Predicting Depression in Screening Interviews from Interactive Multi-Theme Collaboration

Xianbing Zhao[*1,4], **Yiqing Lyu**[1], **Di Wang**[2], **Buzhou Tang**[1,3],

[1]Harbin Institute of Technology, Shenzhen, [2]Xidian University,
[3]Peng Cheng Laboratory, [4]Jiangnan University,
zhaoxianbing_hitsz@163.com, kosmischer@stu.hit.edu.cn,
wangdi@xidian.edu.cn, tangbuzhou@gmail.com

Correspondence: tangbuzhou@gmail.com

Abstract

Automatic depression detection provides cues for early clinical intervention by clinicians. Clinical interviews for depression detection involve dialogues centered around multiple themes. Existing studies primarily design end-to-end neural network models to capture the hierarchical structure of clinical interview dialogues. However, these methods exhibit defects in modeling the thematic content of clinical interviews: 1) they fail to explicitly capture intra-theme and inter-theme correlation, and 2) they do not allow clinicians to intervene and focus on themes of interest. To address these issues, this paper introduces an interactive depression detection framework, namely **P**redicting **D**epression in Screening Interviews from **I**nteractive **M**ulti-Theme **C**ollaboration (PDIMC). PDIMC leverages in-context learning techniques to identify themes in clinical interviews and then models both intra-theme and inter-theme correlation. Additionally, it employs AI-driven feedback to simulate the interests of clinicians, enabling interactive adjustment of theme importance. PDIMC achieves absolute improvements of 12% on Recall and 35% on F1-dep. metrics, compared to previous state-of-the-art model on the depression detection dataset DAIC-WOZ, which demonstrates the effectiveness of capturing theme correlation and incorporating interactive external feedback.

1 Introduction

Depression stands as one of the primary factors affecting individuals' mental health (Wei et al., 2022), exerting negative impacts on their work and daily life through its influence on cognitive processes and behavioral patterns. Statistics indicate a year-on-year increase in the prevalence of depressive symptoms within populations in both China and the United States (Rinaldi et al., 2020), a trend that continues to escalate with advancements in detection methods. However, it remains a challenging task as it requires accurately capturing the relationships across hundreds of turns in clinical interview dialogues and predicting the depressive state (Mallol-Ragolta et al., 2019; Yang et al., 2024; Yao et al., 2024). Specifically, clinical interview dialogues are structured around multiple themes, such as family, work, and medical history. These themes exhibit inter-dependencies, and their correlations with depressive states vary in strength. As illustrated in Figure 1, clinical interview dialogues begin with natural dialogues and then evolve around multiple themes (Gratch et al., 2014; Rinaldi et al., 2020), which are closely related to the depressive state of participant.

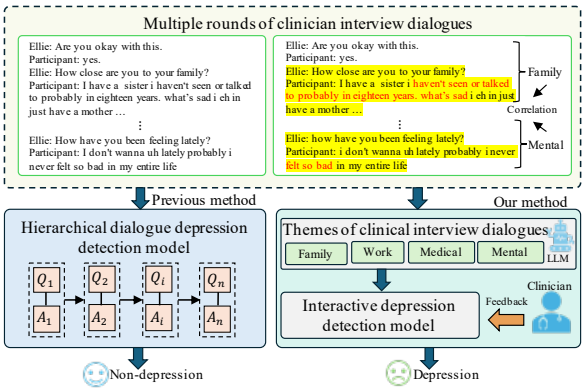

Figure 1: Previous method only focused on modeling the sequential clinical interview dialogue information. Our method learns themes from clinical interview dialogues, models both intra-theme and inter-theme correlation, and finally introduces interactive feedback to guide the model in diagnosing depression.

In recent years, significant efforts have been made to address the aforementioned challenges by modeling the hierarchical structure of clinical interview dialogues. Based on the type of hierarchical neural network employed, existing approaches can be broadly classified into two categories: 1) modeling the hierarchical dependencies in multi-

[*]Remote visiting at Harbin Institute of Technology, Shenzhen

Findings of the Association for Computational Linguistics: ACL 2025, pages 23025–23035

turn question-answering (Wu et al., 2023; Anshul et al., 2024; Zhang et al., 2024). To facilitate the understanding of complex clinical interview dialogues, this approach focuses on capturing the correlation between individual question-answer pairs as well as the sequential dependencies across multiple turns. 2) capturing implicit themes (Gong and Poellabauer, 2017; Rinaldi et al., 2020). This approach relies solely on learnable parameters, modeling each question-answer pair's category as an implicit theme. To smoothly introduce themes, clinical interview dialogues often include dialogues that are unrelated to the depressive state. The former approach tends to trap the model in trivial details, while the latter fails to explicitly identify theme content. Moreover, neither of these methods allows clinicians to intervene in the model to focus on themes of interest.

