

# Enhancing Depression-Diagnosis-Oriented Chat with Psychological State Tracking

Anonymous ACL submission

## Abstract

Depression-diagnosis-oriented chat aims to guide patients in self-expression to collect key symptoms for depression detection. Recent work focuses on combining task-oriented dialogue and chitchat to simulate the interview-based depression diagnosis. Whereas, these methods can not well capture the changing information, feelings, or symptoms of the patient during dialogues. Moreover, no explicit framework has been explored to guide the dialogue, which results in some useless communications that affect the experience. In this paper, we propose to integrate Psychological State Tracking (POST) within the large language model (LLM) to explicitly guide depression-diagnosis-oriented chat. Specifically, the state is adapted from a psychological theoretical model, which consists of four components, namely Stage, Information, Summary and Next. We fine-tune an LLM model to generate the dynamic psychological state, which is further used to assist response generation at each turn to simulate the psychiatrist. Experimental results on the existing benchmark show that our proposed method boosts the performance of all subtasks in depression-diagnosis-oriented chat.

## 1 Introduction

Depression remains an escalating mental health threat globally, due to the severe scarcity and limited access to professionals (Sharma et al., 2020; Pérez et al., 2023). To alleviate such situations, computational methods are emerging tools for large-scale depression detection and diagnosis (Zhang et al., 2021; Pacheco-Lorenzo et al., 2021). Conversations agents related to mental health become a promising technological solution for early depression detection unobtrusively, due to the traditional detection mechanisms being invasive (Chaytor and Schmitter-Edgecombe, 2003). The most popular implementation of chatbots is based on self-rating scales (Jaiswal et al., 2019; Arrabales,

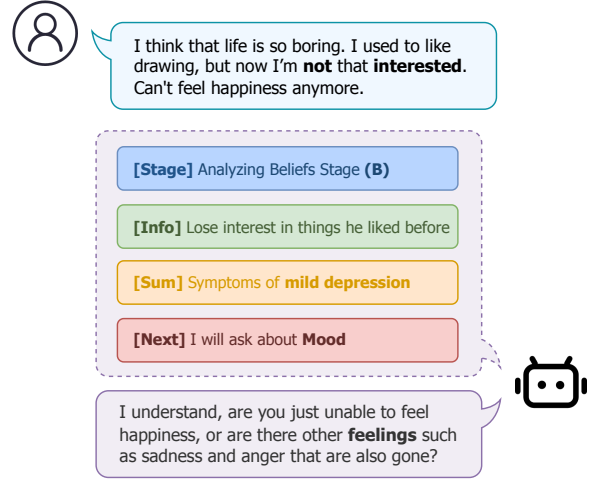


Figure 1: Chat guided by psychological state tracking.

2020) or diagnostic criteria (Philip et al., 2017) by calibrating conversational interactions to cover the targeted topics on the scales, which in turn limits patients to express their feelings (Yao et al., 2022). Therefore, there is still a significant gap for such systems to provide chat-like depression detection in clinical settings.

In practical psychotherapy, psychiatrists calibrate the dialogue flow to gather and synthesize key symptom information from patients while providing appropriate strategies like emotional support. To simulate this process, Yao et al. (2022) defined this type of dialogue as *Task-oriented Chat* and collected the first dialogue dataset D<sup>4</sup> for depression diagnosis. However, existing work mostly focuses on shallow heuristic attempts such as predicting topics and generating empathetic responses, falling short of capturing the changing information, feelings or symptoms of the patient during dialogues.

Recently, Large Language Models (LLMs) have achieved remarkable success in various text reasoning tasks (Kojima et al., 2022; Bang et al., 2023). In the field of psychology, ChatGPT and GPT-4 have shown promising performance in attributing mental states (Bubeck et al., 2023; Chen et al., 2023c).

While initial researchers envisioned further harnessing this capability for complex psychological tasks, the majority still focused on developing chatbots for emotional support purposes. Moreover, no explicit framework has been explored to guide the dialogue, which results in some useless communications that affect the experience.

To enhance depression-diagnosis-oriented chat, we propose the **Psychological State Tracking (POST)** to link the patient’s current symptom with the doctor’s next strategy. Inspired by the ABC Model (Ellis, 1991) in Cognitive-behavioral therapy (CBT), we define POST into four components: *Stage*, *Information*, *Summary* and *Next*. Figure 1 shows an illustrative example. First, we figure out at which stage of the current diagnosis procedure is; Then, we distinguish the key symptoms that the patient is exhibiting; After that, we document the current diagnostic results of the patient; Finally, in *Next*, we introduce an auxiliary targeted stimulus prompt to align with specific counseling strategies. We jointly optimize the POST model and the response generation model by an LLM. We conduct comprehensive experiments using the trained LLM. Experimental results on the existing benchmark show that our proposed method achieves the best performance of all subtasks in depression-diagnosis-oriented chat. Furthermore, psychological state tracking, as the explicit thought behind response generation, provides professional-compliant interpretability to the diagnostic process.

Overall, we make the following contributions:

- We annotate a fine-grained dataset by augmenting the D4 dataset, which annotates the psychological state of each conversation round guided by Albert Ellis’ ABC Model.
- We propose a joint model to explicitly guide depression-diagnosis-oriented chat, which integrates psychological state tracking into an LLM to learn the connection between patients’ state changes and doctors’ strategic planning.
- Extensive experiments on the existing benchmark show that our proposed method boosts the performance of all subtasks in depression-diagnosis-oriented chat. In particular, psychological state tracking, as the explicit thought behind response generation, provides professional-compliant interpretability.

## 2 Related Work

### 2.1 Depression Diagnosis

Depression diagnosis aims to use diagnostic tools to identify symptoms and determine the severity of depression. Early research was conducted by psychiatrists in controlled settings through self-questionnaires, such as the PHQ-9 (Kroenke et al., 2001) and GAD-7 (Spitzer et al., 2006), to assess patients’ cognitive or emotional states. However, in face-to-face settings, individuals often hesitate to express their mental state. Some researchers explore using different network structures to automatically identify mental health status in social media content (Bucur et al., 2023; Gui et al., 2019; An et al., 2020). The poor interactivity of these approaches also limits patients’ self-expression. Recent efforts focus on developing conversational agents to perform diagnosis in interactive environments (Jaiswal et al., 2019; Arrabales, 2020). Yao et al. (2022) propose to combine task-oriented dialogue and chitchat to simulate the interview-based depression diagnosis. Whereas, these methods struggle to well capture the changing information, feelings, or symptoms of the patient during the dialogue process. Moreover, there has been no exploration of an explicit framework to guide the response generation. Thus, we aim to explore a more personalized and professional depression diagnostic chatbot.

