at the subject level. This setting is included to examine whether explicit question context provides additional benefit under the current lightweight framework.

Historical comparison references

Several preliminary sample-level experiments were also conducted, including LR[34], FastText[35], Attention-Hybrid[36], CBA[37], and AMCDD[38]. As shown in Tables 1 and 2, no single sample-level architecture consistently dominates across datasets. On DepInter-CN, FastText achieves the best Accuracy, F1-score, and AUC, whereas CBA yields the highest Recall. On CMDC, LR achieves the best Accuracy, Recall, F1-score, and AUC, whereas CBA achieves the highest Precision.

These results suggest that under the earlier sample-level formulation, model performance remains sensitive to dataset characteristics and does not support a uniformly superior architecture. For this reason, the present study shifts the focus from sample-level comparison to a more clinically appropriate subject-level formulation.

Main results on DepInter-CN

Table 3 summarizes the repeated subject-level results on DepInter-CN. Among the compared methods, answer mean pooling + logistic regression achieves the best overall performance on most metrics, reaching an Accuracy of 0.8600 ± 0.1190, a Precision of 0.9100 ± 0.1280, an F1-score of 0.8378 ± 0.1609, and an AUC of 0.9296 ± 0.0827. These results indicate that a lightweight subject-level representation formed by averaging answer embeddings captures clinically relevant information effectively in the private dataset.

Question-answer mean pooling performs slightly worse overall, suggesting that explicit question context does not yield consistent gains under repeated evaluation. Probability averaging attains the highest Recall (0.8400 ± 0.2000) but remains weaker than answer mean pooling on Accuracy, Precision, and F1. The overall pattern indicates that when the interview structure is regular and each participant contributes a fixed set of units, representation-level aggregation is more effective than decision-level averaging.

External validation on the public dataset

Table 4 reports the external validation results on CMDC. Here the best-performing strategy is question-level probability averaging, which achieves an Accuracy of 0.9790 ± 0.0367, a Precision of 1.0000 ± 0.0000, an F1-score of 0.9633 ± 0.0655, and an AUC of 0.9716 ± 0.0604. Answer mean pooling remains strong on this dataset, but it is consistently below probability averaging on the main metrics.

The contrast between Tables 3 and 4 is one of the author's central findings. DepInter-CN has a fixed interview structure with nine units per subject, whereas CMDC contains a variable number of units and much longer texts. The results therefore suggest that the best aggregation strategy is dataset-dependent: representation-level pooling is advantageous when subject structure is regular, whereas decision-level averaging is more robust when the quantity and composition of subject-level evidence vary across participants.

Question-type analysis on DepInter-CN

Question-type analysis on DepInter-CN in Table 5 and Figure 4 further shows that interview design matters. The question concerning recent physical condition and its impact on life (Q4) yields the strongest single-question performance, with an Accuracy of 0.8800 ± 0.1000, an F1-score of 0.8903 ± 0.0884, and an AUC of 0.9280 ± 0.0712. The self-evaluation question and the recent-emotion question also perform relatively well. In contrast, picture-description tasks are generally less informative, especially the neutral picture-description task (Q8), which produces the weakest results overall.

This pattern indicates that symptom-oriented and situational medical inquiry questions elicit more diagnostically useful language than picture-description tasks. Interview structure should therefore be treated as part of the modeling problem rather than as a superficial formatting detail.

Explainability analysis

The explainability analysis focuses on `answer_meanpool_lr` because it is the best model on DepInter-CN. At the global level in Figure 5, UMAP projections of the subject-level answer embeddings show a meaningful geometry that aligns with both the true labels and the out-of-fold predictions: control samples cluster mainly in two regions, whereas depression samples lie along an intermediate curved band, and the probability map reveals a smooth transition between low- and high-risk areas. This supports the view that the classifier is exploiting structure already present in the latent subject representation.

At the local level in Figures 6 and 7, the model relies primarily on semantically coherent symptom-related expressions rather than isolated sentiment words. Positive attributions are associated with reduced motivation, negative affect, sleep-related complaints, somatic discomfort, and difficulties in family or interpersonal communication. Negative attributions more often reflect preserved daily functioning, relatively normal condition, or comparatively positive self-evaluation. Span-level explanations are more interpretable than token-level explanations because they preserve local context.

Discussion

The first implication of the study is that interview-based depression detection should be evaluated at the subject level. In real clinical settings, labels are assigned to participants, not to isolated answers, and multiple interview units jointly characterize the same subject. A subject-level formulation therefore better matches the diagnostic setting and provides a cleaner evaluation protocol by preventing leakage across answers from the same participant.

The second implication is that aggregation strategy should be matched to dataset structure. When the interview protocol is fixed and regular, as in DepInter-CN, averaging question-level representations is an efficient way to summarize the participant as a whole. When the number and type of available interview units vary across participants, as in CMDC, decision-level probability averaging becomes more robust because it does not require the pooled representation to remain structurally comparable across subjects.

A third finding is that question-answer joint encoding does not consistently outperform answer-only encoding. Under the current small-sample setting, adding question context introduces extra variability but does not reliably improve the final subject-level representation. This suggests that explicit question context should be treated as a design choice that requires empirical validation rather than as a default enhancement.

