D^4: A Psychiatrist-proofread Dialogue Dataset for Depression Diagnosis

Anonymous ACL submission

Abstract

Depression has affected large populations and become a significant threat to life expectations globally. Automatic depression diagnosis methods have been a new research focus. In particular, automatic dialogue-based diagnosis systems are desired since depression diagnosis highly relies on clinical consultation. Based on clinical diagnosis criteria, doctors initiate a conversation with ample emotional support that guides the patients to expose their symptoms. Such a dialog is a combination of task-oriented and chitchat, different from traditional single-purpose human-machine dialog systems. However, due to the social stigma associated with mental illness, the dialogue data related to the diagnosis of actual patients are rarely disclosed. The lack of data has become one of the major factors restricting the research on the consultation dialogue system of depression. Based on clinical depression diagnostic criteria ICD-11 and DSM-5, we construct a *Psychiatrist-proofread Dialogue Dataset for Depression Diagnosis* which simulates the dialogue between the doctor and the patient during the diagnosis of depression and provides diagnosis results and symptom summary given by professional psychiatrists for each dialogue. Finally, we finetune on state-of-art pre-training models and respectively give our dataset baselines on response generation, topic prediction, dialog summary, and severity classification of depression and suicide risk.

1 Introduction

According to the World Health Organization (The World Health Organization, 2021a), approximately 280 million people worldwide have depression, a leading cause of disability worldwide and a significant contributor to the overall global burden of disease. The worst thing is, depression can lead to suicide. However, in contrast to the high prevalence and severe social harm, limited medical resources exhibit difficulties for people in remote or low-income areas to receive diagnosis and treatment. The social stigma (Ben-Zeev et al., 2010) associated with mental illness is also an obstacle to the inability of depressed patients to seek medical treatment in time.

Researchers have been exploring effective methods for depression detection and diagnosis. Besides online self-rating scales, such as 9-item Patient Health Questionnaire(PHQ-9) (Kroenke and Spitzer, 2002) and the Beck Depression Inventory (BDI) (Beck et al., 1996), there is research work on automatic depression detection from posts of social media (Orabi et al., 2018), speech (Cummins et al., 2015) and multi-modality (Cummins et al., 2013), trying to find objective clues applicable for diagnosis from a patient's language, speech, and facial expressions. However, they greatly differ from clinical practice, lacking interpretability and transferring ability. Since the golden standard in depression diagnosis is drawn from a qualified psychiatrist through clinical consultations, these signal-based objective detection methods play a partial role. They are hardly possible to directly apply in depression screening for large-scale populations.

Building an automated dialogue system imitating such a consultation procedure would be ideal for the aforementioned problems. Through dialogue, we can better understand the patient's mood and cognitive state from an objective perspective, which is closely related to the diagnosis of depression and can provide emotional support at the same time. Further, the dialogue system can provide a generalizable carrier for the diagnosis methods based on text, speech and multi-modality. The interpretability of diagnosis can be effectively enhanced through detailed dialogue annotation. Therefore, we believe that a consultation dialogue system is of great value for large-scale depression screening and regular return visits to depressed patients.

Nevertheless, clinical diagnosis is a complex procedure, with the purpose to collect and summarize key symptom information about one patient while providing a chat-like conversation experience. In clinical practice, psychiatrists communicate with patients and provide diagnosis results based on practical experience and multiple diagnostic criteria. The most clinically-adopted criteria involves ICD-11 (The World Health Organization, 2021b), DSM-5 (American Psychiatric Association, 2013), etc., which defines core symptoms for depression diagnosis. At the same time, psychiatrists provide emotional support such as empathy and comfort during the consultation to better prompt patients' self-expression. This explains why even using the same diagnosis scale, self-rated results are only a reference while clinical consultation draws the conclusive diagnosis.

Such a diagnosis-directed dialogue system is extremely challenging hence still under-investigated. Currently, there are no datasets specified for this purpose, partially due to the high correlation between the specific conditions of mental illness and patients' privacy. Some mental health-related dialogue system research (Saha et al., 2021) endeavored to crawl patients' blog posts and comment data from public websites to obtain dialogues, while this kind of dataset lacks professionalism to support the rigorous depression diagnosis procedure.

To construct a dialogue system capable of professional depression diagnosis and emotional support, we conduct dialogue collection through consultation dialogue simulation. However, due to medical data privacy protection, it is highly impossible to collect large-quatity dialogues between real patients and psychiatrists. Traditional dialogue collection methods such as pure crowdsourcing (Daniel et al., 2018) might diverge from clinical protocol and inhabit limited depressive symptoms. To overcome these difficulties, we propose **D**^{4}: a **D**ialogue **D**ataset for **D**epression **D**iagnosis, for which we endeavor to mimic a real-world clinical consultation scene and involve three phases in collecting diagnostic dialogues (see Figure 1). **P1:** To *simulate medical records*, we collect actual patients' portraits with a consultation chatbot web app that asks users fixed questions abstracted from clinical depression diagnosis criteria ICM-11 and DSM-5. **P2:** To *restore psychiatric consultation conversations*, we employ workers to conduct consultation the dialogue simulation based on the collected portraits. The workers are divided into patients and doctors for separate training by professionals. **P3:** To *reinforce the clinical setting*, professional psychiatrists and psychotherapists supervise the whole process and filter out unqualified dialogues. In addition, they provide diagnosis summaries based on the portrait and dialogue history.

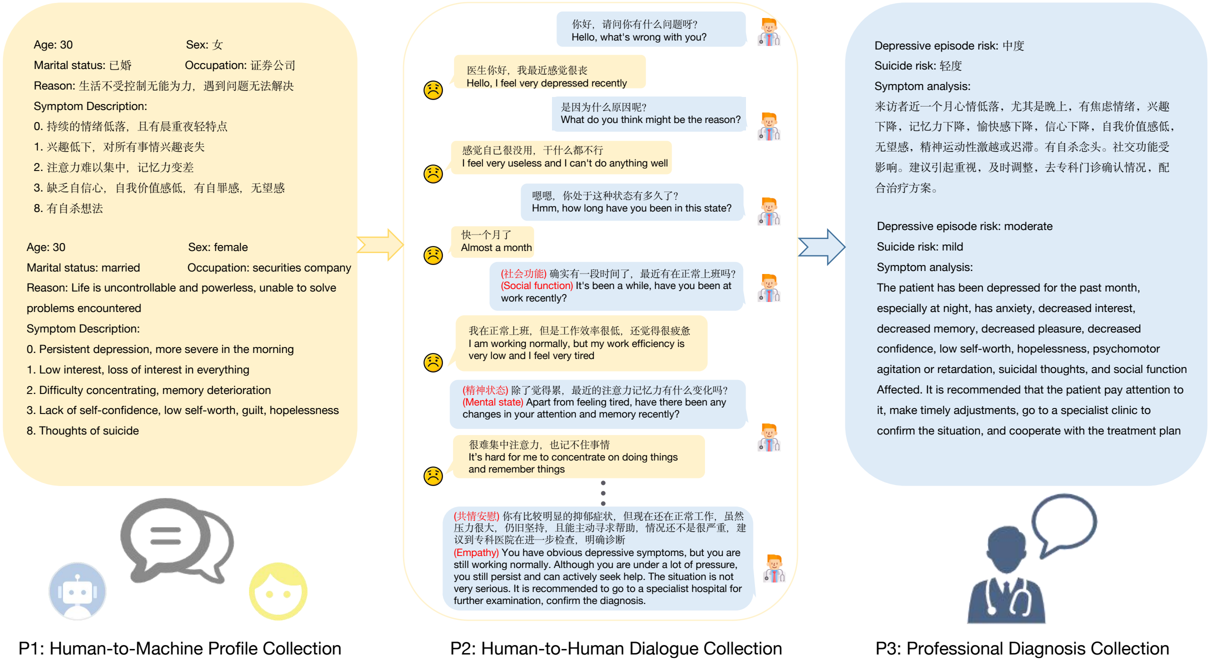
Hereby, we construct a psychiatrists-proofread depression diagnosis dataset with 1,339 conversations as well as doctor-prescribed diagnosis results (including the severity of depression episode and symptom summary). We annotate the conversation procedure with 10 topic tags (grouped by core depressive symptoms listed in DSM-5 and ICD-11). Multi-dimensional analysis suggests that our simulated diagnosis data are reliable and up to professional standards. Experiments on generation, summarization and classification tasks further validate the dataset's purpose in constructing a clinical-practical human-to-machine diagnosis dialogue system.