### 2.2 Large Language Models for Psychology

Large Language Models (LLMs) have demonstrated remarkable performance in interactions in zero-shot settings, igniting a fervent interest in their application within the field of psychology. Researchers adopt methodologies from human psychology to explore the behavior of LLMs (Hagendorff, 2023b), observing that the responses from LLMs often show a similarity to human responses in most scenarios (Almeida et al., 2023). Additionally, LLMs have been utilized as auxiliary tools to advance specific psychological tasks (Chen et al., 2023b); for example, Qin et al. (2023) utilized an LLM to develop an interpretable and interactive system for detecting depression. A few more recent works have begun to enhance the performance of LLMs with deeper professional psychological knowledge (Chen et al., 2023c). These approaches not only enrich the capabilities of LLMs but also open new avenues for psychological research and applications. However, with the knowledge gap

between the two communities, depending only on LLM-based systems for psychology can lead to unpredictable outcomes (Hagendorff, 2023a). Hence, we aim to further fine-tune LLMs combined the professional psychological theories to achieve more professional capabilities for psychology.

### 2.3 Dialogue State Tracking

Dialogue State Tracking (DST) is essential in task-oriented dialogue systems for monitoring conversation states. Previous studies have utilized pre-trained models to improve DST. For example, Lin et al. (2021) and Yu et al. (2022) employed large-scale pre-trained models for zero-shot cross-domain DST and for enhancing dialogue state tracking with external knowledge, respectively. Sun et al. (2022) revolutionizes dialogue state tracking with a Mentioned Slot Pool (MSP) to improve accuracy, while Xu et al. (2023) combines visit sequences and clinical notes through contrastive learning and adaptive language modeling to enhance diagnosis predictions. However, these approaches can not well handle complex dialogues and capture fine-grained semantic relationships.

The rise of large language models has led to advancements in DST methods, such as T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020). Recent research focuses on parameter-efficient fine-tuning (PEFT) to enhance DST performance. Notable techniques include prompt tuning by Lee et al. (2021) and Yang et al. (2023), as well as prefix tuning (Ma et al., 2023) and meta-learning (Chen et al., 2023a). Despite these progresses, there remains a gap when compared to more advanced models like ChatGPT, and additional support from psychological theories is necessitated to accomplish our task. Therefore, our proposed methodology combines the strengths of open-source LLMs and fine-tuning techniques like LoRA (Hu et al., 2021) to improve the fundamental generation ability, and utilizes the ABC model in Cognitive-behavioral therapy to design the DST framework.

## 3 Data Annotation

### 3.1 Annotation Procedure

We adapt the D<sup>4</sup> (Yao et al., 2022) dataset with additional psychological state annotations, which contains 1,339 clinically standardized conversations about depression. The clinical data can facilitate a generation and diagnosis process that closely simulates real-life clinical consultations for depression.

Source	Criteria	Total
D <sup>4</sup>	Dialogues	1,339
	Dialogue turns	28,977
	Average turns per dialogue	21.67
	Average tokens per dialogue	577.12
	Average tokens per utterance	13.31
	Average patient tokens per utterance	11.87
	Average doctor tokens per utterance	14.76
POST	Stage-A per dialogue	5.84
	Stage-B per dialogue	4.11
	Stage-C per dialogue	11.72
	Average Info tokens per turn	47.12
	Average Summary tokens per turn	22.69
	Average POST tokens per turn	75.81

Table 1: Statistics of D<sup>4</sup> annotated with POST

However, the original data lacked tracking of the patient’s conditions. To transform the raw data into a sample set that can be used for psychological state tracking, we annotate the conversations following three steps: (1) First, the patient’s known information and current stage are captured in real-time as the conversation occurs. Special tokens representing the current stage and information are added to each utterance. (2) The annotation of summary and next action is then performed using in-house software. Our linguistic annotators are familiar with real-life clinical consultation scenarios, having completed training courses in depression diagnosis. We followed an annotation protocol defined by authoritative consultants for manual labeling. Subsequently, the manually labeled samples were used as input examples for GPT-4 to carry out the remaining annotations. (3) To ensure the annotation quality of GPT-4, we randomly sample 200 instances from the dataset annotated by GPT-4 and conduct manual verification. Out of these, 163 annotations are deemed accurate, resulting in an accuracy rate of 81.5%. The prompt template of annotation can be found in Appendix A.

### 3.2 Data Analysis

**Statistics** The basic statistics of the annotated dataset are shown in Table 1. The total Dialogues of the dataset is 1,339, while the total turns of dialogues is 28,977. Due to the data cleaning procedure, dialogue turns remaining may be less than the number of the original D<sup>4</sup> dataset. The average doctor tokens per utterance is 14.76, which is approximately 3 tokens more than the average patient tokens, indicating that the doctors often speak more due to the need for consultation or empathetic consolation. The second part concerns the statistical analysis of annotated POSTs. The average num-

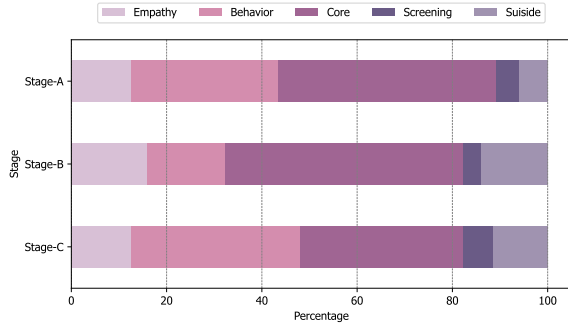


Figure 2: Distribution of topics in different stages

ber of POST tokens per turn is 75.81, suggesting that POSTs contain more information compared to utterances. Notably, the Info part, which contains patient information from dialogue history, is lengthier with an average length of 47.12. In contrast, the Summary, which involves further inference of the patient’s diagnostic results, is almost half the length of the Info part, averaging only 22.69.

**Stage Analysis** The distribution of topics in different stages is illustrated in Figure 2. The chart reveals that Stage-A primarily focuses on Core and Behavior, aligning with its objective of identifying triggering events. In Stage-B, there is an emphasis on Empathetic Comfort and a notable increase in attention to Suicidal tendencies. This reflects the stage’s pivotal role in responding to patient disclosures and probing the beliefs of patients linked to identified symptoms and events. Conversely, Stage-C, as the terminal phase, primarily focuses on evaluating behavioral outcomes and intensifies screening to facilitate final diagnoses.

