D^4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat

Binwei Yao[1,2,3], **Chao Shi**[3], **Likai Zou**[3], **Lingfeng Dai**[1,2,3]
Mengyue Wu[1,2,3*], **Lu Chen**[1,2,3*], **Zhen Wang**[4,5] and **Kai Yu**[1,2,3]
[1]SJTU X-LANCE Lab, Department of Computer Science and Engineering
[2]MoE Key Lab of Artificial Intelligence, SJTU AI Institute
[3]Shanghai Jiao Tong University, Shanghai, China
[4]Shanghai Mental Health Center
[5]Shanghai Jiao Tong University School of Medicine, Shanghai, China
{yaobinwei, mengyuewu, chenlusz, kai.yu}@sjtu.edu.cn

Abstract

In a depression-diagnosis-directed clinical session, doctors initiate a conversation with ample emotional support that guides the patients to expose their symptoms based on clinical diagnosis criteria. Such a dialogue system is distinguished from existing single-purpose human-machine dialog systems, as it combines task-oriented and chit-chats with uniqueness in dialogue topics and procedures. However, due to the social stigma associated with mental illness, the dialogue data related to depression consultation and diagnosis are rarely disclosed. Based on clinical depression diagnostic criteria ICD-11 and DSM-5, we designed a 3-phase procedure to construct D^4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat[1], which simulates the dialogue between doctors and patients during the diagnosis of depression, including diagnosis results and symptom summary given by professional psychiatrists for each conversation. Upon the newly-constructed dataset, four tasks mirroring the depression diagnosis process are established: response generation, topic prediction, dialog summary, and severity classification of depressive episode and suicide risk. Multi-scale evaluation results demonstrate that a more empathy-driven and diagnostic-accurate consultation dialogue system trained on our dataset can be achieved compared to rule-based bots.

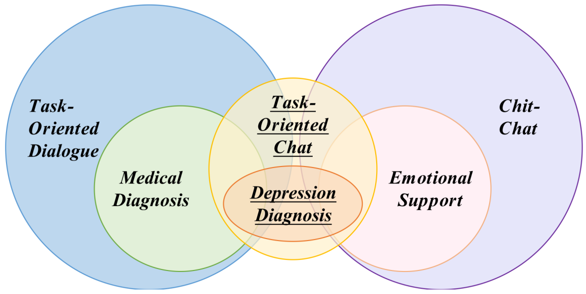

Figure 1: Comparison of Different Dialogue Types

1 Introduction

Given the increasing worldwide health threat brought by depression, researchers have been exploring effective methods for depression detection and diagnosis. Besides automatic depression detection from posts on social media (Orabi et al., 2018), speech (Zhang et al., 2021b) and multi-modality (Cummins et al., 2013), the dialogue system is considered an effective tool for large-scale depression detection (Pacheco-Lorenzo et al., 2021). It is believed that conversation agents could reduce the concealment of sensitive information such as suicidal thoughts caused by social expectations (Schuetzler et al., 2018) and the emotional hindrance due to the pressure of being judged in face-to-face conversation (Hart et al., 2017). In past research, chatbots initiated for depression diagnosis are generally implemented based on self-rating scales (Jaiswal et al., 2019; Arrabales, 2020) or diagnostic criteria (Philip et al., 2017). The final diagnosis results are obtained by asking fixed questions on the scale and corresponding the user's answers to each question to the scale options. These chatbots present good sensitivity and specificity in diagnosis and are more attractive and acceptable (Vaidyam et al., 2019; Abd-Alrazaq et al., 2019) than the original self-rating scales. Nevertheless, the fixed dialogue flow limiting the user's expressions to specific answers can not realize personalized consultation and give emotional support at an appropriate time, for which there still exists a big gap between the conversation experience current depression diagnosis agents provide and the face-to-face interview in the process of clinical diagnosis.

Interview-based clinical diagnosis in psychiatry is a complex procedure with the purpose of collecting and summarizing key symptom infor-

*Mengyue Wu and Lu Chen are the corresponding authors.

[1]To get access to D^4, please look at https://x-lance.github.io/D4.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2438–2459

mation about one patient while providing a chat-like conversation experience. In clinical practice, psychiatrists communicate with patients and offer diagnosis results based on practical experience and multiple diagnostic criteria. The most clinically-adopted criteria involve ICD-11 (The World Health Organization, 2022), DSM-5 (American Psychiatric Association, 2013), etc., which define core symptoms for the depression diagnosis. At the same time, psychiatrists provide emotional support such as empathy and comfort during the consultation to better prompt patients' self-expression. The practice of clinical depression diagnosis displays the possibility of the depression diagnosis dialogue system in further improving the accuracy of diagnosis and user engagement.