To tackle these downsides, we introduce a novel interactive depression detection framework, namely **P**redicting **D**epression in Screening Interviews from **I**nteractive **M**ulti-Theme **C**ollaboration (PDIMC), which is the first depression detection framework incorporating explicit theme correlation learning and external feedback intervention. As illustrated in Figure 1, clinical interview dialogues revolve around themes such as *family*, *work*, *mental*, and *medical*. To provide a more comprehensive evaluation, we additionally introduce a virtual theme *overall*. Based on the above observations, we first design a theme-oriented in-context learning (TICL) module to extract theme content from complex clinical interview dialogues, preventing the model from being overwhelmed by trivial details. To capture both intra-theme and inter-theme correlation, we develop a theme correlation learning (TCL) module. Subsequently, we introduce a interactive theme adjustment strategy (ITAS), which leverages the large language model (LLM) to simulate clinician feedback, emphasizing key information from the feedback to further adjust the importance of different themes. Extensive experimental results on the well-known clinical interview depression dataset DAIC-WOZ demonstrate the effectiveness and superiority of our approach. Our main contributions are threefold:

- We introduce an interactive depression detection framework for automated depression detection. To the best of our knowledge, this is the first framework that explores interactive depression detection using clinical interview data.
- We introduce a theme-oriented in-context learning technique to extract themes from clinical interview dialogues and design a theme correlation learning module to model both intra-theme and inter-theme correlation.
- We propose an interactive theme adjustment strategy, which leverages the LLM to simulate clinician feedback, dynamically adjusting the importance of theme. This enables the model to focus on clinician-preferred themes for more effective depression detection.

2 Related Work

2.1 In-context Learning

In-context Learning (ICL) (Brown et al., 2020; Dong et al., 2024) is a technique that uses few-shot in-context learning samples to guide the pre-traind autoregressive LLM to produce satisfying results, without additional training or fine-tuning. The in-context learning samples are usually specifically designed for designated downstream task and can serve as auxiliary parameters attached to the model to guide the generation process. Liu et al. (2022) explored the way to better design the context sample through distance metrics, and impact of different kinds of the distance metrics like Euclidean distance and so on. Levy et al. (2023) proposed a method to reinforce generalization ability through sample diversification. Chung et al. (2022) introduced a perplexity based method for designing in-context samples. Sorensen et al. (2022) conducted cross-lingual contextal learning experiments using clustering methods. Tanwar et al. (2023) used ICL techniques to perform the alignment task of different languages.

2.2 Automatic Depression Detection

Plenty of works have been dedicated to automating the process of depression detection by implementing methods like natural language processing, machine learning, multimodal model LLM. To start with, researchers used traditional methods to tackle with the issue. For instance, Abdurrahim and Fudholi (2024) introduced a Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) based deep learning model to perform depression detecting tasks upon social media posts. Cai et al. (2023) and Misgar and Bhatia (2022) have posed time series based

LSTM to detect suicide risk. Meanwhile, there are also some works (e.g. Wang et al., 2024; Yao et al., 2024; Ying et al., 2024; Dai et al., 2021), combining feature engineering with machine learning algorithms and deep neural network methods for diagnosing mental disorder. Inspired by advances in multimodal sentiment analysis (Jia et al., 2024; Zhao et al., 2022, 2025), a growing number of researchers have applied multimodal learning approaches to the task of multimodal depression detection. Ali et al. (2024) leveraged audio and text modalities to analyze sentiment and mental health, with a method of prompt engineering. Ye et al. (2024) extracted features of audio and video respectively and used the technique of mamba to fuse them and made collaborative classification for depression detection. Jung et al. (2024) proposed HiQuE, hierarchically modeled the question and answer series in the interview dialogues for depression detection. Furthermore, LLMs like BERT, LLaMA and GPT have also been implemented by many works due to its reasoning ability. Lan et al. (2024), Shah et al. (2024), Kuzmin et al. (2024) implemented some prevailing LLMs and got decent performance on social media based dataset. Danner et al. (2023) developed a chat based method using GPT for interactive depression. Yang et al. (2024) proposed MentaLLaMA, a fine tuned version of LLaMA-2, which concentrated on interpretability of issues upon mental disorder.

To the best of our knowledge, interactive depression detection methods based on explicit theme learning have not yet been explored in the field of automated depression detection. Unlike existing related work, our model is the first to learn theme correlation and design an interactive strategy to incorporate clinical feedback for preference learning.

3 METHODOLOGY

Our goal is to learn multiple themes $T_i, i \in \mathcal{D}_{\{family,work,mental,medical,overall\}}$ from multi-turn clinical interview dialogues S, then model both intra-theme and inter-theme correlation $(X_i^{inter}, X_i^{intra}, i \in \mathcal{D})$. Additionally, we introduce feedback simulated by the LLM to imitate feedback of clinician, enabling preference learning to adjust themes. Finally, we fuse the adjusted themes $X_i^{fd}, i \in \mathcal{D}$ to obtain final representation X^{final} for depression prediction $\hat{y}$. In this section, we introduce each component of the proposed model, as illustrated in Figure 2. Specifically, we first present the theme-oriented in-context learning module in Section 3.1, which extracts themes from clinical interview dialogues. Afterwards, we introduce the theme correlation learning module, which captures both intra-theme and inter-theme correlation in Section 3.2. Finally, we describe the interactive theme adjustment strategy in Section 3.3, which enables the model to focus on clinician-preferred themes for depression detection.