The key contribution of this paper are as follows:

- A up-to-clinical-standard depression diagnosis dataset with 1,339 conversations generated from actual populations' portraits, accompanied by psychiatrists' diagnosis summaries, under the framework of most applied clinical diagnosis criteria ICD-11 and DSM-5;
- Experimental validation on multiple tasks: response generation, topic prediction, dialog summary, and severity classification of depression and suicide risk;
- To the best of our knowledge, this is the first diagnosis dialogue dataset for mental health, enabling the realization of an avante-garde clinical diagnosis dialogue system that combines symptom inquiry and emotional support.

2 Data Collection

In this section, we respectively detail our 3-phase collection paradigm: 1) real populations' portraits (in particular depressive patients) collected to form pre-diagnosis records; 2) simulated natural diagnostic consultations based on the portraits; 3) psychiatrists proofread dialogue history and prescribed professional diagnose summaries.



Age: 30 Sex: 女
Marital status: 已婚 Occupation: 证券公司
Reason: 生活不受控制无能为力，遇到问题无法解决
Symptom Description:
0. 持续的情绪低落，且有晨重夜轻特点
1. 兴趣低下，对所有事情兴趣丧失
2. 注意力难以集中，记忆力变差
3. 缺乏自信心，自我价值感低，有自罪感，无望感
8. 有自杀想法

Age: 30 Sex: female
Marital status: married Occupation: securities company
Reason: Life is uncontrollable and powerless, unable to solve problems encountered
Symptom Description:
0. Persistent depression, more severe in the morning
1. Low interest, loss of interest in everything
2. Difficulty concentrating, memory deterioration
3. Lack of self-confidence, low self-worth, guilt, hopelessness
8. Thoughts of suicide

P1: Human-to-Machine Profile Collection

你好，请问你有什么问题呀？
Hello, what's wrong with you?

医生你好，我最近感觉很丧
Hello, I feel very depressed recently

是因为什么原因呢？
What do you think might be the reason?

感觉自己很没用，干什么都不行
I feel very useless and I can't do anything well

嗯嗯，你处于这种状态有多久了？
Hmm, how long have you been in this state?

快一个月了
Almost a month

(社会功能) 确实有一段时间了，最近有在正常上班吗？
(Social function) It's been a while, have you been at work recently?

我在正常上班，但是工作效率很低，还觉得很疲惫
I am working normally, but my work efficiency is very low and I feel very tired

(精神状态) 除了觉得累，最近的注意力记忆力有什么变化吗？
(Mental state) Apart from feeling tired, have there been any changes in your attention and memory recently?

很难集中注意力，也记不住事情
It’s hard for me to concentrate on doing things and remember things

(共情安慰) 你有比较明显的抑郁症状，但现在还在正常工作，虽然压力很大，仍旧坚持，且能主动寻求帮助，情况还不是很严重，建议到专科医院在进一步检查，明确诊断
(Empathy) You have obvious depressive symptoms, but you are still working normally. Although you are under a lot of pressure, you still persist and can actively seek help. The situation is not very serious. It is recommended to go to a specialist hospital for further examination, confirm the diagnosis.

P2: Human-to-Human Dialogue Collection

Depressive episode risk: 中度
Suicide risk: 轻度
Symptom analysis:
来访者近一个月心情低落，尤其是晚上，有焦虑情绪，兴趣下降，记忆力下降，愉快感下降，信心下降，自我价值感低，无望感，精神运动性激越或迟滞。有自杀念头。社交功能受影响。建议引起重视，及时调整，去专科门诊确认情况，配合治疗方案。

Depressive episode risk: moderate
Suicide risk: mild
Symptom analysis:
The patient has been depressed for the past month, especially at night, has anxiety, decreased interest, decreased memory, decreased pleasure, decreased confidence, low self-worth, hopelessness, psychomotor agitation or retardation, suicidal thoughts, and social function Affected. It is recommended that the patient pay attention to it, make timely adjustments, go to a specialist clinic to confirm the situation, and cooperate with the treatment plan

P3: Professional Diagnosis Collection

Figure 1: The 3-Phases Data Collection: P1, P2, and P3 denotes the three phases in data collection

Risk	Control	Mild	Moderate	Severe
Depression	264	49	95	70
Suicide	338	46	75	19

Table 1: Risk Estimation of portraits: "control" represents no risk, "mild", "moderate", and "severe" represent the severity of the risk respectively

2.1 Profile Collection

To overcome the impracticability in obtaining patients’ medical records covered by doctor-patient confidential protocol, we designed a consultation chatbot based on a state machine, which utilizes fixed questions from clinical criteria to document each user’s depressive symptoms and *demographic information such as age, gender, marital status and occupation*. Core *depression symptoms* are prompted accordingly, including *mood, interests, mental states, sleep, appetite, social functions and suicidal tendency*. The users are invited to respond concisely, e.g., yes/no answer and severity estimation. Combined, we obtain a voluntary, legit depression portrait. As of the submission of the paper, we have collected a total of 478 patient portraits. We estimate the severity of depressive episodes and suicide risk based on clinical criteria ICD-11 and DSM-5 for each patient portrait and the result is shown in Table 1. 68 portrait providers reported that they had been diagnosed with depression in an authorized clinic. Among these providers, 53 are currently experiencing a depressive episode.

2.2 Consultation Conversation Simulation

To guarantee the quantity, quality and professionalism of our consultation dialogues, we conducted conversation simulation under the guidance of psychiatrists, following true depression patients’ self-reports of patients. In particular, we first gathered a small number of dialogues between doctors and patients in real scenarios. Based on above mentioned prerequisites and clinical depression diagnosis criteria ICD-11 and DSM-5, we released the simulation tasks to crowdsourcing workers. The whole procedure is introduced accordingly: 1) Design and Training: the workers first go through specialized training and then divided into doctor and patient roles; 2) Annotation: during the conversation, they are required to annotate topic/symptom transitions; 3) Peer Assessment: doctor and patient roles rate each other on multiple dimensions after the conversation.

2.2.1 Design and Training

Acting Patients Most of our patient actors are not depressed. To help them understand the symptoms in the patient portraits, we provide detailed explanations of the symptoms, including the severity and duration, and some patient self-reports to help them understand patients’ inner feelings. Based on the accurately expressed symptoms, they need to imagine possible life events of the portrait's provider and talk with a doctor about it to express the patient's inner feelings in the process of telling the events.

Acting Doctors Firstly, we invite psychiatrists and clinical psychotherapists to initiate consultation conversations with actual depressive patients, from which we collect reference conversations. Then based on what they asked, combined with ICD-11 and DSM-5, we compile the information that doctors need to know when diagnosing depression. The doctor actors are required to obtain sufficient information in their conversations with the patient and empathize and reassure the patients when they confide in them about what they are experiencing. At the end of the conversation, they need to remind patients who might be depressed to seek timely medical attention.

2.2.2 Annotation

Topic Annotation According to core symptoms covered in clinical criteria, we categorized the topics into *mood, interests, mental status, sleep, appetite, somatic symptoms, social functioning, suicidal tendency, and screening*. Notably, we included *empathy* as a special topic since it is an essential part in clinical practice. The doctor actors were asked to annotate the topic of their messages during the chat.

2.2.3 Peer Assessment

Patient Role Assessment We assess the patient task performance on three dimensions: the naturalness of the expression, the consistency of the narrative, and the extent to which the severity of the symptoms described and the expression match. The first two scores are given by the other participant in the conversation, and the third score is given by the medical professional examining the conversation.

Doctor Role Assessment We assess the completion of the doctor's task by the degree to which the actor resembles a real doctor. This degree is assessed by the other participant of the conversation and a professional doctor, respectively.

2.3 Quality Control

Hierarchical screenings are conducted to control the data quality: whether it is up to clinical standard and can satisfy training purpose.

Aspects	Rating Content	Minimum
Patient	expression naturalness	3(5)
	narrative consistency	3(5)
	matching extent of symptom severity and expression*	3(5)
Doctor	degree of similarity to the doctor	3(5)
	degree of similarity to the doctor*	3(5)
	Avg.length of utterances	8
Total	Avg. utterances per dialogue	30

Table 2: Quality Control Criteria: Scores* is given by psychiatrists, the rest are obtained by peer assessment; Numbers in parentheses = the highest score

Category	Total	Patient	Doctor
Dialogues	1339	-	-
Avg. turns	21.6	-	-
Workers	201	127	74
Avg. utterances per dialogue	60.9	30.9	29.9
Avg. utterance length	12.5	10.4	14.6
Avg. diagnosis length	83.1	-	-

Table 3: D^4 Statistics

Psychiatrists' Clinical Protocol Screening To ensure the accordance with clinical protocol, we further invite professional psychiatrists and clinical psychotherapists for dialogue assessment. They screen the dialogues that meet the diagnostic needs and provide psychiatric diagnostic results and symptom summaries. They score the doctors and the patients separately with the real-scenario resemblance degree.

Objective Quality Screening For better training purpose, we adopt a variety of paradigms to conduct quality examination. We set minimum limits on the length of the dialogue, the average utterance length per dialogue of the doctor, the mutual scores and the scores given by the psychiatrist shown in the Table 2. The unqualified dialogues are excluded.