Figure 3 delineates the stage transitions in dialogues, highlighting the proportional changes every two turns starting from the second turn. Initially, dialogues commence with Stage-A, focusing on recent events potentially triggering the patient’s condition. After a thorough understanding of the precipitating events, the dialogue phase gradually transitions to Stage-B. This stage involves detailed inquiries into the patient’s thoughts and beliefs, and their attitudes towards these events, aiming to gain a deeper insight into the patient’s internal thoughts, including any tendencies towards depression or suicide. As the dialogue progresses, it shifts to Stage-C, where the clinician examines the patient’s core behaviors and symptoms, such as sleep and eating habits, to identify specific manifestations.

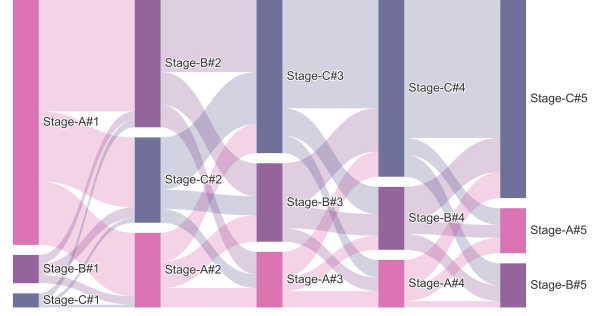


Figure 3: Transitions of stages

## 4 Method

Our approach formalizes depression-diagnosis-oriented chat by representing the user’s psychological state as a set of task attributes, as shown in Figure 1. The goal is to generate doctors’ probable responses based on the dialog context, taking into account the current state and next planning. As shown in Figure 4, following the task-specific fine-tuning paradigm, we build a joint model of psychological state tracking and response generation by equipping a transformer-based language backbone with functional modules.

### 4.1 Task Formulation

For a depression-diagnosis-oriented chat task, there is a  $t$  turn dialogue between a patient and a doctor that can be represented as:

$$h_t = (u_1^p, s_1, u_1^d, \dots, u_t^p, s_t, u_t^d) \quad (1)$$

where  $s_i$  is the patient’s psychological state at turn  $i$ ,  $u_i^p$  is the patient’s utterance, and  $u_i^d$  is the doctor’s response at turn  $i$ . The psychological state of the patient  $s_i$  represents the psychological diagnosis of thought, which consists of four subsequences  $\langle \text{Stage}, \text{Info}, \text{Sum}, \text{Next} \rangle$ . The entire depression-diagnosis-oriented chat procedure can be split into 4 subtasks: **Psychological State Tracking** predicts the patient’s current psychological state and the next action based on the dialogue context. **Response Generation** generates the most likely response based on the dialogue history and current psychological state. Similar to D<sup>4</sup>, we jointly optimize the psychological state tracking model and the response generation model by an LLM. **Dialogue Summary** concentrates on generating summaries of symptoms, drawing from the complete dialogue context; **Severity Classification** assesses the severity of depression and suicide risk independently, utilizing information obtained from dialogue history and symptom summary.



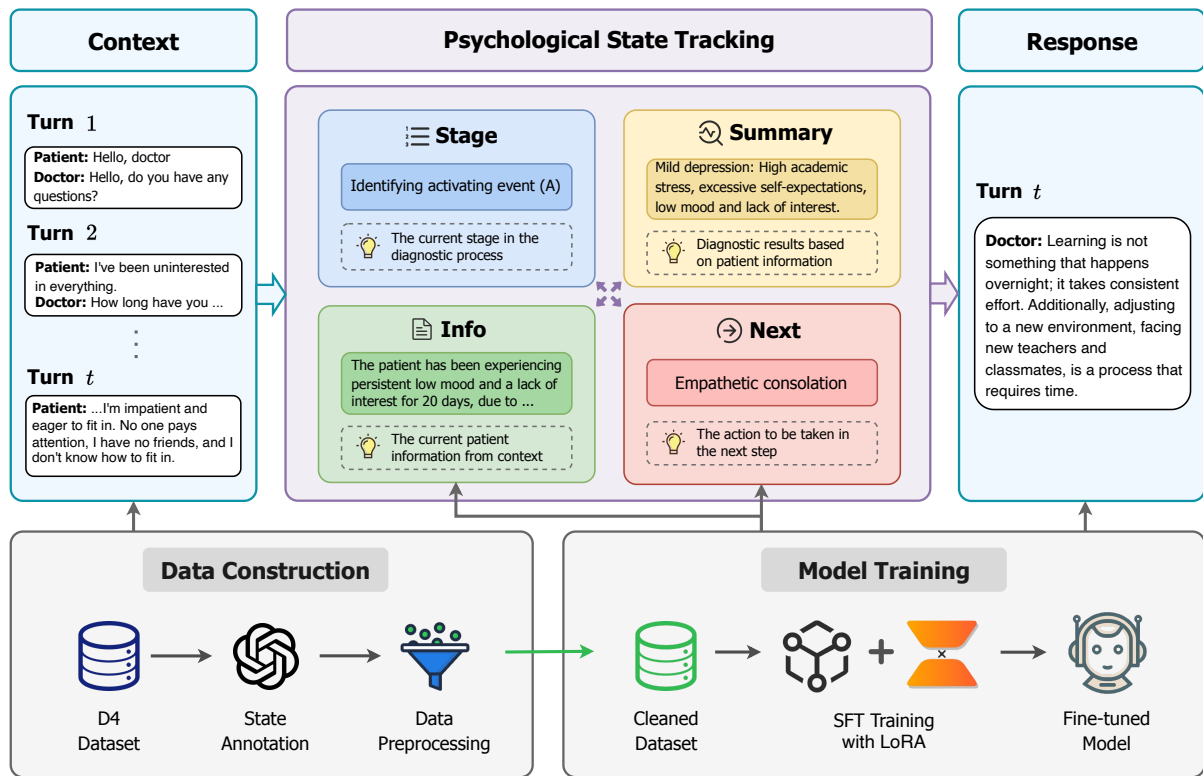


Figure 4: The overall framework of depression-diagnosis-oriented chat with psychological state tracking.

## 4.2 Psychological State Tracking

Clinically, interview-based depression diagnosis need to collect and summarize key symptom information about one patient while providing a chat-like conversation experience. Psychiatrists need to obtain enough information from patients with multiple colloquial rhetoric based on the patient’s speech or written content. Thus, psychiatrists need to design the questioning logic between questions of symptoms from mild to severe during the consultation. To model such finer-grained relationships between patients’ symptoms and questions, we perform psychological state tracking, where the state consists of the following components.

**Stage.** In clinical practice, a consultation follows a gradual in-depth manner and diagnosis topics consistently occur across turns. For example, doctors usually start by asking about core symptoms such as mood and interests, and then gradually turn to behavior symptoms. To perform a deep analysis of the topic transition, we first need to find out at which stage of the current diagnosis procedure. After an assessment, we summarize the dialogue stage into the context that served as the evidence base for asking about different types of symptoms.