3.1 Theme-oriented In-context Learning

Clinical interview dialogues are structured around multiple themes, encompassing both thematic content and trivial details. The thematic content is closely related to the depressive state. To effectively capture these theme-related contents, we design a theme-oriented in-context learning module. This module leverages in-context learning techniques and LLM to extract depression-related themes while discarding irrelevant details. The in-context learning technique guides the LLM to generate output text y_i based on the in-context template I and user input sequence X. Formally, this process can be represented as:

$$P(y_j|S, I) \triangleq p_\theta(X, I), \tag{1}$$

$$\hat{y_j} = argmax P(y_j|S, I), \tag{2}$$

where P represents the token probabilities and $\hat{y_i}$ denotes the token with the highest probability. The operation of the theme-oriented in-context learning technique can be formally summarized as:

$$T_i = p_\theta(S, I), i \in \mathcal{D}, \tag{3}$$

$$\mathcal{D} = \{family, work, mental, medical, overall\}, \tag{4}$$

where the LLM p utilizes in-context prompt I and model parameters θ to extract theme content T from the clinical interview dialogues, filtering out trivial details and preserving information relevant to the depressive state. The in-context template, as illustrated in Figure 2, consists of theme content (family, work, mental, medical, and overall) along with system prompts.

3.2 Theme Correlation Learning

To fully leverage the advantages of themes, we model both intra-theme and inter-theme correlation. The intra-theme correlation aims to capture how key tokens within a theme influence the depressive state, while the inter-theme correlation focuses on learning how semantic relationships across themes

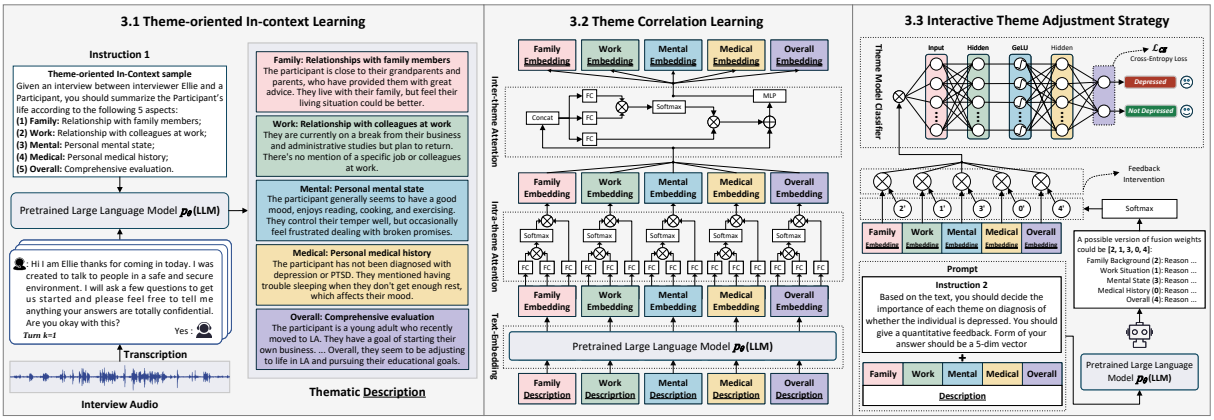

Figure 2: Schematic illustration of the proposed PDIMC framework with three components. The theme-oriented in-context learning technique leverages the LLM to learn theme content from clinical interview dialogues. Theme correlation learning captures the intra- and inter-theme semantics related to depressive states. The interactive theme adjustment strategy utilizes the LLM to simulate clinical feedback, dynamically adjusting theme importance.

contribute to depression assessment. This module enhances the model's ability to highlight key semantics within a theme while also capturing semantic dependencies across themes. Figure 2 illustrates the proposed learning framework, demonstrating how these correlations contribute to a more comprehensive depression detection process. We employ the attention mechanism to learn both intra-theme and inter-theme correlation. Formally,

$$X^* = f_\phi^{corr}(X), * \in \{intra, inter\}, \quad (5)$$

where $f^{corr}(\cdot)$ represents the correlation function with parameter ϕ, while X and X^* denote the input before correlation and the output after correlation, respectively. We first calculate the correlation matrix A. The intra-theme correlation matrix is computed by calculating the token similarity scores, while the inter-theme correlation matrix is determined by calculating the semantic similarity across themes. Formally,

$$A(X) = softmax(\frac{X \cdot W_Q \cdot W_K^T \cdot X^T}{\sqrt{d}}), \quad (6)$$

The computational operation of the correlation function $f^{corr}(\cdot)$ is as follows,

$$f_\phi^{corr}(X) = A(X) \cdot X \cdot W_V, \quad (7)$$

where $A(\cdot)$ represents the correlation matrix function, while $W_{\{Q,K,V\}}$ denotes the learnable parameter matrix of ϕ. We first use the pre-trained large language model p_θ to extract the features X_i of the i-th theme T_i. Formally,

$$X_i = p_\theta(T_i) \in \mathbb{R}^{L_i \times d}, i \in \mathcal{D}, \quad (8)$$