Ultimately, we collect a total of 4,428 conversations and finally retained 1,339(30%) after the uncompromised quality screenings.

3 Data Characteristics

3.1 Statistics

The overall statistics of the dataset are shown in Table 3. As seen in such a diagnosis scenario, sufficient dialogue turns are required: our diagnosis dialogue exhibit avg. 21.6 turns per dialogue. The symptom summaries provided by psychologists summit to 83.1 words on average. These statistics are significantly longer than previous related

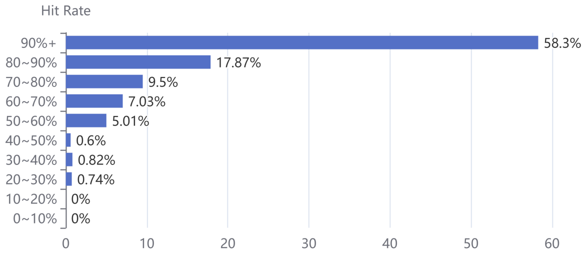


Figure 2: Consistency between doctor's diagnosis and patient's portraits. We convert each dialogue's symptoms of the patient portrait into the category of psychologists' symptom summary, then calculate the hit rate. The hit rates are divided into ten ranges in vertical axis, and the horizontal axis is the percentage of corresponding range.

datasets, suggesting the discrepancies of a diagnosis dialogue task along with its distinguished data requirements.

3.2 Consistency Evaluation

It should be noted that the symptoms simulated in our dialogue data are based on true populations' portraits. In order to verify that the simulated dialogue does reflect the symptoms in the patient's portrait, and effective diagnostic conclusions can be drawn from such dialogues, we analyze the consistency of the patient portrait and the corresponding content of the psychiatrist's symptom summary, as well as consistency of the diagnosis results from different doctors for the same portrait.

Portrait-Diagnosis Consistency The patient portrait contains depressive symptoms, based on which, the patient actor added more details in the actual description, leading to a diagnosis summary covering more content. Thus, We utilize the hit rate of the doctor's diagnosis summary (Figure 2) to measure the consistency. It can be seen that most diagnoses have high accuracy with an average of 86.1%, demonstrating the authenticity of patient imitation and the comprehensiveness of the summary. Besides, we ask psychiatrists to rate the matching extent of patients' expressions and the severity of their symptoms, 3.9 on average with the total score of 5, meaning that the degree of conformity reached 78%.

Doctors' Consistency The diagnosis of the same patient portrait from different doctors can have slightly different results. For the three indicators shown in Table 4, we compute the mean of different portraits' standard derivations (portraits with sole diagnosis is excluded), suggesting a high agreement on the diagnosis results, slightly affected by workers' subjectivity.

	Depression Severity	**Suicide Severity**	**Symptom Number**
Avg. S.D of doctors	0.500	0.366	0.516

Table 4: Consistency Analysis of Doctors' Diagnosis for the Same Patient portrait. The value ranges of depression severity and suicide severity = (0,1,2,3). Max value of symptom number = 9.

Category	**Control**	**Mild**	**Moderate**	**Severe**
Dialogues	430	342	368	199
Avg. turns	17.9	21.3	23.7	26.0
Avglen. of doctor dialogues	25.2	29.5	32.5	36.3
Avglen. of patient dialogues	25.1	30.7	34.7	37.1
1st frequent topic	Emp.	Emp.	Emp.	Emp.
2nd frequent topic	MS	MS	MS	Suicide
3rd frequent topic	Sleep	Senti.	Suicide	MS
Avglen. of diagnosis	58.1	80.9	99.5	110.6

Emp.:Empathy **Senti.**:Sentiment **MS**:Mental State

Table 5: Depression Severity Statistics in D^4

3.3 Analysis of Different Depression Severity

Distribution Feature We present the statistics of patients with different severity of depression episodes in Table 5. As the degree of depression worsens, the turns and dialog lengths get longer due to doctors' more in-depth questions on specific topics. The content in diagnosis summaries becomes longer to list more depressive symptoms. The most frequent topics are also subject to change with severity: "Suicide Tendency" is more likely to be questioned among severer patients.

Lexical Analysis of Symptom Summary From Figure 3, we observe a great difference on hot words from diagnosis summaries of different severity. As shown in chart (a), control patients mostly have superficial symptoms like decision difference and confidence declination, commonly exists in healthy populations. As the condition worsens, more obvious symptoms like sleep difficulty appear frequently, and doctors will suggest patients take timely measures. Among the severest patients, suicide risk and hopelessness become frequent from chart (d).

3.4 Analysis of Topics

Topic Distribution The statistics of different topics' features are shown in Table 6. Core depressive symptoms occupy 68.3% of the conversation, followed by Empathy at 23.1%. By analysing the

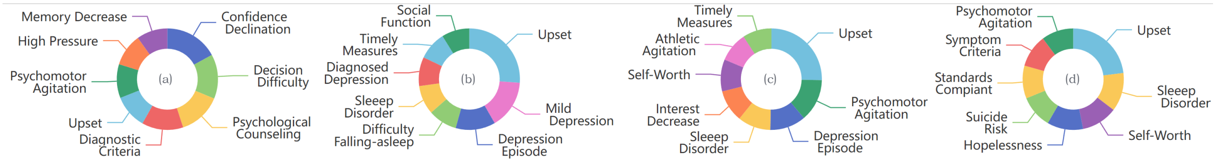


Figure 3: The most frequent word distribution in symptom summary with different depression severity. We compute the top-eight-frequent words, severity increases from chart (a) to chart(d)

topic first-appears we can see that minor symptoms like *sentiment*, *interest* are usually inquired in the beginning, gradually move to *suicide* and *somatic symptoms*, which are usually experienced by severe patients. This echoes clinical practice where a consultation follows a gradual in-depth manner and provides emotional support from time to time.

Topic	Proportion	Avg. Turn of First Appear
Sleep	9.96%	12.0
Sentiment	9.1%	6.10
Screening	4.55%	20.4
Interest	6.35%	6.75
Mental State	13.0%	10.4
Social Function	8.01%	10.8
Appetite	8.19%	15.1
Suicide	9.18%	15.6
Empathy	23.1%	16.7
Somatic Symptoms	8.56%	16.4

Table 6: Statistics of Topics

Topic Transition Figure 4 illustrates the topic-transition process. Different from other commonly seen dialogues where topic rarely extends over one turn, diagnosis topics consistently occur across turns. Empathy stays a stable ratio throughout all phases due to the particularity of the depression diagnosis scene. Suicide is usually conducted at a later phase due to its sensitivity.

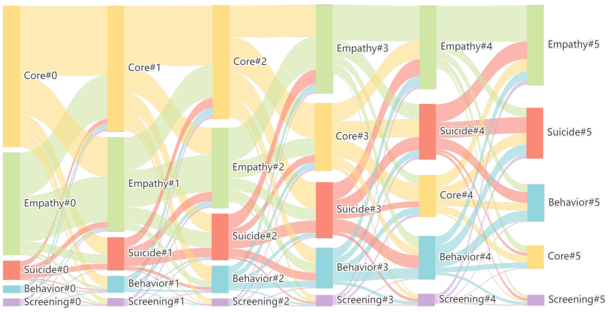


Figure 4: Topic transitions. For brief, Sleep, Appetite, Somatic Symptom grouped into Behavior, Sentiment, Interest,Mental State,Social into Core. Topics over every three turns are visualized. The height represents the absolute number of this category.

Dataset	Domain	Dialogues	Avg.turns	Avg.utterances
MultiWOZ (Budzianowski et al., 2018)	Restaurants, Hotels, etc	8,438	13.46	-
MotiVAte (Saha et al., 2021)	Mental Health	4,000	-	3.70
ESConv (Liu et al., 2021)	Emotinal Support	1,053	-	29.8
MedDialog (Zeng et al., 2020)	Medical Dialogue	3,407,194	-	3.3
DAIC-WOZ (Gratch et al., 2014)	Distress Analysis	189	-	-
D^4(Ours)	**Depression Diagosis**	**1,339**	**21.55**	**60.91**

Table 7: Comparison with Related Datasets

4 Comparison with Related Datasets

Related datasets are introduced and compared with the proposed diagnosis dataset, see Table 7. Depression diagnosis requires precise symptom information collection via more dialogue turns, longer utterances, and sufficient emotional support.

Task-Oriented Dialogue Datasets Task-oriented dialogue dataset is one of the most essential components in dialogue systems study (Ni et al., 2021), consisting of various datasets for this purpose, i.e. MultiWOZ (Budzianowski et al., 2018), MSR-E2E (Li et al., 2018), CamRest (Wen et al., 2016) , Frames (Asri et al., 2017). However, these dialogue datasets aim at common scenarios in life. Thus, the number of dialogue turns is small. Moreover, little attention is paid to the user's emotions and empathize or comfort them in the dialogue.