**Information.** This part aims to discover the psychological information that the patient is exhibiting in the dialogue history. Based on the presented symptoms, we derive the patient’s varying severity. For healthy individuals, the conversation typically manifests surface symptoms such as changes in sleep. As the condition worsens, patients tend to exhibit an increasing array of symptoms. By documenting the content and quantity of symptoms exhibited by the patient as a form of memory, we can gain a clearer understanding of the extent of their depression and devise strategies for planning.

**Summary.** This subsequence aims to document the current diagnosis of the patient. Depression diagnoses are primarily employed for preliminary screening. For milder cases, empathetic strategies are generally used to encourage the patient’s self-expression, while for severe cases, immediate crisis intervention is required. Therefore, we have incorporated real-time diagnostic results into the process of depression diagnosis.

**Next.** Given the inherent ambiguity in the depression diagnosis chat, defining clear planning as in other task-oriented conversations is challenging. To facilitate dialogue generation, we introduce an auxiliary targeted stimulus prompt to the psy-

chological state. These prompts serve as nuanced, patient-specific hints and clues, guiding LLMs to align with specific counseling strategies.

We train LLMs to generate the above components in a psychological state. With psychological state tracking, we obtain a fully interpretable thought process for generating responses focused on depression diagnosis. As a diagnostic aid, such interpretability is crucial for professionals in clinical practice to substantiate the rationale behind the results, which can derive intricate thought patterns behind the entire diagnostic process from the generated thought processes.

### 4.3 Fine-tuning

In this work, we fine-tune ChatGLM3<sup>1</sup> with a parameter efficient approach, i.e., Low-Rank Adaptation (LoRA). LoRA maintains the weights of pre-trained LMs while introducing trainable rank decomposition matrices into each transformer layer, making it feasible to fine-tune LLMs with much fewer computational resources.

We fine-tune an LLM to track the psychological state and generate the response generation jointly, given the crafted example and the annotated label. Specifically, the objective is to predict the next token based on language modeling:

$$\min_{\theta} \sum_{t=1}^T -\log p_{\theta}(s_t, u_t^d | h_{<t}, u_t^p; \theta) \quad (2)$$

where  $\theta$  represents the parameters for a language model and  $T$  is the total turns of the dialogue. Ideally, the objective encourages the model to learn the target distribution by predicting tokens in the sequence. By placing the psychological state before the doctor’s response, the model learns to fuse the distribution from thought to response in an in-context language modeling manner. We only compute the loss of tokens on the psychological state  $s$  and the doctor utterance  $u^d$ .

## 5 Experiments

### 5.1 Baselines

We leverage the original Transformer architecture (Vaswani et al., 2017) fine-tuned on MedDialog dataset (Zeng et al., 2020) as a foundational model. BART model (Lewis et al., 2019) and the CPT model (Shao et al., 2021) are employed as our primary baselines, with the same configuration as

previous studies (Yao et al., 2022). ChatGPT (gpt-3.5-turbo)<sup>2</sup> is listed as a baseline as well, which is optimized for chat by introducing Reinforcement Learning with Human Feedback (RLHF).

### 5.2 Implementation Details

Our model is implemented in PyTorch, using ChatGLM3-6b as the base architecture. In the supervised fine-tuning process, we apply LoRA to all linear layers of the model, where LoRA rank is set to 64. We set the batch size, max context length, and learning rate to 32, 1024, and 2e-4 respectively. The model is trained on one A800 GPU for 5 epochs, which costs about 8 hours.

### 5.3 Automatic Evaluation

**Response Generation** Metrics like BLEU-2 (Papineni et al., 2002), Rouge-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) are employed to assess the response generation quality. In addition, we calculate DIST-2 (Li et al., 2015) to demonstrate the diversity of responses. Tokenization is performed using the jieba<sup>3</sup> library, following the same approach as in D<sup>4</sup>.

Table 2 shows the results of the automatic evaluation for the response generation task. We have the following observations. **First**, without incorporating POSTs, the model achieves considerably better performances on BLEU-2, ROUGE-L and METEOR. It is mainly attributed to the larger scale of the language model and optimization for Chinese dialogue generation. **Second**, the integration of POSTs markedly enhanced diagnostic response capabilities, particularly in BLEU-2 and METEOR, by effectively assessing the diagnosis stage, summarizing patient information, and inferring potential diagnoses. By tracking psychological states, the POST guides the model in generating more rational and interpretable responses, and a further enhancement is observed while the golden POSTs are given. **Third**, The results from ChatGPT indicate relatively lower BLEU-2 and ROUGE-L scores due to the lack of dataset-specific fine-tuning. The introduction of POSTs notably improved ChatGPT’s performance by 7%, with substantial improvements upon using golden POSTs, underscoring the effectiveness of the POST strategy.

**Summarization** Table 3 presents the results for dialog summary task. We provide the model with

<sup>1</sup><https://github.com/THUDM/ChatGLM3>

<sup>2</sup><https://openai.com/blog/chatgpt>

<sup>3</sup><https://github.com/fxsjy/jieba>

Model	BLEU-2	ROUGE-L	METEOR	DIST-2	Topic ACC.
Transformer-	7.28%	0.21	0.1570	<b>0.29</b>	-
ChatGPT-	10.01%	0.19	0.3254	0.15	-
BART-	19.29%	0.35	0.2866	0.09	-
CPT-	19.79%	0.36	0.2969	0.07	-
Ours-	<b>30.85%</b>	<b>0.47</b>	<b>0.4524</b>	0.17	-
Transformer	13.43%	0.33	0.2620	0.04	36.82%
ChatGPT	17.52%	0.30	0.4080	<b>0.19</b>	14.62%
BART	28.62%	0.48	0.4053	0.07	59.56%
CPT	29.40%	0.48	0.4142	0.06	<b>59.77%</b>
Ours	<b>39.76%</b>	<b>0.50</b>	<b>0.5305</b>	0.11	56.98%
Transformer*	25.37%	0.41	0.3905	0.04	-
ChatGPT*	29.46%	0.42	0.5158	<b>0.23</b>	-
BART*	37.02%	0.54	0.4920	0.07	-
CPT*	37.45%	0.54	0.4943	0.06	-
Ours*	<b>45.28%</b>	<b>0.56</b>	<b>0.5794</b>	0.10	-

Table 2: Evaluation results of response generation and next topic prediction. Note that '-' means POSTs are excluded, and \* means golden POSTs are given.