The study also reinforces the practical value of lightweight modeling. The strongest subject-level results on both datasets are achieved without highly parameterized end-to-end architectures. Instead, careful combinations of a strong encoder, simple aggregation, and logistic regression produce competitive and interpretable results. This is particularly attractive in healthcare NLP, where robustness, transparency, and ease of analysis are all important.

Several limitations remain. DepInter-CN is still small at the subject level. The current aggregation mechanisms are deliberately simple and do not model unequal question importance explicitly. In addition, the framework is text-only, whereas real clinical interviews also contain prosodic, acoustic, and interactional cues. Future work should therefore expand clinically

grounded Chinese interview datasets, explore adaptive question-aware aggregation, and integrate multimodal evidence such as speech.

Conclusion

This study investigated depression detection from Chinese clinical interview texts and addressed two central gaps: the scarcity of clinically grounded Chinese resources and the mismatch between interview data and ordinary sample-level classification. the author constructed DepInter-CN, reformulated the task as subject-level classification, and evaluated a lightweight framework based on question-level encoding, cross-question aggregation, and subject-level prediction.

The results support three main conclusions. First, subject-level evaluation is more appropriate for interview-based depression detection because the diagnostic target is the participant rather than the isolated response. Second, the optimal aggregation strategy is dataset-dependent: answer mean pooling performs best on the fixed-structure DepInter-CN dataset, whereas probability averaging is more robust on the irregular CMDC dataset. Third, question-answer joint encoding does not consistently outperform answer-only encoding under repeated evaluation.

Beyond the main classification results, question-type analysis shows that symptom-oriented and situational inquiry questions are more informative than picture-description tasks, while explainability analysis shows that the best model relies on clinically meaningful symptom-related language patterns rather than superficial polarity cues alone. Together, these findings suggest that effective interview-based depression detection depends on the alignment among task formulation, aggregation design, and interview structure, and they support the value of lightweight and interpretable modeling for small-sample clinical NLP.

References

1. Chisholm, D. *et al.* Scaling-up treatment of depression and anxiety: a global return on investment analysis. *The Lancet Psychiatry* **3**, 415–424 (2016).
2. Vos, T. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet* **396**, 1204–1222 (2020).
3. He, L. *et al.* Deep learning for depression recognition with audiovisual cues: A review. *Inf. Fusion* **80**, 56–86 (2022).
4. Valstar, M. *et al.* Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 3–10 (2016).
5. Chung, C. & Pennebaker, J. The psychological functions of function words. In *Social communication*, 343–359 (Psychology Press, 2011).
6. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, vol. 7, 128–137 (2013).
7. Tsugawa, S. *et al.* Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 3187–3196 (2015).
8. Benton, A., Mitchell, M. & Hovy, D. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538* (2017).
9. Orabi, A. H., Buddhitha, P., Orabi, M. H. & Inkpen, D. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 88–97 (2018).
10. Ansari, L., Ji, S., Chen, Q. & Cambria, E. Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Comput. Soc. Syst.* **10**, 211–219 (2022).
11. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186 (2019).
12. Wang, X. *et al.* Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis. *JMIR medical informatics* **8**, e17958 (2020).
13. Joyce, D. W., Kormilitzin, A., Smith, K. A. & Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj digital medicine* **6**, 6 (2023).
14. Arrieta, A. B. *et al.* Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. fusion* **58**, 82–115 (2020).

15. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

16. McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).

17. Gratch, J. *et al.* The distress analysis interview corpus of human and computer interviews. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 3123–3128 (2014).

18. Zhang, Z. *et al.* Natural language processing for depression prediction on sina weibo: method study and analysis. *JMIR mental health* **11**, e58259 (2024).

19. Shen, Y., Yang, H. & Lin, L. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6247–6251 (IEEE, 2022).

20. Zou, B. *et al.* Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affect. Comput.* **14**, 2823–2838 (2022).

21. Yates, A., Cohan, A. & Goharian, N. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2968–2978 (2017).

22. Stankevich, M., Isakov, V., Devyatkin, D. & Smirnov, I. V. Feature engineering for depression detection in social media. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, 426–431 (2018).

23. Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N. & Schuller, B. W. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. In *Proc. Interspeech 2019*, 221–225 (2019).

24. Xezonaki, D., Paraskevopoulos, G. & Potamianos, A. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Proc. Interspeech 2020*, 4556–4560 (2020).

25. Matero, M. *et al.* Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, 39–44 (2019).

26. Cui, Y. *et al.* Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 657–668 (2020).

27. Niu, M., Chen, K., Chen, Q. & Yang, L. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4235–4239 (IEEE, 2021).

28. Alghowinem, S., Gedeon, T., Goecke, R., Cohn, J. F. & Parker, G. Interpretation of depression detection models via feature selection methods. *IEEE transactions on affective computing* **14**, 133–152 (2020).

29. Villatoro-Tello, E., Ramírez-de-la Rosa, G., Gática-Pérez, D., Magimai.-Doss, M. & Jiménez-Salazar, H. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In *Proceedings of the 2021 international conference on multimodal interaction*, 557–566 (2021).

30. Sheehan, D. V. *et al.* The mini-international neuropsychiatric interview (m.i.n.i.): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *The J. clinical psychiatry* **59 Suppl 20** (1998).