Psychological Counseling Datasets Some dialogue studies related to mental health have addressed the emotions in the dialogue process and try to motivate users. For example, Saha et al. (2021) presents the dialogue dataset MotiVAte of imparting optimism, hope, and motivation for distressed people. Recently, works like ESConv (Liu et al., 2021) start to pay attention to Emotional Support Dialog Systems. However, they are mainly concerned with providing encouragement and advice to patients without giving professional diagnostic advice.

Medical Dialogue Datasets There are some medical dialogue datasets aiming at diagnosis before such as MedDG (Zeng et al., 2020) and MedDialog (Liu et al., 2020). However, these efforts focus mainly on somatic symptoms and physical diseases. MedDialog, although it has a small amount of psychiatric data, lacks professional annotation and cannot be used for a depression diagnosis dialogue system. Furthermore, the diagnosis process of depression largely differs from that of somatic disorders. According to ICD-11 (The World Health Organization, 2021b), in addition to somatic symptoms, patients often have multiple dimensions of symptoms such as mood, interest, mental status, and social function disorder. For this reason, psychiatrists need comprehensive information to provide unbiased diagnosis, so the dialogue will be longer and includes multiple knowledge domains.

4.1 Depression-Related Dialogue Dataset

Some dialogue datasets are strongly related to depression, such as DAIC-WOZ (Gratch et al., 2014), a multi-modal dataset. The dataset consists of face-to-face counseling conversations between the interviewer and patient suffering from depression, anxiety, etc, which researchers use to diagnose depression. However, there are only 189 dialogues, which is not enough for the dialogue generation task.

5 Experiments

5.1 Tasks

Based on the dataset D^4, we propose 4 tasks: ***generation*** (response and topic), ***summarization***, and ***classification***.

Response Generation The response generation task aims at generating the probable response of doctors based on the dialog context.

Topic Prediction This task predicts the topic of the response based on the dialogue context. In our experiments, we jointly optimize the model of topic prediction and the model of response generation. We take the topic as a special first token of dialogue response.

Dialogue Summary The dialogue summary task generates symptom summaries based on the entire dialog history.

Severity Classification The classification task separately predicts the severity of depression episodes and the suicide risk based on the dialogue context. Binary (positive/negative) and fine-grained 4-class (positive further classed into mild, medium and severe) classification are both investigated.

5.2 Backbone Models

Transformer We use the classic sequence-to-sequence model (Vaswani et al., 2017) to conduct the response generation and topic prediction experiment. The implementation used is HuggingFace[1]. The parameters are loaded from the transformer pretrained on MedDialog (Zeng et al., 2020), a Chinese Medical Dialogue Dataset.

BART BART (Lewis et al., 2019) is a denoising sequence-to-sequence pre-trained model which is a start-of-art model on both text generation and summary task. For this reason, we use Bart pertrained on Chinese datasets (Shao et al., 2021) to conduct the response generation and dialog summary task.

CPT CPT (Shao et al., 2021) is a novel Chinese pre-trained un-balanced transformer model, which is not only effective in generation task, but also has powerful classification ability, so we choose it as our backbone model to conduct the generation task and also compare its performance of classification task with BART.

BERT Bert (Devlin et al., 2019) is effectively used for a wide range of language understanding tasks, such as question answering and language inference. Thus, we use the version[2] which is pretranied on Chinese datasets (Cui et al., 2020) to conduct the classification task.

5.3 Objective Evaluation

Generation and Summarization We evaluate the *response generation task* and *dialog summary task* with objective metrics including BLEU-2 (Papineni et al., 2002), Rouge-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) to measure the similarity between model generated responses and labels. And to show the diversity of generation, we compute DIST-2 (Li et al., 2015). We implement jieba[3] for tokenization and compute the metrics at word-level.

Results for the *response generation task* are presented in Table 8. Three observations can be drawn:

[1] https://github.com/huggingface/transformers

[2] https://huggingface.co/hfl/chinese-bert-wwm

[3] https://github.com/fxsjy/jieba

Model	BLEU-2	ROUGE-L	METEOR	DIST-2	Topic ACC.
Transformer	3.76%	0.14	0.0964	0.34	23.79%
BART	11.02%	0.24	0.1870	0.06	41.94%
CPT	11.52%	0.25	0.1893	0.05	41.72%
BART*	14.15%	0.29	0.2242	0.09	-

Table 8: Evaluation Results of Response Generation and Topic Prediction: models denoted with * means golden topics are given as a part of the input.

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

Table 9: Evaluation Results of Dialog Summary Task

Task	Model	Precision	Recall	F1
Depression 2-class	BERT	0.82	0.80	0.78
	BART	0.80	0.80	0.80
	CPT	0.80	0.78	0.78
Depression 4-class	BERT	0.53	0.50	0.47
	BART	0.57	0.47	0.48
	CPT	0.52	0.50	0.50
Suicide 2-class	BERT	0.86	0.84	0.84
	BART	0.76	0.76	0.76
	CPT	0.84	0.81	0.82
Suicide 4-class	BERT	0.73	0.64	0.67
	BART	0.60	0.60	0.59
	CPT	0.76	0.70	0.71

Table 10: Evaluation Results of Severity Classification

Metric	Ground-Truth	CPT	BART
Fluency	2.72	2.40	2.36
Reasonableness	2.66	1.46	1.61
Doctor-likeness	2.58	1.82	1.62
Comfort	2.46	1.53	1.53

Table 11: Human Evaluation Results of Response Generation: the range of the 4 metrics is [0, 3].

1) BART and CPT exhibit similar generation performance on our dataset; 2) Both models largely outperforms Transformer, which is pretrained on medical corpus, suggesting that depression diagnosis are different from traditional somatic-oriented medical dialogues; 3) Given golden topics (BART*), generation performance can be further enhanced.

Topic Prediction accuracy results are shown as Topic ACC. in Table 8. Similar trend is observed: BART $\approx$ CPT $>$ Transformer. Since the ten topics are categorized based on core symptoms and emotional support, the uncertainty and linguistic ambiguity of the depression diagnosis dialogue has undoubtedly increased the prediction difficulty.

Results for *Dialog Summary* are listed in Table 9, CPT is on par with BART in terms of the N-gram overlap with human references. Nevertheless, CPT exhibits higher DIST-2 score, suggesting its superiority on generation diversity.

Severity Classification Binary and 4-class classification are evaluated by average weighted precision, recall and F1 by skrlearn[4], shown in Table 10. Results of 4-classification tasks are relatively poor compared with the performance in 2-classification tasks, indicating that the fine-grained classification of depression severity is still challenging for current models.

5.4 Human Evaluation

In order to better evaluate the performance of responses generated by the model, we employ 15 workers to rate the responses of ground truth, generated by the BART and CPT separately. We randomly selected 100 responses from different topic for each model and let 3 workers evaluated the same response from 4 aspects: **Fluency** measures fluency of generated sentences; **Reasonableness** measures how reasonable to give this response based on the dialog history; **Doctor-likeness** measures what extent does the response resemble the words of a doctor; **Comfort** measures how comforting the response is. The evaluation result is in Table 11. Generally, human evaluation is in accordance with objective measures: CPT and BART exhibit similar performances, though both fall behind Ground-Truth. With regard to detailed human evaluation metrics, both models can generate fluent responses. However, towards generating reasonable, comforting and doctor-like responses, both models are still facing great challenges in conducting a professional diagnosis.

6 Conclusion

In this paper, we construct an up-to-clinical-standard depression diagnosis dataset with 1,339 conversations accompanied by psychiatrists' diagnosis summaries. Further, we conduct experimental validation on multiple tasks with state-of-art models and compare the results with objective and human evaluation. Although models could generate fluent and human-like response, diagnosis dialogue generation remains a challenging task.

References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*,

[4] https://scikit-learn.org

5th ed. edition. Autor, Washington, DC.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

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.

Aaron T Beck, Robert A Steer, and Gregory K Brown. 1996. *Beck depression inventory (BDI-II)*, volume 10. Pearson.

Dror Ben-Zeev, Michael A Young, and Patrick W Corrigan. 2010. Dsm-v and the stigma of mental illness. *Journal of mental health*, 19(4):318–327.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. 2013. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 11–20.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Gratch, Ron Artstein, Gale 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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128.

Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: a new depression diagnostic and severity measure.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *arXiv preprint arXiv:2010.07497*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387*.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tulika Saha, Saraansh Chopra, Sriparna Saha, Pushpak Bhattacharyya, and Pankaj Kumar. 2021. A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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. *arXiv preprint arXiv:2109.05729*.

The World Health Organization. 2021a. Depression.