Model	BLEU-2	ROUGE-L	METEOR	DIST-2
BART	16.44%	0.26	0.25	0.19
CPT	16.45%	0.26	0.24	0.21
Ours	<b>18.32%</b>	<b>0.30</b>	<b>0.29</b>	<b>0.26</b>

Table 3: Evaluation results of dialog summary task

all the dialogue context and concatenate it with the patient's portrait information. The model then generates a summary of the patient's symptoms. The results show that our model surpasses both BART and CPT in text generation and also exhibits better diversity according to the DIST-2 scores.

**Severity Classification** Results for depression severity and suicide risk severity classification are listed in Tables 4 and 5. To be consistent with D<sup>4</sup>, we use average weighted precision, recall, and F1 scores computed by sklearn<sup>4</sup> to evaluate classification results.

The results demonstrate that our approach outperforms the baseline models in most scenarios, with a notable performance enhancement in 4-classification tasks when utilizing summaries compared to dialog history. However, the difference is less pronounced in the 2-classification tasks due to the increased complexity of the 4-classification tasks. Summaries provide concise and effective diagnostic information about the patient's condition, which leads to a more substantial improvement in performance. Additionally, the variance in the 4-classification results based on dialog history is relatively high, which might be partly due to un-

Task	Input	Model	Precision	Recall	F1
2-class	dialog	BERT	0.81±.04	0.80±.03	0.80±.03
		BART	0.80±.02	0.79±.03	0.79±.03
		CPT	0.79±.02	0.78±.03	0.78±.03
		Ours	<b>0.88±.00</b>	<b>0.86±.01</b>	<b>0.84±.01</b>
	summary	BERT	0.90±.02	0.90±.02	0.90±.02
		BART	0.89±.03	0.89±.03	0.89±.03
		CPT	<b>0.92±.01</b>	<b>0.92±.02</b>	<b>0.92±.01</b>
		Ours	0.91±.01	0.91±.02	0.91±.01
4-class	dialog	BERT	0.49±.05	0.45±.04	0.45±.04
		BART	0.53±.04	0.53±.04	0.52±.04
		CPT	0.49±.04	0.47±.04	0.46±.05
		Ours	<b>0.57±.07</b>	<b>0.56±.04</b>	<b>0.53±.04</b>
	summary	BERT	0.67±.04	0.66±.04	0.66±.04
		BART	0.68±.03	0.67±.02	0.66±.02
		CPT	0.73±.03	0.72±.03	0.72±.03
		Ours	<b>0.79±.02</b>	<b>0.76±.02</b>	<b>0.76±.02</b>

Table 4: Results of depression severity classification

even label distribution in the dataset. Besides, the 2-classification accuracy of suicide severity based on dialog history have reached 90%, indicating this approach's practical viability for diagnosing suicidal tendencies from dialog history.

## 5.4 Human Evaluation

To simulate realistic depression diagnosis scenarios for evaluation, we prompt ChatGPT to act as patients, based on patient backgrounds from the dataset. Correspondingly, Our model and baselines played the role of doctors, conducting diagnosis conversations with the patients. Each conversation consists of a minimum of 15 turns and will terminate at an appropriate round. Then, we assigned annotators with dialogue pairs to evaluate

<sup>4</sup><https://scikit-learn.org>

Task	Input	Model	Precision	Recall	F1
2-class	dialog	BERT	0.81 $\pm$ .02	0.78 $\pm$ .02	0.79 $\pm$ .02
		BART	0.77 $\pm$ .02	0.75 $\pm$ .02	0.75 $\pm$ .02
		CPT	0.84 $\pm$ .02	0.82 $\pm$ .03	0.82 $\pm$ .03
		Ours	<b>0.91<math>\pm</math>.01</b>	<b>0.90<math>\pm</math>.02</b>	<b>0.91<math>\pm</math>.01</b>
4-class	dialog	BERT	0.72 $\pm$ .03	0.64 $\pm$ .04	0.66 $\pm$ .03
		BART	0.70 $\pm$ .05	0.66 $\pm$ .04	0.65 $\pm$ .03
		CPT	0.76 $\pm$ .02	0.68 $\pm$ .02	0.70 $\pm$ .02
		Ours	<b>0.79<math>\pm</math>.01</b>	<b>0.71<math>\pm</math>.02</b>	<b>0.73<math>\pm</math>.02</b>

Table 5: Results of suicide severity classification

the performance of the doctor model following four aspects: 1) *Fluency (Flu.)* assesses the smoothness of the whole conversation; 2) *Comforting (Com.)* measures the ability to empathize and comfort; 3) *Doctor-likeness (Doc.)* gauges the adaptability in shifting topics based on the patient’s situation; 4) *Engagingness (Eng.)* measures if the model sustains attention throughout the conversation.

As shown in Table 6, our model surpasses the baseline in the majority of aspects, particularly in *doctor-likeness* and *engagingness*, indicating significant improvements compared to other models. This suggests that incorporating the POST enhances our model’s ability to emulate doctors’ interactions, capturing the patient’s state in real-time, and allowing for flexible topic transitions throughout the dialogue. However, we also note that our model did not outperform ChatGPT in terms of *comforting*. This could be attributed to ChatGPT’s ability to generate extensive comforting and advisory responses without specific tuning. The dialogue examples of these models can be found in Appendix B.

Comparisons	Aspect	Win	Lose	Tie
Ours vs. CPT	Flu.	<b>44.7</b>	21.2	34.1
	Com.	<b>68.9</b>	15.2	15.9
	Doc.	<b>62.1</b>	14.4	23.5
	Eng.	<b>56.8</b>	11.4	31.8
Ours vs. ChatGPT	Flu.	<b>39.4</b>	34.8	25.8
	Com.	33.3	55.3	11.4
	Doc.	<b>53.8</b>	34.1	12.1
	Eng.	<b>41.7</b>	28.8	29.5

Table 6: Results of human evaluation

## 5.5 Ablation Studies

To verify the effectiveness of our method, we conducted ablation studies. We used the model incorporating the golden POST as the baseline. Then, we removed each component of the POST in turn to study the effect of each part of the POST. The obtained results are demonstrated in Table 7.