Accordingly, the depression diagnostic conversation belongs to a distinguished dialogue from previously defined dialogue typologies, which is a combination of task-oriented dialogue and chit-chat. Such a compound dialogue type could be defined as *Task-Oriented Chat* as shown in Figure 1. This type of dialogue requires multiple assessments regarding task completion and chit-chat experience, which are extremely challenging and still under-investigated. As a specific domain of Task-Oriented Chat, the depression diagnosis dialogue has a clear purpose of the task-oriented dialogue aiming at medical diagnosis: to collect the patient's symptom information and draw a diagnosis conclusion while simultaneously bearing the needs of a chit-chat dialogue with emotional support: to start a user-oriented chat and provide emotional support from time to time. Currently, no datasets are specified for depression diagnosis, mainly due to the social stigma associated with clinical privacy and the complexity of the diagnosis process.

To construct a clinically sound and empathetic depression-diagnosis-oriented dialogue system close to clinical practice, we conduct dialogue collection through consultation dialogue simulation. We devise a three-phase approach to collect depression diagnostic dialogues (see Figure 2). **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 the consultation dialogue simulation based on the collected portraits. The workers are divided into patients and doctors for separate training by professionals. The doctor actor is required to obtain fixed symptom information involved in the diagnostic criteria in the chat, while the patient actor needs to express according to the symptoms in the portrait. **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. We further annotate the conversation procedure with 10 topic tags and the symptom summaries with 13 symptom tags (grouped by core depressive symptoms listed in DSM-5 and ICD-11). In this way, we propose $\mathbf{D}^4$: a Chinese **D**ialogue **D**ataset for **D**epression-**D**iagnosis-Oriented Chat. The key contribution of this paper is as follows:

- A close-to-clinical-practice 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, with multi-dimensional analysis suggesting that our simulated diagnosis data are reliable and up to professional standards.
- Experimental validation on four tasks that mirror the real-life diagnosis process: 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, aiming to advance the realization of an Avante-Garde clinical diagnosis-oriented dialogue system that combines characteristics of task-oriented dialogue and chit-chat.

2 Data Collection

To maximize doctor and patient authenticity in a diagnosis dialogue, we devise a 3-phase collection paradigm (see Figure2) instead of the commonly-adopted vanilla crowdsourcing scheme: **P1.** We collected natural populations' portraits (in particular actual depressive patients) to form pre-diagnosis records; **P2.** Simulated natural diagnostic consultation dialogues based on the portraits; **P3.** Psychiatrists proofread dialogue history and prescribed professional symptom summaries.