The World Health Organization. 2021b. Depression.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. Conditional generation and snapshot learning in neural dialogue systems. *arXiv preprint arXiv:1606.03352*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

A Ethical Considerations

When we collected patients' portraits, we had informed the patients of the purpose of the data and obtained their consent to use the anonymous data for research. When collecting dialogue data, we also informed all participants of the purpose. Our job is not to explore how machines can replace human doctors but to use dialogue systems to alleviate the lack of medical resources.

B Training Details

B.1 Response Generation

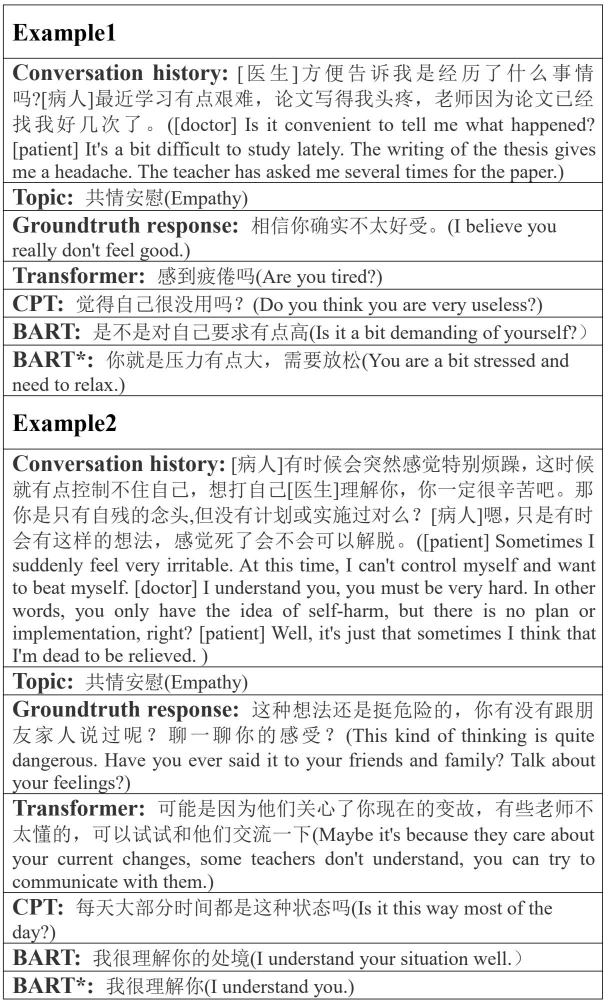
For BART and CPT models, the initial parameters are pretrained on Chinese datasets (Shao et al., 2021). We use a cosine learning rate scheduler with the initial learning rate of 1e-5, 100 warm-up steps and the AdamW optimizer (Loshchilov and Hutter, 2019). Beam search where the number of beams is 4 is used in response generation. Models are trained for 30 epochs, the one with the best BLEU-2 metric on the evaluation set is selected for test.

For the Transformer, we use the implementation by HuggingFace[5]. We load the parameters of the Transformer pretrained on MedDialog (Zeng et al., 2020). The weight parameters were learned with Adam and a linear learning rate scheduler with the initial learning rate of 1.0e-4 and 100 warm-up steps. The batch size was set to 16. Top-k random sampling (Fan et al., 2018) is used in response generation. The model is trained for 20 epochs. The one with highest BLEU-2 score on evaluation set is chosen for test.

Due to the limitation of models' positional embedding, we intercept data with a length over 512. In the response generation task, we try to keep the most recent conversations as they are more instructive to the current response. To further minimize the context size, we replace the utterences of the doctor with the doctor's topic.

B.2 Dialog Summary

Both BART and CPT models are trained for 50 epochs. We use a cosine learning rate scheduler with the initial learning rate of 1e-5 and 100 warm-up steps and the AdamW optimizer. The one with the highest rouge-1 metric on the evaluation set is selected for test.



Example1
Conversation history: [医生]方便告诉我是经历了什么事情吗?[病人]最近学习有点艰难，论文写得我头疼，老师因为论文已经找我好几次了。([doctor] Is it convenient to tell me what happened? [patient] It's a bit difficult to study lately. The writing of the thesis gives me a headache. The teacher has asked me several times for the paper.)
Topic: 共情安慰(Empathy)
Groundtruth response: 相信你确实不太好受。(I believe you really don't feel good.)
Transformer: 感到疲倦吗(Are you tired?)
CPT: 觉得自己很没用吗？(Do you think you are very useless?)
BART: 是不是对自己要求有点高(Is it a bit demanding of yourself?）
BART*: 你就是压力有点大，需要放松(You are a bit stressed and need to relax.)
Example2
Conversation history: [病人]有时候会突然感觉特别烦躁，这时候就有点控制不住自己，想打自己[医生]理解你，你一定很辛苦吧。那你是只有自残的念头,但没有计划或实施过对么？[病人]嗯，只是有时会有这样的想法，感觉死了会不会可以解脱。([patient] Sometimes I suddenly feel very irritable. At this time, I can't control myself and want to beat myself. [doctor] I understand you, you must be very hard. In other words, you only have the idea of self-harm, but there is no plan or implementation, right? [patient] Well, it's just that sometimes I think that I'm dead to be relieved.)
Topic: 共情安慰(Empathy)
Groundtruth response: 这种想法还是挺危险的，你有没有跟朋友家人说过呢？聊一聊你的感受？(This kind of thinking is quite dangerous. Have you ever said it to your friends and family? Talk about your feelings?)
Transformer: 可能是因为他们关心了你现在的变故，有些老师不太懂的，可以试试和他们交流一下(Maybe it's because they care about your current changes, some teachers don't understand, you can try to communicate with them.)
CPT: 每天大部分时间都是这种状态吗(Is it this way most of the day?)
BART: 我很理解你的处境(I understand your situation well.）
BART*: 我很理解你(I understand you.)

Figure 5: Examples of generated response

If the input dialog history is longer than the model's input size, we retain the utterences in the middle of the dialog.

C Examples of generated response

As shown in Figure 5, we selected two representative examples of the generated responses by different models. Their responses are fluent and precisely predict the topic of the response according to the context.

D Examples of generated summary

In Figure 6, we present an example of the generated summary by different models. In spite of divergence in diagnosis results, the models all list the main symptoms of the patient.

[5] https://github.com/huggingface/transformers

Groundtruth Diagnosis Sheet: 来访者注意力下降；精神运动性激越和迟滞，有躯体反应：头晕；自我价值感低；睡眠质量不好，睡眠浅；没有抑郁风险和自杀风险。(The visitor’s attention drop; psychomotor agitation and retardation, physical reactions: dizziness; low sense of self-worth; poor sleep quality, shallow sleep; no risk of depression and suicide.)

BART: 病人最近情绪低落，上课集中不了注意力，学习效率下降，睡眠障碍，入睡困难，躯体不适感，思维迟缓，诊断抑郁发作。(The patient has recently been depressed, unable to concentrate in class, decreased learning efficiency, sleep disturbance, difficulty falling asleep, physical discomfort, slow thinking, and diagnosis of depressive episodes.)

CPT: 来访者近一个月出现：注意力下降，自我评价降低，认知功能受损，兴趣减退，精力不足，睡眠障碍，躯体化症状，头晕，考虑抑郁发作。(In the past month, the visitor experienced: decreased concentration, decreased self-evaluation, impaired cognitive function, decreased interest, lack of energy, sleep disturbance, somatization symptoms, dizziness, depressive episodes considered.)

Figure 6: Examples of generated summary

E Samples of topics

In Table 7, we present the 10 topics with their typical examples and hot words.

F Explanation of symptoms

Figure 8 and Figure 9 are the detailed explanations of depression symptoms. Both patients and doctors are asked to get acquaintance with the two figures before starting their conversations.

G Doctors’ questions

Figure 10 and Figure 11 are the diagnosis tutorial for doctors, which helps them adapt to this role faster.

H Screenshot of the data collection system

Figure 12 is doctors’ user interface, and Figure 13 is the patients’.