Model	BLEU-2	ROUGE-L	METEOR	DIST-2
Ours*	<b>45.28%</b>	<b>0.56</b>	<b>0.5794</b>	0.10
w/o Info	44.43%	0.54	0.5708	0.16
w/o Stage	43.57%	0.53	0.5681	0.16
w/o Sum	44.25%	0.54	0.5722	0.16
w/o Next	36.24%	0.45	0.5107	<b>0.18</b>

Table 7: Results of ablation studies

The study reveals that omitting the next topic (**w/o Next**) significantly impacts response generation, as it is directly correlated with the subsequent response’s relevance and effectiveness. This omission led to an increase in the DIST-2 value, suggesting a reduction in response content diversity due to the focused nature of topic-specific responses. Furthermore, excluding the stage (**w/o Stage**) notably affects the quality of generation, indicating the importance of the diagnostic dialogue phase in tailoring responses to be stage-appropriate and rational. In contrast, the absence of information (**w/o Info**) and summary (**w/o Sum**) components showed lesser impact on response generation quality, due to their roles in summarizing dialogue history rather than directly influencing responses. Nonetheless, these elements are crucial for their interpretability and utility in clinical settings, aiding doctors in summarizing patient symptoms and providing preliminary diagnostic insights, thereby facilitating further diagnosis and having a substantial role in clinical psychological consultations.

## 6 Conclusions

In this paper, we incorporate Psychological State Tracking (POST) within LLM to guide response generation for doctors during depression diagnosis consultations. In particular, the state is defined based on the Albert Ellis’ ABC Model in psychology, which illuminates a profound connection between patients’ information changes and doctors’ strategic planning. Extensive experiments show that the integration of psychological state tracking significantly enhances the performance of LLMs to generate responses in depression-diagnosis-oriented chat. Furthermore, our approach also provides explicit interpretations for using appropriate strategies in different situations to collect information or comfort patients for depression diagnosis. In the future, we will explore more specialized and fine-grained state tracking methods and incorporate patient personalized information to guide diagnosis-oriented chat.



## Limitations

The main limitation of our approach lies in the evaluation metrics. In the field of diagnostic chat, traditional automatic evaluation methods may not accurately reflect the true effectiveness of the models. As highlighted by Liu et al. (2016), there exists a certain discrepancy between the results and human evaluations. Developing more comprehensive, considerate, and task-specific evaluation metrics remains a challenge to be addressed in future research.

## Ethics Statement

In this study, we annotate the psychological state collaboratively by professional labors and GPT-4, followed by a thorough data cleaning and filtering process. This step ensured the removal of any private or sensitive information from the dataset.

## Acknowledgements

We acknowledge the contributions of the SJTU X-LANCE Lab<sup>5</sup> in constructing the original depression-diagnosis-oriented chat dataset (i.e., D<sup>4</sup>). We cite the work and conduct further research by adapting the dataset and exploring LLM-based method from the psychological views.

## References

- Guilherme F. C. F. Almeida, José Luiz Nunes, Neele Engelman, Alex Wiegmann, and Marcelo de Araújo. 2023. Exploring the psychology of gpt-4’s moral and legal reasoning.
- Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *proceedings of the 28th international conference on computational linguistics*, pages 1078–1089.
- Raúl Arrabales. 2020. Perla: a conversational agent for depression screening in digital ecosystems. design, implementation and validation. *arXiv preprint arXiv:2008.12875*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

<sup>5</sup><https://x-lance.sjtu.edu.cn/en>

- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenhao Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrmke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P Dinu. 2023. It’s just a matter of time: Detecting depression with time-enriched multimodal transformers. In *European Conference on Information Retrieval*, pages 200–215. Springer.
- Naomi Chaytor and Maureen Schmitter-Edgecombe. 2003. The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology review*, 13:181–197.
- Derek Chen, Kun Qian, and Zhou Yu. 2023a. Stabilized In-Context Learning with Pre-trained Language Models for Few Shot Dialogue State Tracking. *ArXiv:2302.05932 [cs]*.
- Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023b. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation.
- Zhiyu Chen, Yujie Lu, and William Wang. 2023c. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.
- Albert Ellis. 1991. The revised abc’s of rational-emotive therapy (ret). *Journal of rational-emotive and cognitive-behavior therapy*, 9(3):139–172.
- Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative

683	multimodal approach to depression detection in twitter. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 33, pages 110–117.	737
684		738
685		739
686	Thilo Hagendorff. 2023a. <a href="#">Deception abilities emerged in large language models</a> .	740
687		741
688	Thilo Hagendorff. 2023b. <a href="#">Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods</a> .	742
689		743
690		744
691	Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. <a href="#">LoRA: Low-rank adaptation of large language models</a> . In <i>International Conference on Learning Representations</i> .	745
692		746
693		747
694		748
695		749
696	Shashank Jaiswal, Michel Valstar, Keerthy Kusumam, and Chris Greenhalgh. 2019. Virtual human questionnaire for analysis of depression, anxiety and personality. In <i>Proceedings of the 19th ACM international conference on intelligent virtual agents</i> , pages 81–87.	750
697		751
698		752
699		753
700		754
701	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>Advances in neural information processing systems</i> , 35:22199–22213.	755
702		756
703		757
704		
705		
706	Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. <i>Journal of general internal medicine</i> , 16(9):606–613.	758
707		759
708		760
709		761
710	Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. <a href="#">Dialogue State Tracking with a Language Model using Schema-Driven Prompting</a> . ArXiv:2109.07506 [cs].	762
711		763
712		764
713		
714	Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. <a href="#">Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension</a> .	765
715		766
716		767
717		768
718		769
719	Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. <a href="#">A diversity-promoting objective function for neural conversation models</a> .	770
720		
721		
722	Chin-Yew Lin. 2004. <a href="#">ROUGE: A package for automatic evaluation of summaries</a> . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	771
723		772
724		773
725		774
726	Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021. <a href="#">Leveraging Slot Descriptions for Zero-Shot Cross-Domain Dialogue State Tracking</a> . ArXiv:2105.04222 [cs].	775
727		776
728		777
729		778
730		779
731		
732	Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. <a href="#">How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation</a> . In <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i> , pages 2122–2132, Austin, Texas. Association for Computational Linguistics.	780
733		781
734		782
735		783
736		784
	Mingyu Derek Ma, Jiun-Yu Kao, Shuyang Gao, Arpit Gupta, Di Jin, Tagyoung Chung, and Nanyun Peng. 2023. <a href="#">Parameter-Efficient Low-Resource Dialogue State Tracking by Prompt Tuning</a> . ArXiv:2301.10915 [cs].	785
		786
		787
		788
		789
		790
		791
	Moisés R Pacheco-Lorenzo, Sonia M Valladares-Rodríguez, Luis E Anido-Rifón, and Manuel J Fernández-Iglesias. 2021. Smart conversational agents for the detection of neuropsychiatric disorders: a systematic review. <i>Journal of Biomedical Informatics</i> , 113:103632.	
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. <a href="#">Bleu: a method for automatic evaluation of machine translation</a> . In <i>Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics</i> , pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.	
	Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2023. <a href="#">Semantic similarity models for depression severity estimation</a> . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 16104–16118, Singapore. Association for Computational Linguistics.	
	Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. 2017. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. <i>Scientific reports</i> , 7(1):42656.	
	Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. <a href="#">Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media</a> .	
	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. <a href="#">Exploring the limits of transfer learning with a unified text-to-text transformer</a> . <i>J. Mach. Learn. Res.</i> , 21:140:1–140:67.	
	Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. <i>arXiv preprint arXiv:2109.05729</i> .	
	Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. <a href="#">A computational approach to understanding empathy expressed in text-based mental health support</a> . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 5263–5276, Online. Association for Computational Linguistics.	

- Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.
- Zhoujian Sun, Zhengxing Huang, and Nai Ding. 2022. [On Tracking Dialogue State by Inheriting Slot Values in Mentioned Slot Pools](#). ArXiv:2202.07156 [cs].
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#).
- Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. [VecoCare: Visit Sequences-Clinical Notes Joint Learning for Diagnosis Prediction in Healthcare Data](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4921–4929, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization.
- Yuting Yang, Wenqiang Lei, Pei Huang, Juan Cao, Jintao Li, and Tat-Seng Chua. 2023. [A Dual Prompt Learning Framework for Few-Shot Dialogue State Tracking](#). ArXiv:2201.05780 [cs].
- Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. [D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Dian Yu, Mingqiu Wang, Yuan Cao, Laurent El Shafey, Izhak Shafran, and Hagen Soltau. 2022. [Knowledge-grounded Dialog State Tracking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3428–3435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021. [Depa: Self-supervised audio embedding for depression detection](#). In *Proceedings of the 29th ACM International Conference on Multimedia*, MM ’21, page 135–143, New York, NY, USA. Association for Computing Machinery.

## A Prompt Template

**Description:** I want you to serve as a psychological counselor, and use rational emotive behavior therapy and ABC theory to mark the current patient situation and the next strategy. Please mark the input dialogues according to the requirements and examples in Chinese. The sequence numbers of the input dialogues must strictly correspond one-to-one with those of the outputs, with no merging or omissions allowed.

**Requirements:** Definitions and marking requirements of four parts of POST:

<Stage> The current diagnostic stage (Activating Event (A) / Belief (B) / Consequence (C), please enter the complete stage, and answers like "same as above" are not allowed).

<Info> Current patient information (keep information from historical conversations as briefly as possible).

<Sum> Diagnosis result of the patient's condition (you need to refine it to the underlying disease).

<Next> Next step strategy (choose the most suitable option from 11 strategies).

**Examples:** [Examples pre-labeled by humans]

**Input:** [Input data]

**Output:**

## B Case Study

Figures 5, 6, 7 respectively display dialogue examples of our model, CPT, and ChatGPT under the same patient background, where these models play the role of doctors. In the diagram, statements of empathetic consolation are highlighted with green wavy lines, while inquiries of diagnosis are underlined in blue. It is evident from the illustration that, in comparison to other models, our model offers more natural empathetic consolation and topics of inquiry that are closely aligned with the diagnostic state. The responses from CPT tend to be relatively rigid, with less empathetic consolation and more straightforward, singular diagnostic inquiries. Although ChatGPT provides commendable empathetic consolation, its replies often are too lengthy, containing extensive analysis and advice, yet they do not closely correlate with the diagnostic content, lacking the distinctive characteristics of a doctor.



Background	
<b>Age:</b> 22 years old	<b>Gender:</b> Female
<b>Occupation:</b> Student	<b>Marital status:</b> Single
<b>Depression risk:</b> Moderate	<b>Suicide risk:</b> Mild
<b>Specific symptoms:</b> Reduced interest, loss of interest in past hobbies; Difficulty concentrating, difficulty making decisions; Lack of self-confidence, feelings of guilt, hopelessness; Impaired personal life functioning, unstable interpersonal relationships, reduced work and study efficiency.	
Ours	
<p><b>Patient:</b> 开始的时候只是觉得有些情绪低落,但随着时间的推移,症状越来越明显,影响到我的日常生活和工作。I initially only felt somewhat down, but over time the symptoms became more evident and affected my daily life and work.</p> <p><b>Doctor:</b> 嗯嗯,如果长时间出现情绪低落的状态,肯定会影响到正常的工作生活。那你觉得是什么原因导致的呢?是发生了什么事情吗? Hmm, long-term low mood would certainly affect normal work and life. What do you think caused it? Did something happen?</p> <p><b>Patient:</b> 我觉得自己的情绪低落可能与一些生活上的压力和挫折有关。最近工作上遇到了一些挑战和困难,导致我失去了对自己能力的信心。同时,我也感到人际关系不稳定,有些关系出现了问题,这也给我带来了一定的心理负担。I think my low mood may be related to some life pressures and setbacks. Recently at work I encountered some challenges and difficulties that caused me to lose confidence in my abilities. At the same time, I felt interpersonal relationships were unstable and some relationships had issues, which also brought me some psychological burden.</p> <p><b>Doctor:</b> 确实,生活中总会有一些不如意的事情发生,我们也要学会调整心态去面对它们。那你最近有没有感觉自己很疲倦? You're right. There will always be some disappointments in life, and we need to learn to adjust our mindset to face them. Have you been feeling very tired recently?</p> <p><b>Patient:</b> 是的,我最近感到非常疲倦。即使我休息好了一晚上,仍然感觉没有精力去做事情。我往往感到身体沉重,动作迟缓,精神也不够集中。这种疲倦感持续了一段时间,对我的日常生活和工作造成了很大的困扰。Yes, I've been feeling extremely tired recently. Even if I've had a good night's rest, I still feel like I don't have the energy to do things. I often feel my body is heavy and my movements are sluggish, and my spirit is not concentrated enough. This tiredness has lasted for some time and caused me great trouble in my daily life and work.</p> <p><b>Doctor:</b> 身体上的疲惫也是抑郁症的一个症状。那最近睡眠怎么样呀? 睡眠时间会比较短吗? Physical fatigue is also a symptom of depression. How has your sleep been recently?</p> <p><b>Patient:</b> 是的,我的睡眠时间确实相对较短。通常情况下,我只能睡 4-6 个小时,而且还是断断续续的。这远远不能满足我身体的休息需求,导致我在白天感到更加疲倦和无精神。无法获得充足的睡眠对我来说是一个很大的困扰。Yes, my sleep time is relatively short. Usually, I can only sleep 4-6 hours, and it's intermittent. This is far from enough to meet the rest needs of my body, causing me to feel even more tired and lack energy during the day. Not being able to get sufficient sleep is a big problem for me.</p> <p><b>Doctor:</b> 睡眠对于人体来说是很重要的,我们还是要保证充足的睡眠时间。那你最近吃饭怎么样呀? Sleep is very important for the body, and we should ensure sufficient sleep time. How has your appetite been recently?</p>	