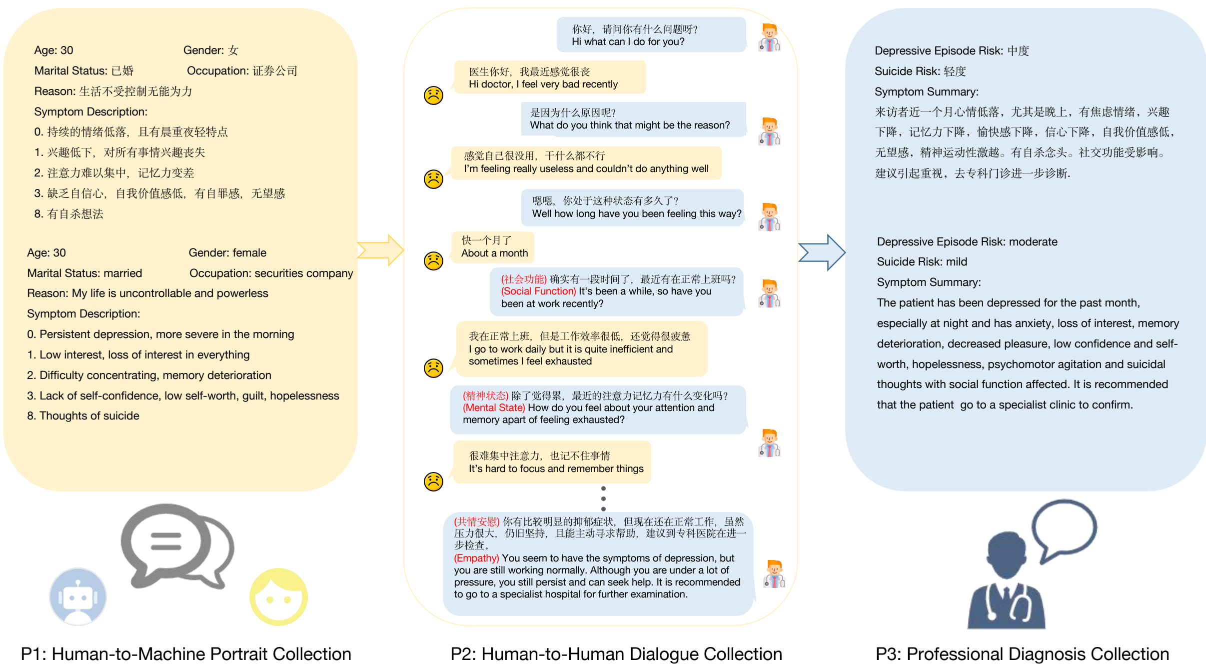

Figure 2: The 3-Phases Data Collection: P1, P2, and P3 denotes the three phases in data collection

2.1 Human-to-Machine Portrait Collection

To overcome the impracticability in obtaining patients' medical records covered by doctor-patient confidential protocol, we designed a consultation chatbot based on the state machine, which utilizes fixed questions from clinical criteria to document each user's depression symptoms and *demographic information such as age, gender, marital status and occupation*. *Depression symptoms* are prompted accordingly, including *mood, interest, mental status, sleep, appetite, social function, and suicidal tendency*. Users are invited to respond concisely, e.g., yes/no answer and severity estimation. Combined, we obtained a voluntary and 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 . The result is shown in Table 1. Sixty-eight portrait providers reported being diagnosed with depression in an authorized clinic. Among these providers, 53 are currently experiencing a depressive episode.

2.2 Human-to-Human Dialogue Collection

To guarantee the quantity, quality, and professionalism of our consultation dialogues, we conducted conversation simulations under the guidance of psychiatrists, following portraits collected in Phase 1. In particular, we first gathered a small number of dialogues between doctors and patients in real scenarios. Based on the prerequisites mentioned above 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 are then divided into doctor and patient roles; 2) Annotation: During the conversation, they are required to annotate topic transitions; 3) Peer Assessment: doctor and patient roles rate each other on multiple dimensions after the conversation.

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

2.2.1 Design and Training

Acting Patients It should be noted that most of our patient actors are not depressive patients. To help them better interpret the symptoms in the patient portraits, we provide detailed explanations, including the severity and duration, and some patients' self-reports to help them understand their inner feelings. Based on the accurately expressed symptoms, they extend the natural expressions of each aspect following doctors' inquiries in the conversation.

Acting Doctors Firstly, we invite licensed psychiatrists and clinical psychotherapists to initiate consultation conversations with actual depressive patients, from which we collect reference conversations. Then based on these essential histories, combined with ICD-11 and DSM-5, we compile 41 symptom items necessary when diagnosing depression and design the questioning logic between questions of symptoms from mild to severe. The inquiries weren't set as specific expressions for data diversity. Thus, the acting doctors needed to use colloquial rhetoric to ask relevant information involved in these questions and obtain enough information from the patient. Meanwhile, to further improve the dialogue experience, we require the acting doctors to conduct a user-oriented dialogue and provide emotional support when necessary. All acting doctors start the dialogue simulation after completing the training process.

2.2.2 Topic Annotation

Considering that the depression diagnostic dialogue has ambiguity between the chat and task-oriented dialogue, it's difficult to define a clear ontology as other task-oriented dialogues(Chen et al., 2022b). To facilitate dialogue generation, we conducted topic annotation on doctors' utterances. According to core symptoms covered in the clinical criteria, we categorized the dialogue topics into *mood, interest, mental status, sleep, appetite, somatic symptoms, social function, suicidal tendency, and screening*. Notably, we included *empathy* as a special topic since it is an essential part of clinical practice. The doctor actors were asked to mark the topics for each utterance during the conversation.

2.2.3 Peer Assessment

After the conversation, both sides are required to rate each other in several dimensions for the need for quality control which will be detailed in 2.4.

2.3 Professional Diagnosis Collection

To ensure the accordance with clinical protocol, we further invite professional psychiatrists and clinical psychotherapists to screen the dialogues that meet the diagnostic standards and provide psychiatric diagnostic results and symptom summaries. At the same time, they score the acting doctors and patients separately with the real-scenario resemblance degree.