The form of intra-theme correlation learning X_i^{intra} is as follows,

$$X_i^{intra} = f_{\phi_{intra}}^{corr}(X_i) \in \mathbb{R}^{L_i \times d}, \quad (9)$$

where L_i represents the sequence length of the i-th theme, and d denotes the feature dimension. The ϕ_{intra} denotes the learnable parameter matrix. After learning the token correlation semantics within the theme, we concatenate multiple themes along the sequence dimension. Formally,

$$X_{con}^{intra} = Concat\{X_i^{intra}\}_{i=1}^{|\mathcal{D}|} \in \mathbb{R}^{(\sum L_i) \times d}, \quad (10)$$

immediately, we utilize the correlation learning function $f_\phi^{corr}(\cdot)$ to learn the semantic of inter-theme correlation. Formally,

$$X^{inter} = f_{\phi_{inter}}^{corr}(X_{con}^{intra}) \in \mathbb{R}^{(\sum L_i) \times d}, \quad (11)$$

where ϕ_{inter} denotes the learnable parameter matrix. The intra-theme and inter-theme respectively highlight the importance of tokens within a theme and the importance of the theme itself.

3.3 Interactive Theme Adjustment Strategy

To address the issue that depression detection models cannot incorporate feedback from clinician, we have designed a customized interactive theme adjustment strategy. This strategy introduces simulated clinician feedback into the depression detection model, enabling the model to focus on the parts of interest to clinician. When incorporating external feedback information containing categories c to constrain the model output in a generative model

(Dhariwal and Nichol, 2021; Sanchez et al., 2024), the model probability likelihood $\hat{\mathbf{P}}_\Theta$ consists of the model prediction $\mathbf{P}_\Theta$ and the prediction of the feedback information $\mathbf{P}_\Phi$, resulting in the approximated modified distribution,

$$\hat{\mathbf{P}}_\Theta(x|c) \propto \mathbf{P}_\Theta(x) \cdot \mathbf{P}_\Theta(c|x)^\gamma, \tag{12}$$

where γ represents the constraint strength, which controls the model's degree of focus on the constraint. By removing the auxiliary classification task without classifier guidance using Bayes rule, the same model $\mathbf{P}_\Theta$ simultaneously supports both conditional and unconditional predictions to reformulate Equation (12) as: $\mathbf{P}_\Theta(c|x) \propto \frac{\mathbf{P}_\Theta(x|c)}{\mathbf{P}_\Theta(x)}$. The sampling process of classifier-free guidance can be reformulated as:

$$\hat{\mathbf{P}}_\Theta(x|c) \propto \frac{\mathbf{P}_\Theta(x|c)^\gamma}{\mathbf{P}_\Theta(x)^{\gamma-1}}, \tag{13}$$

Taking the logarithm of both sides of Equation (13) results in the following form,

$$\log\hat{\mathbf{P}}(x|c) = \gamma\log\mathbf{P}_\Theta(x|c) - (\gamma-1)\log\mathbf{P}_\Theta(x), \tag{14}$$

let $\mathbb{P}(x) = (1-\gamma)\log\hat{\mathbf{P}}_\Theta(x)$, $\beta = \frac{\gamma}{1-\gamma}$, the probability distribution of the model's prediction can be rewritten as follows,

$$\mathbb{P}(x) = \underbrace{\beta\log\mathbf{P}_\Theta(x|c)}_{\text{feedback prediction}} + \underbrace{\log\mathbf{P}_\Theta(x)}_{\text{vanilla prediction}}, \tag{15}$$

when introducing conditional constraints into the model prediction, the model's prediction is composed of the superposition of the prediction from the vanilla information and the prediction from the conditional constraint information.

Based on the above theory, we integrate the classifier-free feedback of clinician on the theme with the vanilla theme information to diagnose depression. Formally,

$$X_i^{fd} = X_i^{inter} + w_i^{fd} X_i^{inter}, X_i^{fd} \in \mathbb{R}^{L_i \times d}, \tag{16}$$

The integration results for each theme are as follows:

$$X_{final} = \sum_{i=1}^{|\mathcal{D}|} X_i^{fd}, i \in \mathcal{D}, w_i^{fd} \in W^{fd}, \tag{17}$$

where w_{fd} represents the weight of the simulated clinical feedback from the LLM. Finally, we use the $Softmax$ function to normalize the weight of $(1 + W^{fd})$ and integrate the weights of multiple themes for depression detection.

3.4 Depression Detection Layer

The final fused representation X_{final} is passed through several fully connected layers and activation functions to obtain the prediction result $\hat{y}$. We compute the cross-entropy loss between the predicted value $\hat{y}$ and the ground truth y to optimize the model:

$$\mathcal{L}_{CE} = -\sum_{i}^{|N|} \mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1-\mathbf{y}_i)\log(1-\hat{\mathbf{y}}_i) \tag{18}$$

where N denotes the number of training samples.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. DAIC-WOZ (Gratch et al., 2014) is a clinical interview dialogue dataset which is collected and released by the University of Southern California to help veterans back to civilian life. For metrics, we have conducted a series of experiments along with previous state-of-the-art hierarchical dialogue models (Jung et al., 2024) upon the indices of Accuracy, Precision, Recall, F1-score and G-means, and we also report the Precision, Recall, F1-score of Weighted Average version (denoted as WA*Prec., WA*Rec., WA*F1 in Table 1). Meanwhile, to compare against methods of latent theme, we reported F1 score of depression and non-depression on test set, following the previous work (Rinaldi et al., 2020) (shown in Table 2).