31. Kroenke, K., Spitzer, R. L. & Williams, J. B. The phq-9: validity of a brief depression severity measure. *J. general internal medicine* **16**, 606–613 (2001).

32. Black, D. W. & Grant, J. E. *DSM-5® guidebook: the essential companion to the diagnostic and statistical manual of mental disorders* (American Psychiatric Pub, 2014).

33. Sun, Y. *et al.* Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137* (2021).

34. Cramer, J. S. The origins of logistic regression. Tech. Rep. No. 02-119/4, Tinbergen Institute discussion paper, Tinbergen Institute, Amsterdam and Rotterdam (2002).

35. Kim, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1746–1751 (2014).

36. Ghosh, T. *et al.* An attention-based hybrid architecture with explainability for depressive social media text detection in bangla. *Expert. Syst. with Appl.* **213**, 119007 (2023).

37. Thekkekara, J. P., Yongchareon, S. & Liesaputra, V. An attention-based cnn-bilstm model for depression detection on social media text. *Expert. systems with applications* **249**, 123834 (2024).

38. Selim, M. M. & Assiri, M. S. Enhancing arabic text-to-speech synthesis for emotional expression in visually impaired individuals using the artificial hummingbird and hybrid deep learning model. *Alex. Eng. J.* **119**, 493–502 (2025).

Acknowledgements

This research was supported by the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences in terms of data collection.

Author contributions statement

S.X. conceived the experiments, drafted the manuscript, analysed the results. All authors reviewed the manuscript.

Data availability statement

The data are available from the corresponding author upon reasonable request.

Additional information

Accession codes

The code and supplementary materials associated with this study are publicly available at https://github.com/xiashj2021/Depression-Detection.

All main experiments were conducted under repeated 5-fold subject-level stratified cross-validation with five random seeds, using subject identifiers as the grouping variable to avoid subject leakage. In the current implementation, the base reproducibility seed was set to 42, and the repeated evaluation used the seed set $\{0,1,2,3,4\}$.

All compared methods used the same pretrained Chinese ERNIE encoder (`nghuyong/ernie-3.0-base-zh`) with mean pooling over the last hidden states. For DepInter-CN, the maximum input length was set to 150 in the current implementation. For CMDC, maximum length was 512 in the corresponding public-dataset experiments. The final classifier in all lightweight settings was logistic regression with `max_iter=1000`, and a fixed probability threshold of 0.5 was used for subject-level prediction.

Global explainability was performed by first standardizing subject-level embeddings and then projecting them into a two-dimensional space using UMAP. The main UMAP settings were `min_dist=0.1`, `n_components=2`, cosine distance, and `random_state=42`. Local explainability was conducted for the `answer_meanpool_lr` method using question-level logit decomposition and text-level attribution analysis. In the current implementation, text-level SHAP explanation used a maximum display size of 15, a maximum evaluation budget of 128, and a batch size of 8. Cross-sample attribution patterns were further summarized through token/span aggregation and optional semantic clustering.

Research funding

This research did not receive funding.