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

2.4 Quality Control

Hierarchical screenings are conducted to control the data quality: whether it is up to clinical standard and can satisfy our model training purpose. Besides psychiatrists' clinical protocol screening mentioned in part 2.3, we adopt a variety of paradigms to conduct quality examinations for better training. 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 collected a total of 4,428 conversations and finally retained 1,339 (30%) after our stringent up-to-clinical-standard 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 and avg. 877.6 tokens per dialogue, significantly longer than previous related datasets, suggesting the discrepancies of a diagnosis dialogue task and its distinguished data requirements. Meanwhile, our dataset has colloquial and diverse expressions shown by the number of n-grams and avg. 14.4 tokens per utterance.

3.2 Depression Severity Analysis

To observe differences in patients with different depression severity, we analyzed conversational and summary symptom statistics by seriousness.

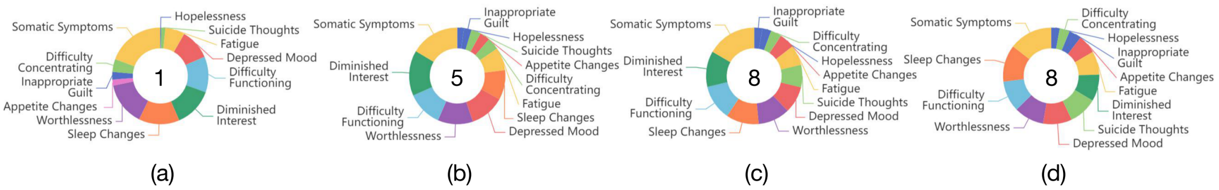

Figure 3: The Symptom Ratio of Summaries: the depressive episode severity increases from (a) to (d) with avg. number of symptoms in the center of each pie chart

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. tokens per dialogue	877.6	381.8	495.8
Distinct 3-grams of dialogues	245,553	148,269	128,203
Distinct 4-grams of dialogues	452,012	251,121	224,476
Distinct 5-grams of dialogues	617,233	324,738	304,128
Avg. tokens per utterance	14.4	12.3	16.6
Avg. tokens per symptom summary	84.4	-	-

Table 3: D^4 Statistics

Distribution Feature We present statistics on patients' severity of depressive episodes in Table 4. As the degree of depression worsens, the turns and dialog lengths get longer due to doctors' more in-depth questions on specific topics. The diagnostic summaries are also longer to include more symptoms. The most frequent topics are also subject to change with severity: *suicidal tendency* is more likely to be questioned among severer patients.

Category	Control	Mild	Moderate	Severe
Dialogues	430	342	368	199
Avg. turns	17.9	21.3	23.7	26.0
1st frequent topic	Emp.	Emp.	Emp.	Emp.
2nd frequent topic	MS	MS	MS	Suicide
3rd frequent topic	Sleep	Mood	Suicide	MS
Avg. tokens of symptom summary	59.8	82.0	100.5	111.9

Emp.:Empathy **MS**:Mental Status

Table 4: Depression Severity Statistics in D^4

Analysis of Symptom Summary We annotated the 13 core symptoms in the symptom summary according to ICD-11. From Figure 3, we observe a difference in the symptom number and ratio from diagnosis summaries of varying severity. As shown in Chart (a), control participants have only a few symptoms, and most are superficial symptoms like sleep changes and worthlessness, commonly in healthy populations. As the condition worsens, the patient has more symptoms, the proportion of each symptom in the summary is gradually averaged, and suicide thoughts become more frequent. The moderate and severe patients share the same average symptom number, indicating that a more fine-grained classification of depression severity requires additional information besides the number of symptoms, such as the duration and severity of each symptom.

3.3 Topic Analysis

To analyze the characteristics of the doctor's consultation method, we provide perspectives on topic distribution, transition, and lexical features of empathy.

Topic Distribution To better analyze the proportion of different symptoms, we regrouped the 10 topics annotated by acting doctors. *mood, interest, mental status, social function* are grouped into *core* and *sleep, appetite, somatic symptom* are grouped into *behavior*. Figure 4 shows the propotion of regrouped topics. *Core* and *behavior* occupy 63.17% of the conversation, followed by *empathy* at 23.1%, indicating that empathy plays an important role in such a psychiatric diagnosis-oriented dialogue.

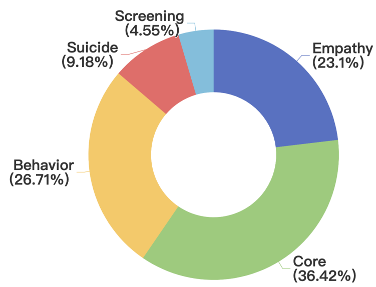

Figure 4: Topic Proportion

Topic Transition Figure 5 illustrates the topic-transition process. Unlike other commonly seen dialogues where the topic rarely extends over one turn, diagnosis topics consistently occur across turns. Further, core symptoms like *mood*, *interest* are usually inquired in the beginning, gradually move to behavior symptom such as *somatic symptom* and *suicide*, which are normally 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.