Implementation Details. We implemented PDIMC on Nvidia A100 GPU. Applying PyTorch library, our Adam Optimizer is configured with batch size of 32, learning rate of 10^{-5}, and training epochs of 80. For pretrained large language model, we employed Qwen2.5, an open source LLM, to extract text feature.

4.2 Performance Comparison

To demonstrate the effectiveness of our poposed PDIMC, we compared our proposed methods with several baselines. These methods are hierarchical dialogue models and latent theme models respectively. The hierarchical dialogue models includes: TFN (Zadeh et al., 2017), BiLSTM-1DCNN (Lin et al., 2020), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Depression Vlog (D-vlog) (Yoon et al., 2022), BC-LSTM (Poria et al., 2017), ERSDL (Satt et al., 2017), ATSM (Hanai et al.,

Method	Accuracy	Precision	Recall	F1-Score	WA*Prec.	WA*Rec.	WA*F1.	G-Mean
TFN	0.81	0.67	0.72	0.68	0.84	0.78	0.81	0.699
BiLSTM-1DCNN	0.78	0.65	0.61	0.62	0.77	0.71	0.73	0.630
MulT	0.84	0.73	0.74	0.74	0.81	0.77	0.77	0.735
MISA	0.85	0.74	0.77	0.74	0.86	0.77	0.79	0.755
D-vlog	0.84	0.73	0.72	0.73	0.82	0.76	0.77	0.725
BC-LSTM	0.76	0.59	0.60	0.59	0.77	0.69	0.72	0.595
EMSDL	0.80	0.65	0.69	0.66	0.69	0.70	0.71	0.670
ATSM	0.81	0.67	0.71	0.70	0.85	0.73	0.77	0.690
TopicModel	0.78	0.63	0.60	0.62	0.81	0.71	0.74	0.615
CADL	0.83	0.71	0.71	0.71	0.85	0.73	0.77	0.710
Speechformer	0.83	0.70	0.72	0.70	0.78	0.76	0.76	0.710
GRU/BiLSTM	0.86	0.75	0.78	0.75	0.86	0.77	0.80	0.765
HiQuE	0.87	0.78	0.80	0.79	0.85	0.80	0.82	0.790
PDIMC	**0.94**	**0.89**	**0.92**	**0.90**	**0.92**	**0.91**	**0.92**	**0.905**

Table 1: Performance comparison between our proposed PDIMC and several baselines on the DAIC-WOZ dataset.

Category	Model	F1 dep.	F1 non-dep.
Latent Theme (Text)	PR	0.45	0.82
	BERT	0.44	0.84
	JLPC	0.53	0.82
	JLPCPost	0.52	0.85
Explicit Theme	**PDIMC**	**0.87**	**0.92**

Table 2: Performance comparison between our proposed explicit theme model and implicit theme models.

2018), TopicModel (Gong and Poellabauer, 2017), CADL (Lam et al., 2019), Speechformer (Chen et al., 2022), GRU/BiLSTM (Shen et al., 2022), Hierarchical Question Embedding Network (HiQuE) (Jung et al., 2024).The latent theme models include: PR, BERT, JLPC, JLPCPost (all reported in Rinaldi et al., 2020).

The comparison results are summarized in Table 1 and 2. By comparing the results, we could draw the following conclusions. 1) Compared to hierarchical dialogue models, our proposed method PDIMC shows remarkable improvement on all metrics. Among these baseline models, HiQuE (Jung et al., 2024) achieved the best performance in nearly every metric, except weighted average precision (WA*Prec.). For macro metrics, our proposed method achieved 11% better in Precision and F1-Score, 12% better in Recall, compared to HiQuE. 2) Compared to model of latent theme JLPC(Rinaldi et al., 2020), our model PDIMC of explicit theme achieved a 35% improvement in the F1 score of depression. The experimental results indicate that explicit themes provide more clues related to depressive states than implicit themes. This result suggests that the hierarchical dialogue model struggles to learn information related to depressive states when modeling hundreds of clinical interview dialogues. It further underscores the importance of learning themes and incorporating clinical feedback.

In addition, we explore the use of different backbone networks to extract theme-related features and incorporate various LLM to simulate clinician feedback. Notably, models utilizing Qwen2.5-7B and LLaMA3-8B both achieved an F1 score of 0.90, indicating that the framework is relatively robust to variations in textual features. However, when Qwen2.5-7B, Gemma-9B, and LLaMA3-8B are used to simulate feedback, the model attained F1 scores of 0.90, 0.79, and 0.71, respectively. This result highlights that the quality of simulated feedback from LLMs substantially affects model performance, suggesting that the reasoning capability of the LLM plays a critical role in the final prediction.

4.3 Ablation Study

To assess the effectiveness of each module we proposed, we conducted a series of decremental ablation study, removing one component once. The results are displayed in Table 3, and we are giving a specific explanation. 1) **Without Single Theme**. In this section we alternately removed each theme and use the rest four themes to perform prediction. According to the results, removing any theme would cause a performance drop, from 5% drop on WA*F1 of work theme to 11% of mental theme

Model	Acc.	WA*Prec.	WA*Rec.	WA*F1
w/o family	0.85	0.85	0.85	0.85
w/o work	0.87	0.87	0.87	0.87
w/o mental	0.81	0.81	0.81	0.81
w/o medical	0.85	0.85	0.85	0.85
w/o overall	0.81	0.81	0.81	0.81
w/o TCL	0.78	0.77	0.77	0.77
w/o ITAS	0.81	0.81	0.81	0.81
PDIMC	**0.94**	**0.92**	**0.91**	**0.92**

Table 3: Ablation study to investigate the effect of different themes and tailored module.