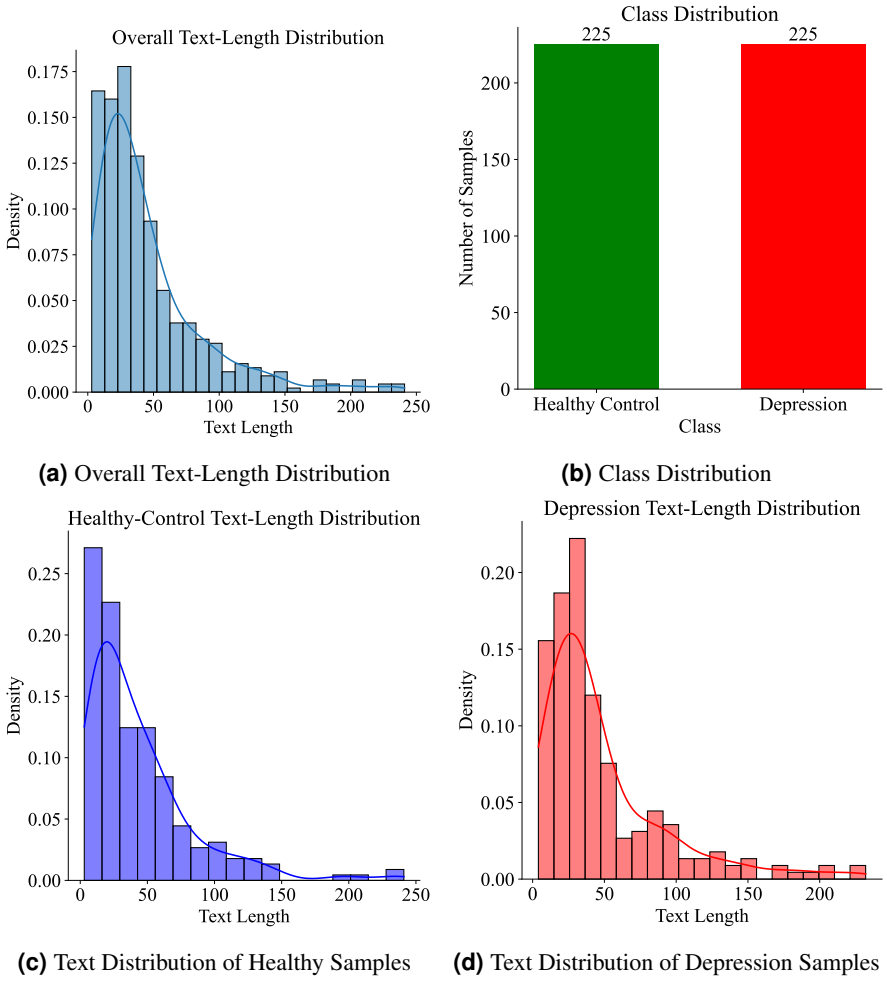
Competing interests

The author declares no competing interests.

Ethics statement

This study has been approved by the ethics committees of Beijing Anding Hospital Capital Medical University, Tianshui Third People's Hospital, and the Second Hospital of Lanzhou University. All patients signed written informed consent forms.

Figures



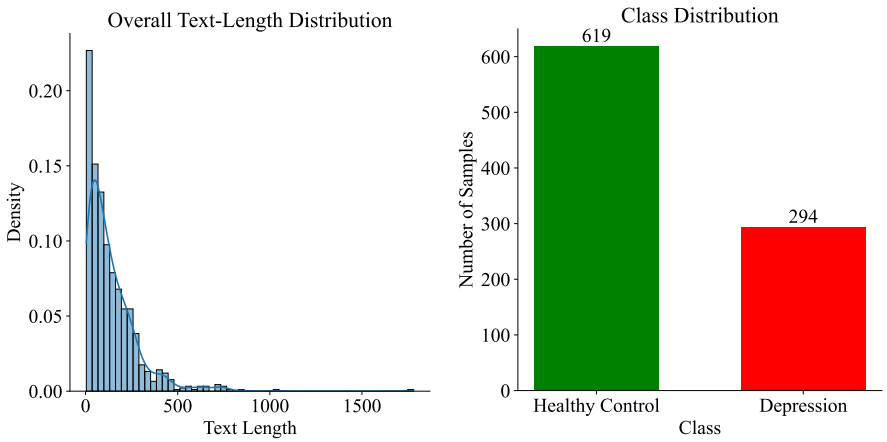
(a) Overall Text-Length Distribution

(b) Class Distribution

(c) Text Distribution of Healthy Samples

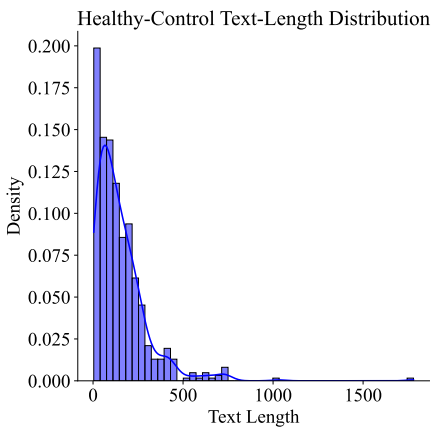
(d) Text Distribution of Depression Samples

Figure 1. Text-length and sample-distribution statistics of DepInter-CN. Panels (a)–(d) show the overall text-length distribution, the class-count distribution, and the class-specific text-length distributions for the two subject groups. The figure illustrates that DepInter-CN is class-balanced and that most interview texts are relatively short, which is consistent with the experimental choice of a moderate maximum sequence length.

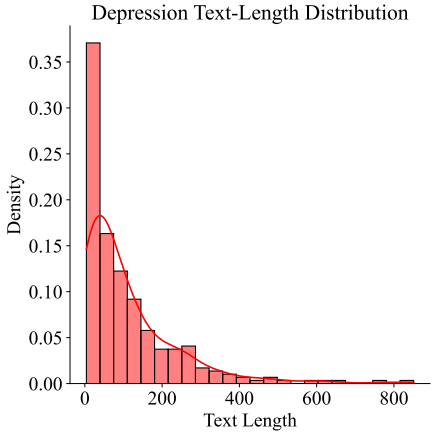


(a) Overall Text-Length Distribution

(b) Class Distribution



(c) Text Distribution of Healthy Samples



(d) Text Distribution of Depression Samples

Figure 2. Text-length and sample-distribution statistics of CMDC. Panels (a)–(d) show the overall text-length distribution, the class-count distribution, and the class-specific text-length distributions in the public dataset. Compared with DepInter-CN, CMDC exhibits substantially longer texts and a broader length distribution, which motivates the use of a larger maximum sequence length in the public-dataset experiments.