Topic	Example	Hot words

Sleep	那你一般需要多久能睡着啊？
	How long do you usually need to fall asleep?
		入睡困难(Difficulty falling-asleep)
		早醒(Wake up early)
		睡着(Falling asleep)

Sentiment	你觉得有影响到你的情绪吗？
	Do you think it affects your mood?
		快乐(Happiness)
		心情(Sentiment)
		低落(Upset)

Screening	你会不会有时候觉得比较兴奋？
	Do you get excited sometimes?
		家族史(Family history)
		有没有(Do you have)
		亲属中有患者吗(Are there patients among the relatives)

Interest	你会觉得对过去的爱好失去兴趣吗？
	Do you feel uninterested in the past hobbies?
		兴趣Interest)
		喜欢(Like)
		爱好(Hobby)

Mental State	会感到每天很疲劳或者精力不足吗？
	Do you feel tired or under-energized every day?
		自信(Self-confidence)
		疲劳(Tired)
		决断(Judge)
		注意力溃散(Broken attention)

Social Function	会和朋友们倾诉自己的问题吗？
	Will you talk to your friends about your problems?
		学习(Study)
		工作(Work)
		生活(Life)
		社交(Social)
		朋友(Friends)

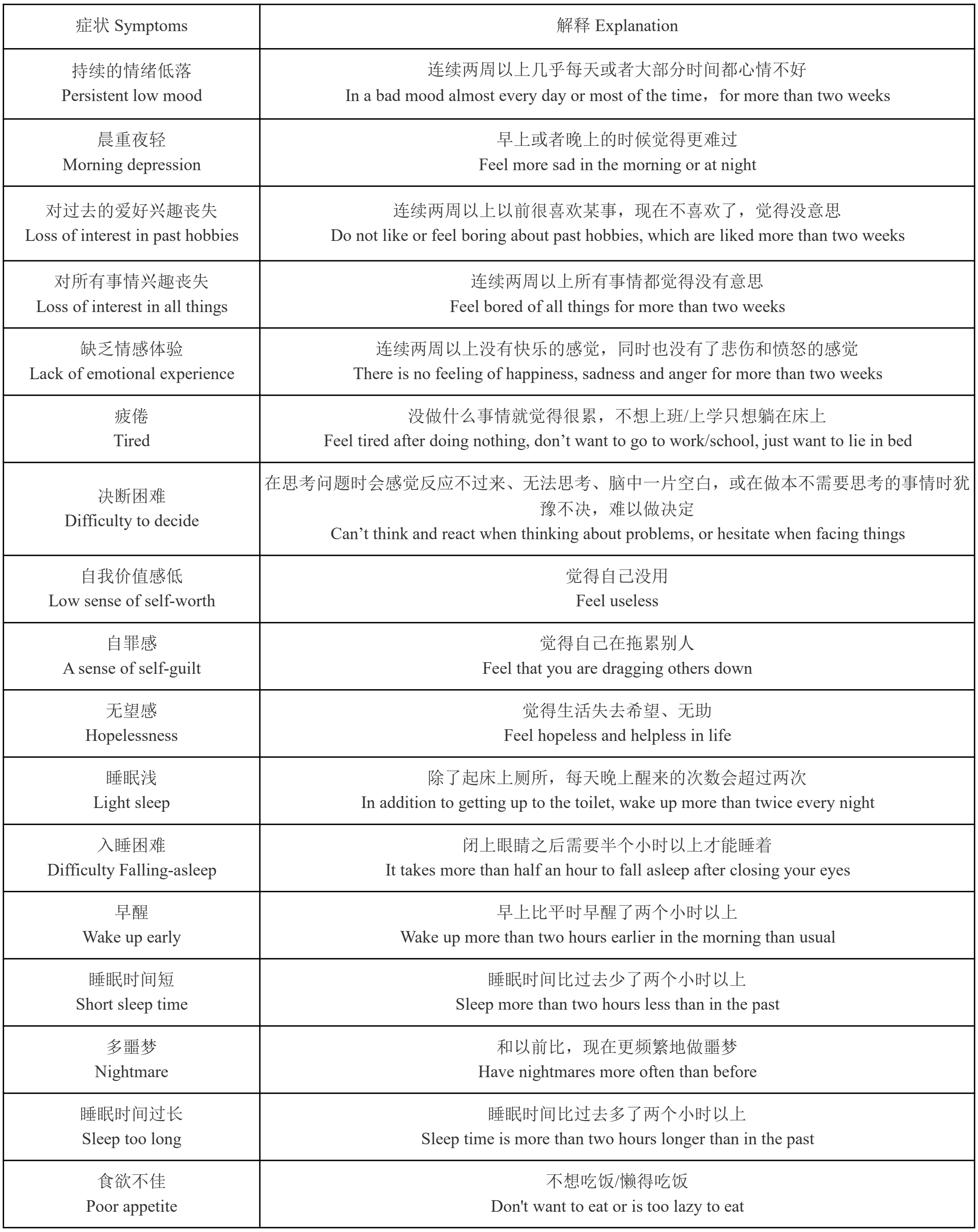
Appetite	那体重跟食欲方面最近有什么变化吗？
	Has there been any recent change in weight and appetite?
		胃口(Appetite)
		食欲(Appetite)
		吃饭(Dine)
		体重(Weight)

Suicide	在你感到绝望的时候有想过伤害自己吗？
	Have you ever wanted to hurt yourself when you're desperate?
		绝望(Despair)
		自杀(Suicide)
		无望感(Hopelessness)
		消极(Negative)
		自责(Self-blame)
		拖累(Encumber)
		悲观(Gloomy)

Empathy	嗯嗯，选择困难症很多人都有哦，不用太烦恼。
	Well, a lot of people have a choice of difficulties, don't worry too much.
		理解(Understand)
		加油(Come on)
		不用担心(Don’t worry)
		会好起来(Will get better)

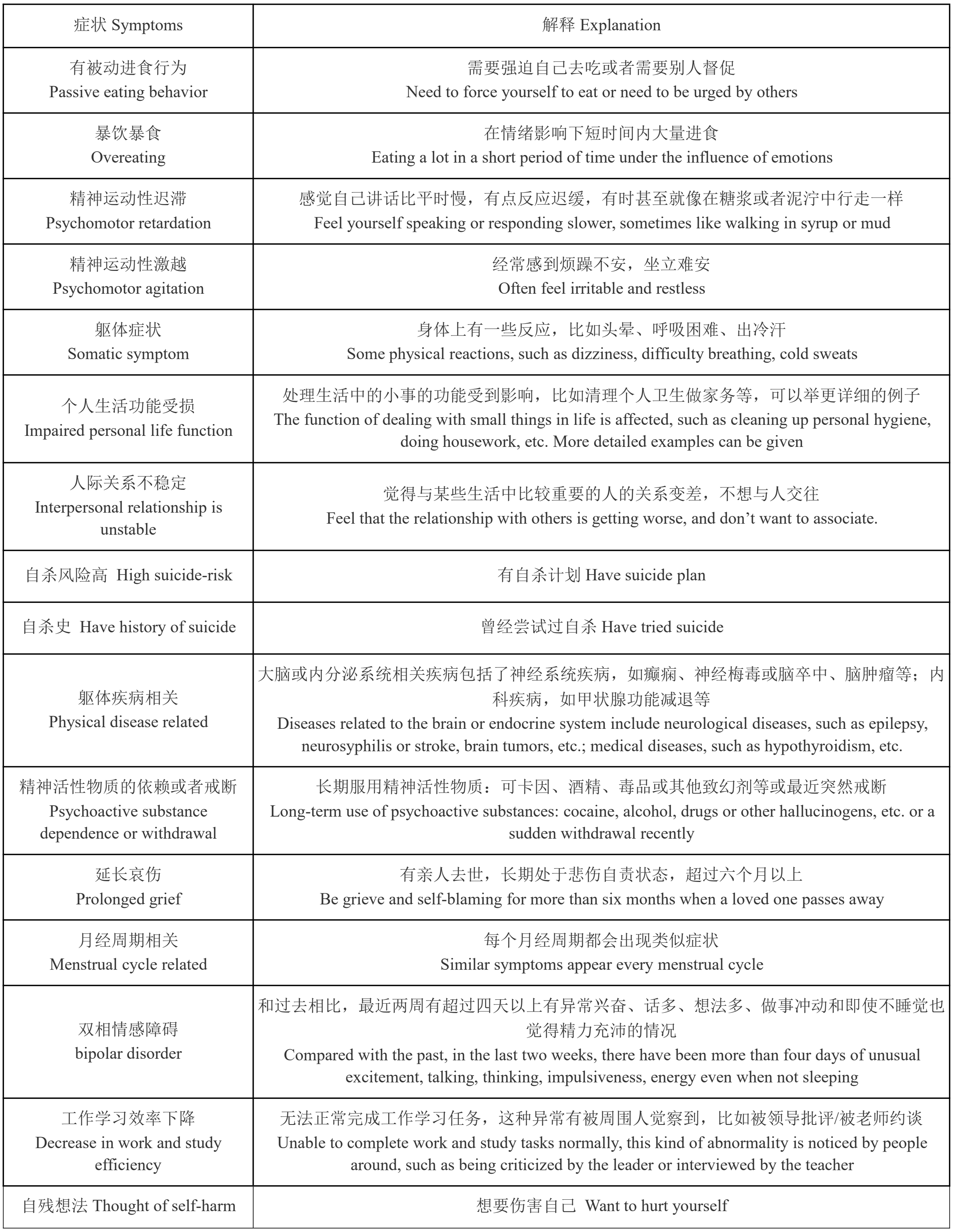
Somatic Symptom	你会觉得头晕冒冷汗什么的吗？
	Do you feel dizzy and sweating or something?
		身体(Body)
		躯体(Body)
		头晕(Dizzy)
		暴躁(Irascible)
		冒冷汗(Sweat)