Figure 5: Dialogue example of our model

CPT
<p><b>Patient:</b> 大部分时间我都是处于这种低落的状态。很难感到开心或者有精神,情绪总是低迷的。For most of the time I am in this low mood. It is difficult to feel happy or spirited, my emotions are always depressed.</p> <p><b>Doctor:</b> <u>好的,我了解了,你最近生活中有发生什么事情吗? Okay, I understand. Has anything happened in your life recently?</u></p> <p><b>Patient:</b> 最近我没有遇到什么特别的事情,生活比较平静。但是由于情绪低落的原因,我对日常生活的一些事情也没有太多兴趣和动力去做。所以有时候觉得自己的生活有些乏味。Recently nothing special has happened, my life has been relatively calm. But due to my low mood, I have little interest or motivation to do some ordinary daily things. So sometimes I feel my life is a bit bland.</p> <p><b>Doctor:</b> <u>有和家人朋友联系倾诉,寻求他们的支持吗? Have you been in contact with family and friends to share and seek their support?</u></p> <p><b>Patient:</b> 我很难和家人朋友联系倾诉,寻求他们的支持。因为我觉得他们可能无法理解我的情绪和困扰,或者我觉得自己给他们带来麻烦和负担。所以我更多地选择独处,很少主动和他们交流。I find it difficult to contact family and friends to share and seek their support. Because I feel they may not understand my emotions and troubles, or I feel I bring them trouble and burden. So I prefer solitude more, and rarely take initiative to communicate with them.</p> <p><b>Doctor:</b> <u>好的,我了解了,这种状态持续多久了呢? Okay, I understand. How long has this state lasted?</u></p> <p><b>Patient:</b> 我这种状态已经持续了几个月的时间了,大约是从半年前开始的。一开始可能只是偶尔感到不开心,但是渐渐地情绪低落的频率和程度都增加了。现在已经达到了一个相对稳定的状态,持续了相当一段时间了。I have been in this state for several months, approximately since half a year ago. At the beginning it may have just been occasionally feeling unhappy, but gradually the frequency and degree of low mood increased. Now it has reached a relatively stable state and lasted for a considerable time.</p> <p><b>Doctor:</b> <u>好的,我了解了,最近你会觉得很疲倦吗? Okay, I understand. Have you been feeling very tired recently?</u></p> <p><b>Patient:</b> 是的,最近我经常感到非常疲倦。即使有足够的睡眠,起床后仍然感到没有精神。进行日常活动也感到非常吃力,经常会出现身体乏力和精神上的疲倦感。Yes, recently I often feel extremely tired. Even with enough sleep, I still feel lacking in spirit after getting up. Daily activities also feel very laborious, and there are often feelings of physical weakness and mental fatigue.</p> <p><b>Doctor:</b> <u>好的,我了解了。最近生活中有发生什么事情吗? Okay, I understand. Has anything happened in your life recently?</u></p>

Figure 6: Dialogue example of CPT

## ChatGPT

**Patient:** 最近两周,我的心情一直都很低落,感觉就像有一种沉重的阴影笼罩在我身上。每天都感觉很沮丧,缺乏动力去做任何事情。For the past two weeks, my mood has been very low, feeling like a heavy shadow is cast over me. Every day I feel dejected and lack motivation to do anything.

**Doctor:** 嗯,我能理解你现在的困扰。那你最近睡眠情况怎么样呢?是不是也出现了问题?  
Hmm, I understand your current troubles. How has your sleep been recently? Have there been any issues?

**Patient:** 是的,我的睡眠情况也很糟糕。我经常难以入睡,即使入睡了也容易醒来,并且早上起床时感觉疲倦不堪。我常常在夜晚反复辗转反侧,睡眠质量很差。Yes, my sleep has also been very bad. I often have difficulty falling asleep, and even if I fall asleep, I easily wake up, and feel extremely tired when I wake up in the morning. I often toss and turn at night, with poor sleep quality.

**Doctor:** 嗯,睡眠问题对于情绪低落来说是相互影响的。那你的食欲情况怎么样呢?最近有没有食欲不振或者过度进食的情况出现?  
Hmm, sleep issues can mutually influence low mood. How has your appetite been? Have there been issues with lack of appetite or overeating recently?

**Patient:** 我的食欲也受到了影响。我最近常常没有食欲,对于平时喜欢的食物也提不起兴趣。有时候,我会因为情绪低落而暴饮暴食,过度进食。My appetite has also been affected. Recently I often have no appetite and cannot work up any interest in foods I usually like. Sometimes when my mood is low, I will overeat or binge eat.

**Doctor:** 嗯,好的,我了解你的情况了。根据你描述的症状,你可能正在经历一段抑郁状态。情绪低落、失去兴趣、睡眠问题、食欲变化以及社交功能下降都是抑郁症的常见症状。接下来,我会进一步提问,确保我们对你的情况有一个全面的了解。请问你最近有没有自杀倾向呢?  
Hmm, okay, I understand your situation now. Based on the symptoms you described, you may be experiencing a period of depression. Low mood, loss of interest, sleep issues, appetite changes and decreased social functioning are all common symptoms of depression. Next, I will ask further questions to ensure we have a comprehensive understanding of your situation. Do you have any suicidal thoughts recently?

**Patient:** 没有,我没有自杀倾向。尽管我感到绝望和无望,但我并没有想过要伤害自己或者结束自己的生命。No, I don't have suicidal thoughts. Although I feel despair and hopelessness, I haven't thought about harming myself or ending my life.

**Doctor:** 好的,我了解你的情况了。抑郁症是一种常见的心理健康问题,但它是可以得到缓解和治疗的。我建议你尽快与专业心理咨询师或医生进行面对面的咨询和评估。他们会根据你的情况给出更具体的建议和治疗方案。同时,你也可以尝试一些自我调节的方法,如保持规律的作息时间、积极参与有意义的活动、与亲友进行交流和寻求支持等。请记住,你并不孤单,我们会一起度过这段困难时期。Okay, I understand your situation now. Depression is a common mental health issue, but it can be relieved and treated. I recommend you see a professional psychologist or doctor for in-person consultation and assessment as soon as possible. Based on your situation, they will provide more specific recommendations and treatment plans. In the meantime, you can also try some self-regulation methods like maintaining regular sleep schedules, actively participating in meaningful activities, communicating with family and friends for support. Remember that you are not alone, we will get through this difficult time together.

Figure 7: Dialogue example of ChatGPT