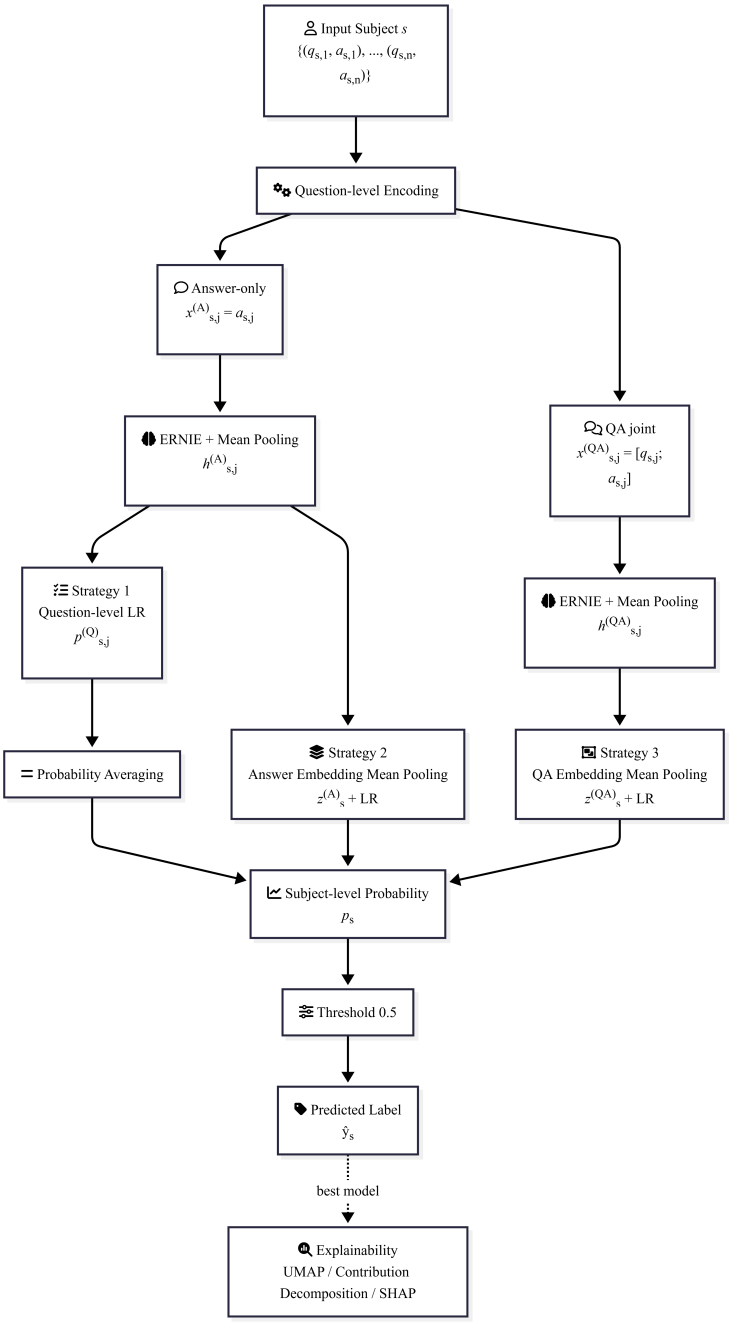


Figure 3. Overview of the proposed subject-level framework for depression detection from Chinese clinical interview texts. Each subject consists of multiple question-answer units. After question-level encoding with ERNIE, subject-level prediction is obtained through three alternative cross-question aggregation strategies: question-level probability averaging, answer-only representation mean pooling, and question-answer representation mean pooling. The best-performing model is further analyzed with global and local explainability methods.