Figure 7: Samples of doctors’ topic



症状 Symptoms	解释 Explanation
持续的情绪低落 Persistent low mood	连续两周以上几乎每天或者大部分时间都心情不好 In a bad mood almost every day or most of the time，for more than two weeks
晨重夜轻 Morning depression	早上或者晚上的时候觉得更难过 Feel more sad in the morning or at night
对过去的爱好兴趣丧失 Loss of interest in past hobbies	连续两周以上以前很喜欢某事，现在不喜欢了，觉得没意思 Do not like or feel boring about past hobbies, which are liked more than two weeks
对所有事情兴趣丧失 Loss of interest in all things	连续两周以上所有事情都觉得没有意思 Feel bored of all things for more than two weeks
缺乏情感体验 Lack of emotional experience	连续两周以上没有快乐的感觉，同时也没有了悲伤和愤怒的感觉 There is no feeling of happiness, sadness and anger for more than two weeks
疲倦 Tired	没做什么事情就觉得很累，不想上班/上学只想躺在床上 Feel tired after doing nothing, don’t want to go to work/school, just want to lie in bed
决断困难 Difficulty to decide	在思考问题时会感觉反应不过来、无法思考、脑中一片空白，或在做本不需要思考的事情时犹豫不决，难以做决定 Can’t think and react when thinking about problems, or hesitate when facing things
自我价值感低 Low sense of self-worth	觉得自己没用 Feel useless
自罪感 A sense of self-guilt	觉得自己在拖累别人 Feel that you are dragging others down
无望感 Hopelessness	觉得生活失去希望、无助 Feel hopeless and helpless in life
睡眠浅 Light sleep	除了起床上厕所，每天晚上醒来的次数会超过两次 In addition to getting up to the toilet, wake up more than twice every night
入睡困难 Difficulty Falling-asleep	闭上眼睛之后需要半个小时以上才能睡着 It takes more than half an hour to fall asleep after closing your eyes
早醒 Wake up early	早上比平时早醒了两个小时以上 Wake up more than two hours earlier in the morning than usual
睡眠时间短 Short sleep time	睡眠时间比过去少了两个小时以上 Sleep more than two hours less than in the past
多噩梦 Nightmare	和以前比，现在更频繁地做噩梦 Have nightmares more often than before
睡眠时间过长 Sleep too long	睡眠时间比过去多了两个小时以上 Sleep time is more than two hours longer than in the past
食欲不佳 Poor appetite	不想吃饭/懒得吃饭 Don't want to eat or is too lazy to eat

Figure 8: Explanation of Symptoms - 1



症状 Symptoms	解释 Explanation
有被动进食行为 Passive eating behavior	需要强迫自己去吃或者需要别人督促 Need to force yourself to eat or need to be urged by others
暴饮暴食 Overeating	在情绪影响下短时间内大量进食 Eating a lot in a short period of time under the influence of emotions
精神运动性迟滞 Psychomotor retardation	感觉自己讲话比平时慢，有点反应迟缓，有时甚至就像在糖浆或者泥泞中行走一样 Feel yourself speaking or responding slower, sometimes like walking in syrup or mud
精神运动性激越 Psychomotor agitation	经常感到烦躁不安，坐立难安 Often feel irritable and restless
躯体症状 Somatic symptom	身体上有一些反应，比如头晕、呼吸困难、出冷汗 Some physical reactions, such as dizziness, difficulty breathing, cold sweats
个人生活功能受损 Impaired personal life function	处理生活中的小事的功能受到影响，比如清理个人卫生做家务等，可以举更详细的例子 The function of dealing with small things in life is affected, such as cleaning up personal hygiene, doing housework, etc. More detailed examples can be given
人际关系不稳定 Interpersonal relationship is unstable	觉得与某些生活中比较重要的人的关系变差，不想与人交往 Feel that the relationship with others is getting worse, and don’t want to associate.
自杀风险高 High suicide-risk	有自杀计划 Have suicide plan
自杀史 Have history of suicide	曾经尝试过自杀 Have tried suicide
躯体疾病相关 Physical disease related	大脑或内分泌系统相关疾病包括了神经系统疾病，如癫痫、神经梅毒或脑卒中、脑肿瘤等；内科疾病，如甲状腺功能减退等 Diseases related to the brain or endocrine system include neurological diseases, such as epilepsy, neurosyphilis or stroke, brain tumors, etc.; medical diseases, such as hypothyroidism, etc.
精神活性物质的依赖或者戒断 Psychoactive substance dependence or withdrawal	长期服用精神活性物质：可卡因、酒精、毒品或其他致幻剂等或最近突然戒断 Long-term use of psychoactive substances: cocaine, alcohol, drugs or other hallucinogens, etc. or a sudden withdrawal recently
延长哀伤 Prolonged grief	有亲人去世，长期处于悲伤自责状态，超过六个月以上 Be grieve and self-blaming for more than six months when a loved one passes away
月经周期相关 Menstrual cycle related	每个月经周期都会出现类似症状 Similar symptoms appear every menstrual cycle
双相情感障碍 bipolar disorder	和过去相比，最近两周有超过四天以上有异常兴奋、话多、想法多、做事冲动和即使不睡觉也觉得精力充沛的情况 Compared with the past, in the last two weeks, there have been more than four days of unusual excitement, talking, thinking, impulsiveness, energy even when not sleeping
工作学习效率下降 Decrease in work and study efficiency	无法正常完成工作学习任务，这种异常有被周围人觉察到，比如被领导批评/被老师约谈 Unable to complete work and study tasks normally, this kind of abnormality is noticed by people around, such as being criticized by the leader or interviewed by the teacher
自残想法 Thought of self-harm	想要伤害自己 Want to hurt yourself