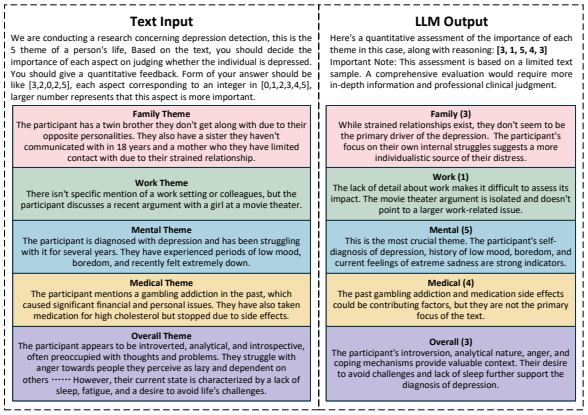

Figure 3: Visualization of the theme-oriented in-context learning technique illustrates the thems learned from clinical interview dialogues.

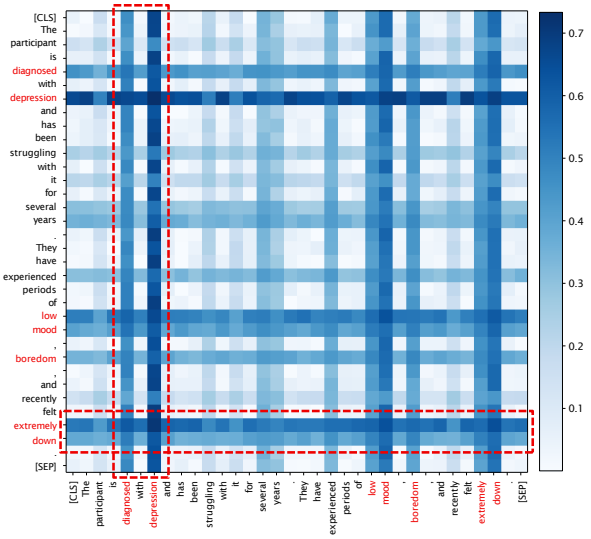

Figure 4: Visualization of intra-theme attention weights.

and overall theme. This demonstrated the irreplaceability of each theme, as they can maximize the utilization of the input text and collaboratively predict depression predisposition. 2) **Without TCL Module**. We removed the Theme Correlation Learning module, which means self-attention mechanism is disabled. From the result, we can see that the performance suffers the most performance decline, about 15% on WA*F1. This manifested the validity of self-attention mechanism which learns correlation both intra-theme and inter-theme and reinforces the accuracy of model predictions. 3) **Without ITAS Module**. This variant removes the Interactive Theme Adjustment Strategy module, and this means no discrepancy of weights would be applied during the multi-theme fusion process. Evidently, the 11% drop on WA*F1 indicated the indispensability of considering the prioritization of each theme when fusing them together. In general, the complete version of our model PDIMC outperformed the others with module removed in depression detection tasks. This result verify the effectiveness and complementary of the three components.

4.4 Theme-oriented In-context Learning Analysis

Apart from achieving the superior performance, the key advantage of our model over other approaches is its ability to learn thematic content from complex clinical dialogues due to theme-oriented in-context learning. To this end, we carried out experiments to explore the ability of the module to extract thematic descriptions from the transcript text of the interview based on the in-context samples. Figure 3 illustrates the themes extracted from hundreds of dialogue rounds, as well as the results of simulated clinical feedback generated by the large language model. By examining Figure 3, we can observe that the theme-oriented in-context learning technique accurately captures content related to family, work, mental health, and medical themes within the dialogue. Furthermore, the feedback simulated by LLM aligns well with common knowledge (for more details, refer to Section 4.6). This result proved the capability of TICL module to extract valid thematic information from complex dialogues.

4.5 Theme Correlation Learning Analysis

To qualitatively validate the effectiveness of the TCL module to capture correlation both intra-theme and inter-theme, we conducted experiments and visualized the map of attention distributions. Figure 4 displays the token weights of intra-theme correlation distribution. From Figure 4, we could observe the token distribution within the

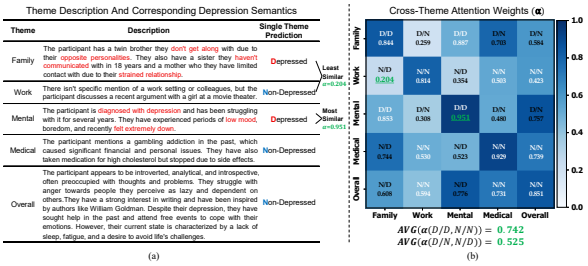

Figure 5: Visualization of inter-theme attention weights. (a) uses description of each single theme respectively to predict depression tendency, and we notate the predicted label upon the attention map of (b) (D for Depressed, N for Non-Depressed). Additionally, average attention scores of themes that share same and different predicted labels are calculated under the attention map in (b).