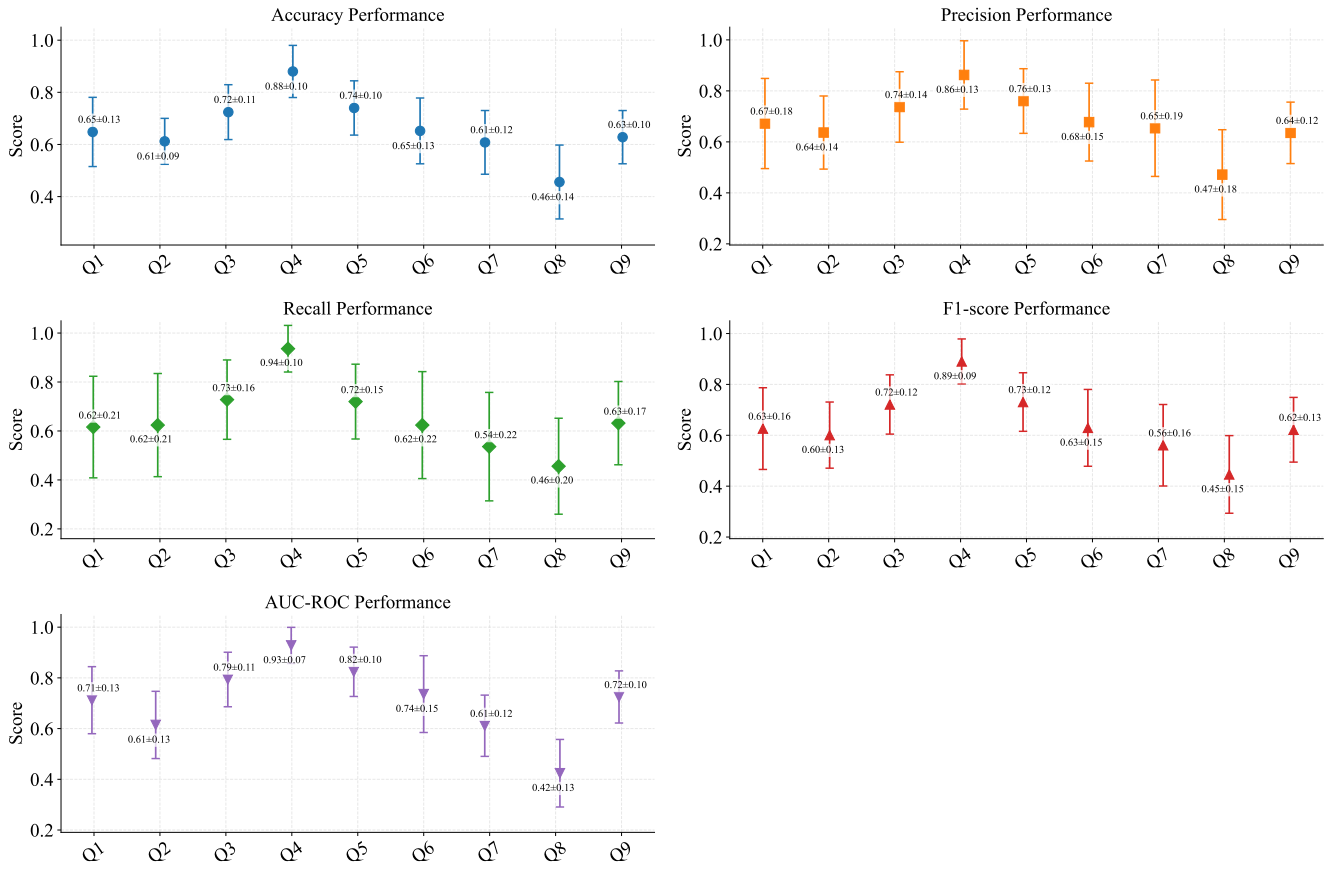


Figure 4. Question-type analysis on DepInter-CN under repeated subject-level evaluation. The figure summarizes the mean and standard deviation of Accuracy, Precision, Recall, F1-score, and AUC for single-question classification across the nine interview units. Each point corresponds to one question, and error bars indicate variability across repeated subject-level cross-validation runs. The results show substantial differences in diagnostic utility across interview questions, with symptom-oriented and situational inquiry questions generally outperforming picture-description tasks.

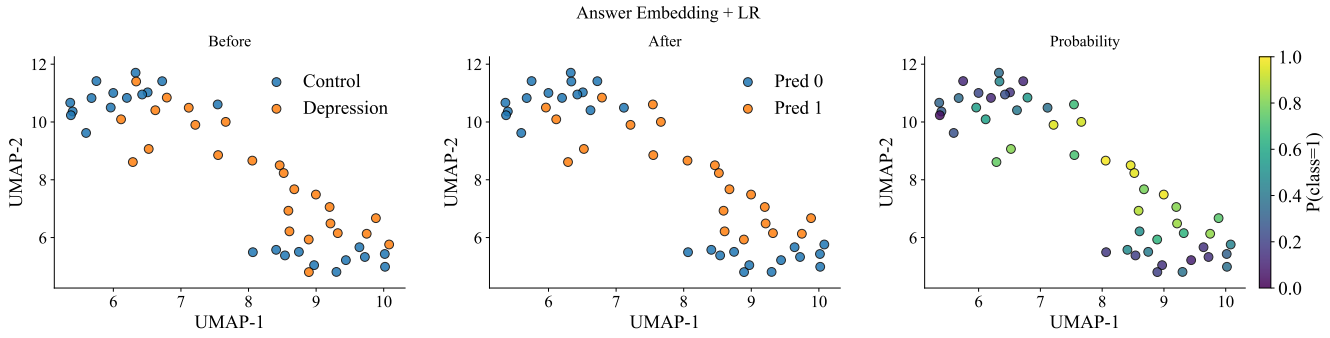
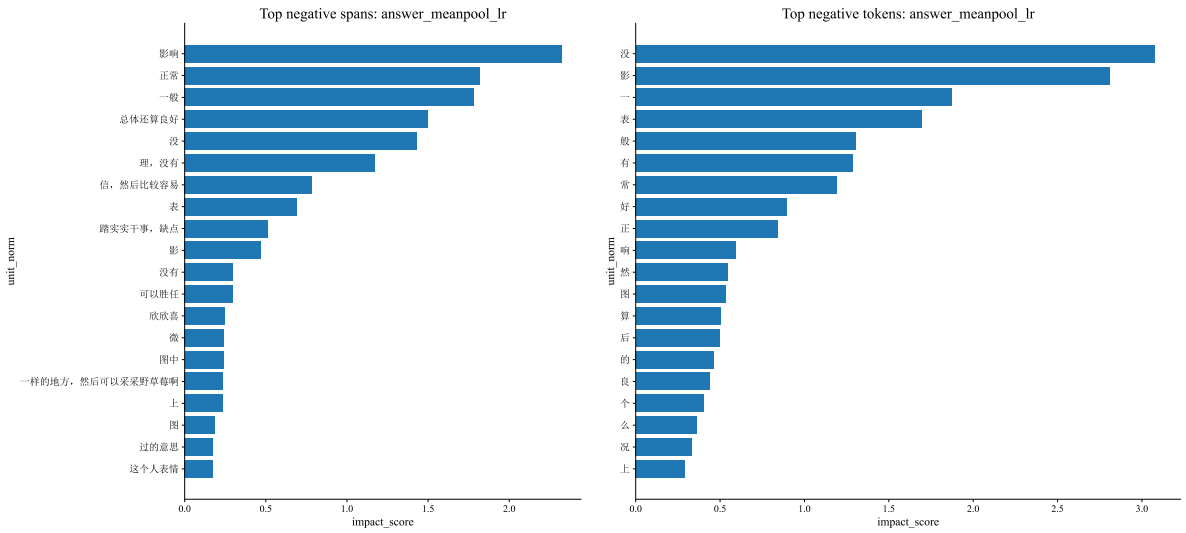