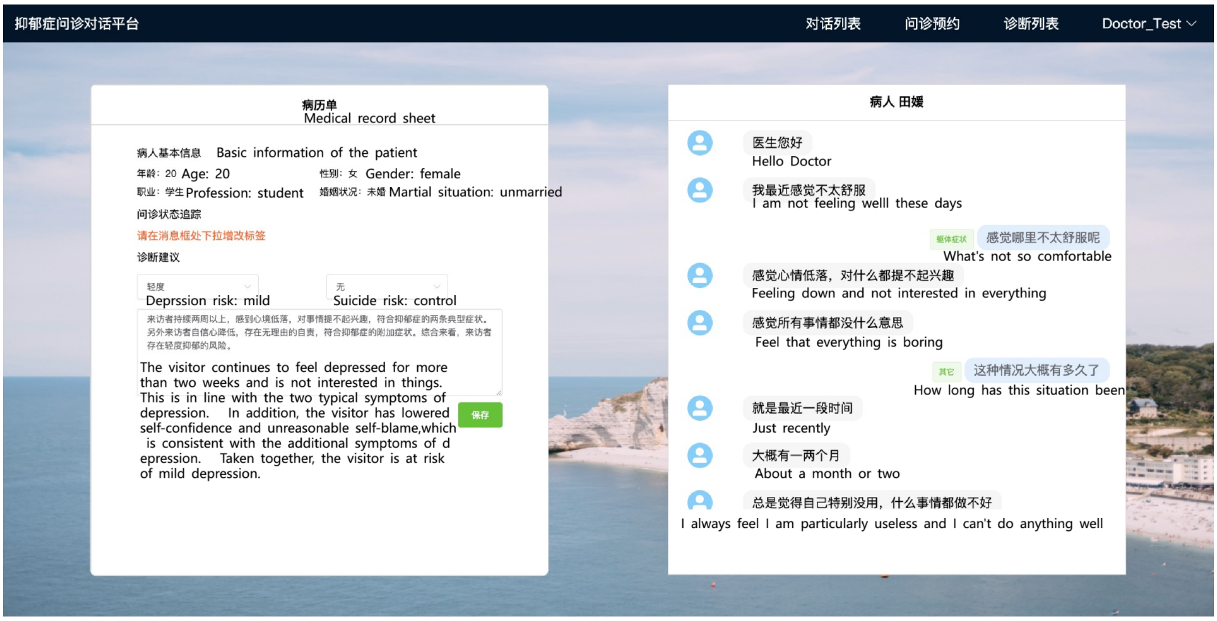
Figure 9: Explanation of Symptoms - 2

症状版块 Symptoms section	询问主题 Consultation topic	备注 Remark
导语 Lead	病人主要诉求 Patient's main appeal	
持续时间 Duration	持续时间 Duration	病人有情绪低落/兴趣低下/疲倦的问题之后提问 Ask the question after the patient has problems with depression/low interest/tiredness
原因 Cause	病因 Cause	病人有情绪低落或兴趣低下问题时提问 Ask if the patient has a problem with depression or low interest
情绪低落 Upset	是否有情绪低落 Whether patients are upset	
	持续时间 Duration	
	早晚差异 The difference between morning and evening	是否在某些特定时段会尤为心情不好 Are you in a particularly bad mood at certain times
兴趣低下 Low interest	是否兴趣低下 Does the patient has low interest	
	不感兴趣的范围 Range for not being interested	
	不感兴趣的原因 Reasons for not being interested	
	是否情感淡漠 Is it emotionally indifferent	
社会功能 Social function	个人生活事务 Personal life affairs	根据不同年龄段提问一些基本的生活事务是否正常 According to different age groups, ask whether some basic life affairs are normal
	学习工作 Study and Work	
	社交 Social contact	是否和家人朋友联系/倾诉，是否获得他们的支持 Whether to contact/talk to family and friends, to get their support
	社交 Social contact	病人是否有意回避社交 Does the patient deliberately avoid social interaction
精神状态 Mental state	注意力下降 Decreased concentration	
	记忆力变差 Memory loss	
	疲倦 Tired	
	决断困难 Difficulty in decision	
	自信心下降 Decline in self-confidence	

Figure 10: Doctors’ questions - 1

症状版块 Symptoms section	询问主题 Consultation topic	备注 Remark
睡眠问题 Sleep problems	睡眠问题 Does the patient has sleep problems	
	入睡困难 Difficulty falling asleep	有睡眠问题逐个问 Ask if the patient has sleep problems
	睡眠浅 Light sleep	有睡眠问题逐个问 Ask if the patient has sleep problems
	早醒 Wake up early	有睡眠问题逐个问 Ask if the patient has sleep problems
	睡眠时间过短 Sleep too short	有睡眠问题逐个问 Ask if the patient has sleep problems
	多梦 Dreamy	有睡眠问题逐个问 Ask if the patient has sleep problems
食欲问题 Appetite problems	食欲问题 Does the patient has appetite problems	
	食欲不振 Loss of appetite	
	暴饮暴食 Overeating	
	体重变化 Weight change	无上述食欲问题时提问 Ask when there is no appetite problem mentioned above
躯体症状 （有严重情绪和兴趣问题时再问） Somatic symptom (Ask when patients have serious emotional and interest issues)	精神运动性激越或迟滞 Psychomotor agitation or retardation	烦躁不安或反应迟缓 Irritability or slow response
	躯体不适 Physical discomfort	
自杀 Suicide	自残倾向 Self-harm tendency	
	自杀倾向 Suicidal tendency	
	无望感 Hopelessness	
	未来的规划 Future plan	
	内疚感/自卑感 Guilt/inferiority complex	
	自我价值感低 Low self-worth	
筛查 Screening	亲人去世导致长期悲伤 The death of a loved one causes long-term grief	病人描述中提到时需要问 Need to ask when mentioned in the patient description
	躁狂 Mania	是否易怒、易发生争执 Is it irritable and prone to disputes
遗传史 Genetic history	遗传 Genetic	如果对方有情绪兴趣症状或者自杀倾向 If the patient has emotional or interest symptoms or suicidal tendencies
结束之前 Before the end	病人是否有其他问题 Does the patient have other problems	

Figure 11: Doctors' questions - 2



病历单
Medical record sheet

病人基本信息 Basic information of the patient

年龄：20 Age: 20　性别：女 Gender: female

职业：学生 Profession: student　婚姻状况：未婚 Martial situation: unmarried

问诊状态追踪

请在消息框处下拉增改标签

诊断建议

轻度 Deprssion risk: mild　无 Suicide risk: control

来访者持续两周以上，感到心境低落，对事情提不起兴趣，符合抑郁症的两条典型症状。另外来访者自信心降低，存在无理由的自责，符合抑郁症的附加症状。综合来看，来访者存在轻度抑郁的风险。

The visitor continues to feel depressed for more than two weeks and is not interested in things. This is in line with the two typical symptoms of depression. In addition, the visitor has lowered self-confidence and unreasonable self-blame,which is consistent with the additional symptoms of d epression. Taken together, the visitor is at risk of mild depression.

病人 田媛

医生您好
Hello Doctor

我最近感觉不太舒服
I am not feeling welll these days

躯体症状 感觉哪里不太舒服呢
What's not so comfortable

感觉心情低落，对什么都提不起兴趣
Feeling down and not interested in everything

感觉所有事情都没什么意思
Feel that everything is boring

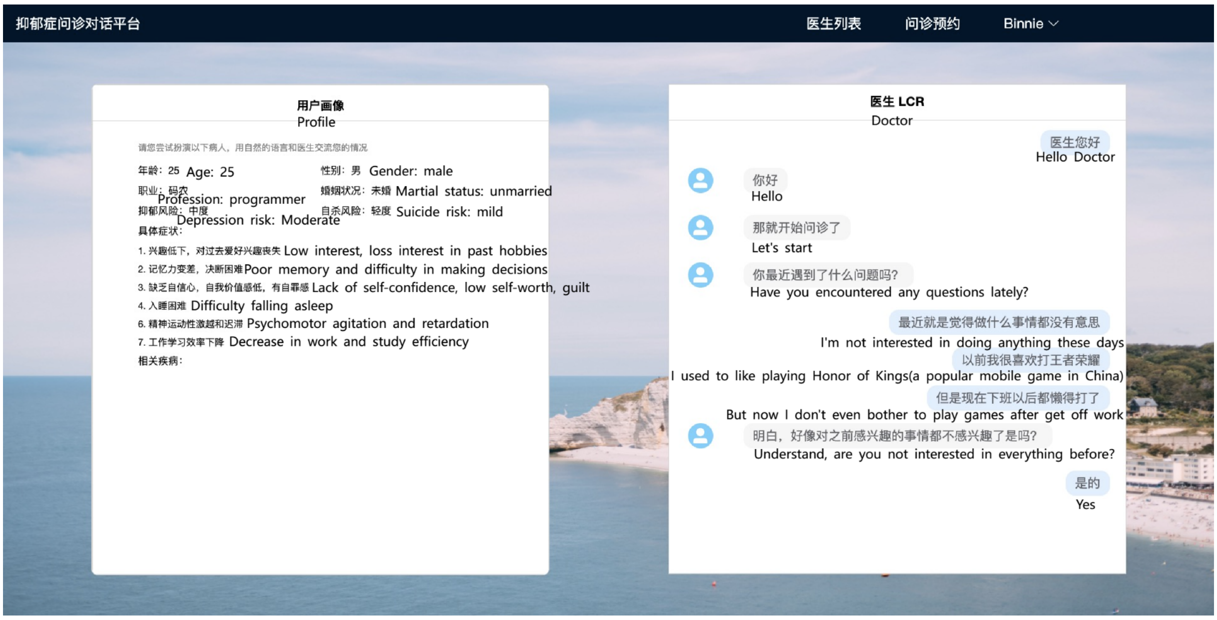
其它 这种情况大概有多久了
How long has this situation been

就是最近一段时间
Just recently

大概有一两个月
About a month or two

总是觉得自己特别没用，什么事情都做不好
I always feel I am particularly useless and I can't do anything well

Figure 12: Page of doctor



用户画像
Profile

请您尝试扮演以下病人，用自然的语言和医生交流您的情况

年龄：25 Age: 25　性别：男 Gender: male

职业：码农 Profession: programmer　婚姻状况：未婚 Martial status: unmarried

抑郁风险：中度 Depression risk: Moderate　自杀风险：轻度 Suicide risk: mild

具体症状：

1. 兴趣低下，对过去爱好兴趣丧失 Low interest, loss interest in past hobbies
2. 记忆力变差，决断困难 Poor memory and difficulty in making decisions
3. 缺乏自信心，自我价值感低，有自罪感 Lack of self-confidence, low self-worth, guilt
4. 入睡困难 Difficulty falling asleep
6. 精神运动性激越和迟滞 Psychomotor agitation and retardation
7. 工作学习效率下降 Decrease in work and study efficiency

相关疾病：

医生 LCR
Doctor

医生您好
Hello Doctor

你好
Hello

那就开始问诊了
Let's start

你最近遇到了什么问题吗？
Have you encountered any questions lately?

最近就是觉得做什么事情都没有意思
I'm not interested in doing anything these days

以前我很喜欢打王者荣耀
I used to like playing Honor of Kings(a popular mobile game in China)

但是现在下班以后都懒得打了
But now I don't even bother to play games after get off work

明白，好像对之前感兴趣的事情都不感兴趣了是吗？
Understand, are you not interested in everything before?

是的
Yes

Figure 13: Page of patient