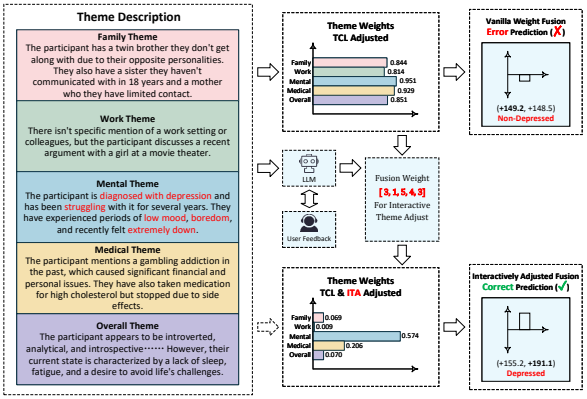

Figure 6: Visualization of the weights of each theme before and after applying the interactive theme adjustment strategy.

"Mental" theme and the corresponding attention weights. Some tokens, such as "diagnosed," "depression","extremely down" and "low mood," show a clear depressive tendency, and their corresponding attention weights are significantly higher than those of other tokens. This result indicates that intra-theme correlation learning can highlight token semantic information related to depressive states, thereby enhancing the performance of the depression detection model.

To gain the deep insights into our proposed inter-theme of theme correlation learning, we analyzed the weights of several inter-theme correlation. We obtained the following observations: 1) Themes with consistent depressive semantics have higher inter-theme correlation weights, while themes with inconsistent depressive semantics have lower correlation weights. For example, the highest similarity score among themes with consistent depressive semantics is 0.951, whereas the lowest similarity score among themes with inconsistent depressive semantics is 0.204. 2)The average weight of themes with consistent depressive semantics is higher than that of themes with inconsistent depressive semantics (0.742 vs. 0.525). 3)The inter-theme correlation weights are highly correlated with and accurately reflect depressive semantic information. These results indicate that inter-theme correlation learning effectively captures cross-theme semantic correlations, aiding the model's decision-making process and ultimately improving prediction accuracy.

4.6 Interactive Theme Adjustment Strategy Analysis

To validate the effectiveness of the interactive theme adjustment strategy, we compared the weights obtained through inter-theme correlation learning with those learned through the interactive theme adjustment strategy. Specifically, we first used the LLM to simulate clinical doctor feedback by scoring each theme, and then converted the scores into weights to intervene in the model's prediction. From Figure 6, we can observe that the LLM's scoring of theme importance is highly intuitive, such as the "Mental" theme receiving a high weight due to its direct correlation with depressive states. This result indicates that using the LLM to simulate clinician feedback is accurate.

From Figure 6, we also could see that inter-theme correlation learning could, to some extent, capture semantic information related to depression, such as the higher weight of the "Mental" theme compared to other themes. However, the distinction is still not sufficiently clear, which may lead to incorrect predictions. After applying the interactive theme adjustment strategy, the weight of the "Mental" theme becomes significantly higher than that of other themes, leading to the correct prediction. This is because, during the multi-theme fusion phase, the distinction between themes is not clear enough, reducing the influence of key themes on the prediction results. These results demonstrate the importance of introducing simulated clinician feedback through the LLM, as it highlights key theme information to guide the model towards making accurate predictions.

5 Conclusion

In this paper, we proposed a novel interactive depression detection framework. The framework includes three components: 1) The theme-oriented in-context learning module learns interview theme-related information from multi-turn dialogues. 2) The theme correlation learning module mines correlation of intra-theme and inter-theme. 3) The interactive theme adjustment strategy introduces clinical feedback simulated by the LLM to guide the model in aligning with the preferences indicated by external feedback. Extensive experiments demonstrate the effectiveness and superiority of our proposed model. It not only achieves accurate depression diagnosis but also allows for clinical intervention, enabling integration of expert feedback into the diagnostic process.

Limitations

There are three limitations in this study. First, although our approach achieves outstanding performance on depression detection datasets, it only utilizes textual data from clinical interview dialogues without incorporating multimodal information. Integrating multimodal data could potentially further enhance the model's performance. Second, in designing the interactive depression detection framework, we use the large language model to simulate clinical feedback as external guidance. However, the professional reliability of this simulated feedback may require further evaluation. Finally, this method lacks validation in real-world clinical settings.

Acknowledgments

This study is partially supported by National Key R&D Program of China (2023YFC3502900), National Natural Science Foundation of China (62276082), Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E09/22), Major Key Project of PCL (PCL2021A06), Shenzhen Soft Science Research Program Project (RKX20220705152815035), Shenzhen Science and Technology Research and Development Fund for Sustainable Development Project (GXWD20231128103819001, No.KCXFZ20201221173613036, 20230706140548006), and the Fundamental Research Fund for the Central Universities (HIT.DZJJ.2023117), and the Outstanding Youth Science Foundation of Shaanxi under Grant 2025JC-JCQN-083, and the Key Research and Development Program of Shanxi Province under Grant 2025CY-YBXM-047.

References

Abdurrahim Abdurrahim and Dhomas Hatta Fudholi. 2024. Mental health prediction model on social media data using cnn-bilstm. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*.