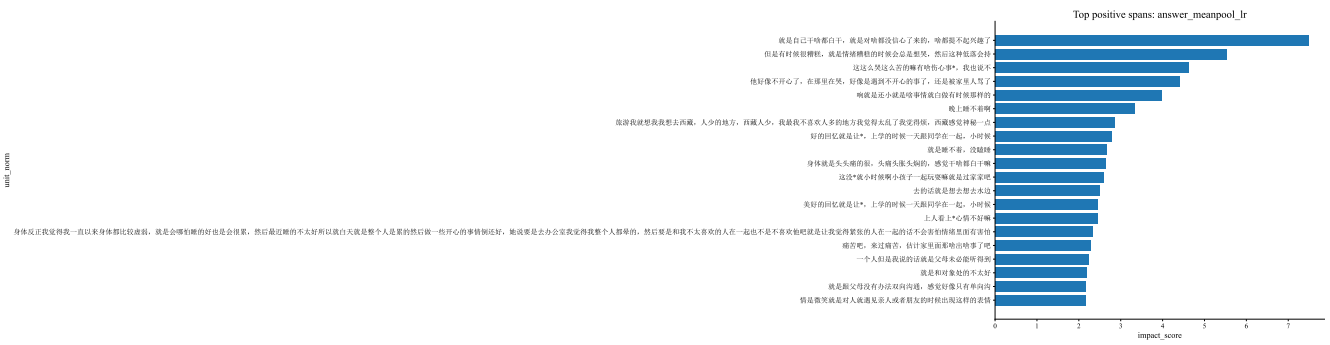
Figure 5. Global representation analysis of the `answer_meanpool_lr` method. The three panels visualize the UMAP projection of subject-level answer embeddings, colored by ground-truth labels (left), aggregated predicted labels (middle), and aggregated depression probabilities (right), respectively. The projection is constructed from out-of-fold subject-level predictions under repeated subject-independent evaluation. The overall alignment between the representation geometry and the prediction patterns indicates that the classifier decisions are supported by the latent answer-derived subject representation.



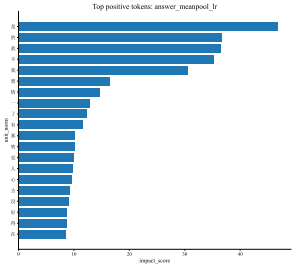
(a) Highest-Impact Negative Spans

(b) Highest-Impact Negative Tokens

Figure 6. Local negative attribution analysis of the `answer_meanpool_lr` method. Panel (a) shows the highest-impact negative spans, and panel (b) shows the highest-impact negative tokens aggregated across subjects. Negative attributions indicate linguistic units that shift the prediction toward the healthy-control class. As in the positive-attribution analysis, span-level explanations are more semantically coherent than token-level explanations, highlighting interpretable language patterns associated with lower depression risk.



(a) Highest-Impact Positive Spans



(b) Highest-Impact Positive Tokens

Figure 7. Local positive attribution analysis of the `answer_meanpool_lr` method. Panel (a) shows the highest-impact positive spans, and panel (b) shows the highest-impact positive tokens aggregated across subjects. Positive attributions indicate linguistic units that increase the model's tendency toward the depression class. The comparison between span-level and token-level results shows that span-level explanations preserve local semantic context more effectively and provide more interpretable evidence for depression-related language patterns.

Tables

Table 1. Repeated sample-level cross-validation results on DepInter-CN. Results are reported as mean ± standard deviation over repeated 5-fold sample-level stratified cross-validation with five random seeds. Best results for each metric are shown in bold.

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
LR	0.6982 ± 0.0401	0.7039 ± 0.0462	0.6907 ± 0.0796	0.6943 ± 0.0484	0.7556 ± 0.0483
FastText	**0.7231 ± 0.0389**	0.7298 ± 0.0427	0.7111 ± 0.0602	**0.7191 ± 0.0428**	**0.7860 ± 0.0460**
Attention-Hybrid	0.7058 ± 0.0483	0.7304 ± 0.0781	0.6853 ± 0.1099	0.6973 ± 0.0502	0.7788 ± 0.0488
CBA	0.7084 ± 0.0463	0.7060 ± 0.0707	**0.7360 ± 0.0815**	0.7157 ± 0.0435	0.7800 ± 0.0578
AMCDD	0.7129 ± 0.0463	**0.7304 ± 0.0521**	0.6862 ± 0.1204	0.7002 ± 0.0692	0.7838 ± 0.0505

Table 2. Repeated sample-level cross-validation results on CMDC. Results are reported as mean ± standard deviation over repeated 5-fold sample-level stratified cross-validation with five random seeds. Best results for each metric are shown in bold.

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
LR	**0.8611 ± 0.0250**	0.7921 ± 0.0451	**0.7735 ± 0.0479**	**0.7819 ± 0.0390**	**0.9181 ± 0.0229**
FastText	0.8324 ± 0.0272	0.7762 ± 0.0588	0.6775 ± 0.0453	0.7226 ± 0.0437	0.8924 ± 0.0239
Attention-Hybrid	0.8559 ± 0.0285	0.8073 ± 0.0570	0.7318 ± 0.0926	0.7637 ± 0.0549	0.9039 ± 0.0327
CBA	0.8580 ± 0.0309	**0.8154 ± 0.0542**	0.7257 ± 0.0860	0.7650 ± 0.0616	0.9050 ± 0.0316
AMCDD	0.8464 ± 0.0331	0.7870 ± 0.0644	0.7244 ± 0.0849	0.7509 ± 0.0569	0.8990 ± 0.0305

Table 3. Repeated subject-level cross-validation results on DepInter-CN. Results are reported as mean ± standard deviation over repeated 5-fold subject-level stratified cross-validation with five random seeds. Best results for each metric are shown in bold.

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
FastText	0.7280 ± 0.1514	0.7471 ± 0.1745	0.7440 ± 0.2123	0.7252 ± 0.1602	0.8160 ± 0.1637
`answer_meanpool_lr`	**0.8600 ± 0.1190**	**0.9100 ± 0.1280**	0.8000 ± 0.2082	**0.8378 ± 0.1609**	**0.9296 ± 0.0827**
`qa_meanpool_lr`	0.8440 ± 0.1227	0.9012 ± 0.1322	0.7920 ± 0.2040	0.8255 ± 0.1447	0.9248 ± 0.0827
`question_probavg_lr`	0.8240 ± 0.1234	0.8367 ± 0.1337	**0.8400 ± 0.2000**	0.8192 ± 0.1366	0.9264 ± 0.0869

Table 4. Repeated subject-level cross-validation results on CMDC. Results are reported as mean ± standard deviation over repeated 5-fold subject-level stratified cross-validation with five random seeds. Best results for each metric are shown in bold.

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
`answer_meanpool_lr`	0.9388 ± 0.0589	0.9607 ± 0.0946	0.8507 ± 0.1441	0.8966 ± 0.1091	0.9603 ± 0.0707
`qa_meanpool_lr`	0.9057 ± 0.0735	0.9132 ± 0.1137	0.8000 ± 0.1851	0.8408 ± 0.1361	0.9446 ± 0.0739
`question_probavg_lr`	**0.9790 ± 0.0367**	**1.0000 ± 0.0000**	**0.9360 ± 0.1114**	**0.9633 ± 0.0655**	**0.9716 ± 0.0604**

Table 5. Repeated single-question subject-level classification results on DepInter-CN. Each model was trained and evaluated using only one interview question at a time. Results are reported as mean ± standard deviation over repeated subject-level cross-validation (50 subjects, 5 stratified folds, and 25 runs per question), allowing a direct comparison of question-specific diagnostic utility. Best results for each metric are shown in bold.

ID	Question Text	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
1	If you had a vacation for traveling, please describe your travel plan.	0.6480 ± 0.1327	0.6720 ± 0.1768	0.6160 ± 0.2075	0.6263 ± 0.1606	0.7120 ± 0.1322
2	Please share a pleasant memory and briefly describe the scene.	0.6120 ± 0.0881	0.6366 ± 0.1434	0.6240 ± 0.2107	0.6007 ± 0.1300	0.6144 ± 0.1326
3	How have you been feeling recently? How has this emotion affected your life?	0.7240 ± 0.1052	0.7369 ± 0.1382	0.7280 ± 0.1621	0.7211 ± 0.1165	0.7936 ± 0.1075
4	How has your physical condition been recently? How has it affected your life?	**0.8800 ± 0.1000**	**0.8625 ± 0.1340**	**0.9360 ± 0.0952**	**0.8903 ± 0.0884**	**0.9280 ± 0.0712**
5	How do you evaluate yourself?	0.7400 ± 0.1041	0.7602 ± 0.1268	0.7200 ± 0.1528	0.7307 ± 0.1151	0.8240 ± 0.0973
6	Please talk about an experience that made you feel very bad. Describe a very painful event.	0.6520 ± 0.1262	0.6775 ± 0.1526	0.6240 ± 0.2185	0.6292 ± 0.1510	0.7360 ± 0.1514
7	Description of a positive-emotion picture.	0.6080 ± 0.1222	0.6536 ± 0.1891	0.5360 ± 0.2215	0.5608 ± 0.1602	0.6112 ± 0.1208
8	Description of a neutral-emotion picture.	0.4560 ± 0.1417	0.4716 ± 0.1761	0.4560 ± 0.1960	0.4461 ± 0.1526	0.4240 ± 0.1332
9	Description of a negative-emotion picture.	0.6280 ± 0.1021	0.6354 ± 0.1205	0.6320 ± 0.1701	0.6217 ± 0.1272	0.7248 ± 0.1028