Abdelrahman A. Ali, Aya E. Fouda, Radwa J. Hanafy, and Mohammed E. Fouda. 2024. Leveraging audio and text modalities in mental health: A study of llms performance. *ArXiv*, abs/2412.10417.

Ashutosh Anshul, Gumpili Sai Pranav, Mohammad Zia Ur Rehman, and Nagendra Kumar. 2024. A multimodal framework for depression detection during covid-19 via harvesting social media. *IEEE Transactions on Computational Social Systems*, 11:2872–2888.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. 2023. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Syst. Appl.*, 217:119538.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. pages 346–350.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048.

Michael Danner, Bakir Hadzic, Sophie Gerhardt, Simon Ludwig, Irem Uslu, Peng Shao, Thomas Weber, Youssef Shiban, and Matthias Ratsch. 2023. Advancing mental health diagnostics: Gpt-based method for depression detection. In *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, pages 1290–1296.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. pages 1107–1128.

Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, page 69–76, New York, NY, USA. Association for Computing Machinery.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.

Tuka Al Hanai, Mohammad Mahdi Ghassemi, and James R. Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *Proceedings of the 28th ACM International Conference on Multimedia*.

Xiangyu Jia, Xianbing Zhao, Buzhou Tang, and Ronghuan Jiang. 2024. Bidirectional multimodal block-recurrent transformers for depression detection. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3323–3328. IEEE.

Juho Jung, Chaewon Kang, Jeewoo Yoon, Seungbae Kim, and Jinyoung Han. 2024. Hique: Hierarchical question embedding network for multimodal depression detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1049–1059, New York, NY, USA. Association for Computing Machinery.

Gleb Kuzmin, Petr Strepetov, Maksim Stankevich, Ivan Smirnov, and Artem Shelmanov. 2024. Mental disorders detection in the era of large language models. *ArXiv*, abs/2410.07129.

Genevieve Lam, Dongyan Huang, and Weisi Lin. 2019. Context-aware deep learning for multi-modal depression detection. pages 3946–3950.

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024. Depression detection on social media with large language models. *ArXiv*, abs/2403.10750.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. pages 1401–1422.

Lin Lin, Xuri Chen, Ying Shen, and Lin Zhang. 2020. Towards automatic depression detection: A bilstm/1d cnn-based model. *Applied Sciences*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? pages 100–114.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews.

Muzafar Mehraj Misgar and M.P.S Bhatia. 2022. Detection of depression from iomt time series data using umap features. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 623–628.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Alex Rinaldi, Jean Fox Tree, and Snigdha Chaturvedi. 2020. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18, Online. Association for Computational Linguistics.

Guillaume Sanchez, Alexander Spangher, Honglu Fan, Elad Levi, and Stella Biderman. 2024. Stay on topic with classifier-free guidance. In *Forty-first International Conference on Machine Learning*.

Aharon Satt, Shai Rozenberg, and Ron Hoory. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*.

Shahid Munir Shah, Syeda Anshrah Gillani, Mirza Samad Ahmed Baig, Muhammad Aamer Saleem, and Muhammad Hamzah Siddiqui. 2024. Advancing depression detection on social media platforms through fine-tuned large language models. *ArXiv*, abs/2409.14794.

Yingli Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. pages 6247–6251.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. pages 819–862.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. pages 6292–6307.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Han-Guang Wang, Huirang Hou, Li-Cheng Jin, Chen-Yang Xu, Zhong-Yi Zhang, and Qinghao Meng. 2024. Sad-time: a spatiotemporal-fused network for depression detection with automated multi-scale depth-wise and time-interval-related common feature extractor. *ArXiv*, abs/2411.08521.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. 2022. Multi-modal depression estimation based on sub-attentional fusion. In *ECCV Workshops*.

Wen Wu, Chao Zhang, and Philip C Woodland. 2023. Self-supervised representations in speech-based depression detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. page 4489–4500.

Xingda Yao, Lingkang Ying, Tianxiang He, Ligang Ren, Ruiji Xu, and Keji Mao. 2024. Depression detection based on multilevel semantic features. In *Artificial Neural Networks and Machine Learning – ICANN 2024: 33rd International Conference on Artificial Neural Networks, Lugano, Switzerland, September 17–20, 2024, Proceedings, Part VIII*, page 44–55, Berlin, Heidelberg. Springer-Verlag.

Jiaxin Ye, Junping Zhang, and Hongming Shan. 2024. Depmamba: Progressive fusion mamba for multimodal depression detection. *ArXiv*, abs/2409.15936.

Ming Ying, Xuexiao Shao, Jing Zhu, Qinglin Zhao, Xiaowei Li, and Bin Hu. 2024. Edt: An eeg-based attention model for feature learning and depression recognition. *Biomedical Signal Processing and Control*, 93:106182.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024. When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. pages 146–158.

Xianbing Zhao, Yinxin Chen, Sicen Liu, and Buzhou Tang. 2022. Shared-private memory networks for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(4):2889–2900.

Xianbing Zhao, Xuejiao Li, Ronghuan Jiang, and Buzhou Tang. 2025. Resolving multimodal ambiguity via knowledge-injection and ambiguity learning for multimodal sentiment analysis. *Information Fusion*, 115:102745.