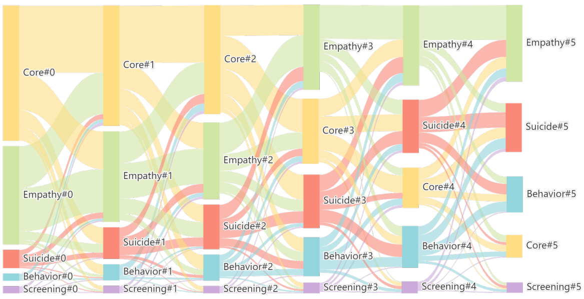

Figure 5: Topic Transitions. Topics over every 3 turns are visualized. The height represents the absolute number of dialogues at this topic.

Lexical Analysis of Empathy As shown in Figure 4, *empathy* accounts for a large proportion, indicating its importance and commonness. We extract its lexical features and observe that the empathy expressions in our dataset could mainly be divided into 4 aspects: 1) `understanding`: "will understand/is normal" to express understanding of the patient's situation; 2) `encouragement`: "is valuable" to help patients regain confidence; 3) `suggestions`: "you can try/try" to encourage patients to make changes and try; 4) `blessings`: "you will get well soon" to express blessings to the patient. In actual practice, providing empathetic and emotional support improves the medical experience and is a critical component of ensuring the success and completion of a diagnostic session(Hardy, 2019).

4 Comparison with Related Datasets

D^4 is compared with related datasets and manifested its characteristics as having more dialogue turns and utterances with a sufficient number of dialogues for model training (see Table 5). This again emphasizes that depression diagnosis is distinguished from current dialogue types and exhibits specific challenges with existing data.

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 (Chen et al., 2022a), 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

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)	Emotional 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 5: Comparison with Related Datasets

dialogue datasets mainly involve daily scenarios instead of clinical practice. Therefore, the number of dialogue turns is relatively small, with little attention paid to providing emotional support.

Emotional Support Datasets A few dialogue studies on mental health address users' emotions in the dialogue process and endeavor to motivate users suffering from a mood disorder. For example, Saha et al. (2021) presents the dialogue dataset MotiVAte to impart optimism, hope, and motivation for distressed people. Recently, works like ESConv (Liu et al., 2021) switch their attention to construct a professional emotional support dialog Systems. However, they are mainly concerned with providing encouragement and advice to patients instead of providing professional diagnoses for screening purposes.

Medical Diagnosis Dialogue Datasets Some medical dialogue datasets target at diagnosis, such as MedDG (Liu et al., 2020) and MedDialog (Zeng et al., 2020). Meanwhile, some datasets aim at biomedical language understanding such as CBLUE (Zhang et al., 2021a). However, these efforts focus mainly on somatic symptoms and physical diseases. MedDialog, although containing a small amount of psychiatric data, lacks professional psychiatric annotations, limiting its usage for a depression diagnosis dialogue system. It should be noted that the diagnosis process of depression essentially differs from that of somatic disorders. According to ICD-11 (The World Health Organization, 2022), 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 extracted from patients' subjective statements to provide unbiased diagnoses, leading to a longer, multi-domain dialogue process.

Depression-Related Dialogue Dataset Along with the worldwide attention on depression, a few dialogue datasets strongly related to depression are constructed, such as DAIC-WOZ (Gratch et al., 2014), a multi-modal dataset. The dataset consists of face-to-face counseling conversations between a wizard interviewer and patients who suffer from depression, anxiety, etc. However, DAIC-WOZ only includes 189 dialogues without specific annotations, which is insufficient for dialogue generation training.

5 Experiments

5.1 Tasks

Upon the construction of D^4 with 1,339 well-annotated and up-to-clinical-standard depression diagnosis conversations, we can support an entire generation and diagnosis process mirroring the real-life clinical consultation scenario. We split the entire depression diagnosis dialogue procedure into 4 subtasks: **Response Generation** aims to generate doctors' probable response based on the dialog context; **Topic Prediction** predicts the topic of the response based on the dialogue context. In our experiments, we jointly optimize the topic prediction model and the response generation model. We take the topic as a special first token of dialogue response; **Dialogue Summary** generates symptom summaries based on the entire dialog history; **Severity Classification** separately predicts the severity of depressive episodes and the suicide risk based on the dialogue context and dialogue summary. Binary (positive/negative) and fine-grained 4-class (positive further classed into mild, medium, and severe) classifications are both investigated.

5.2 Backbone Models

We use Transformer (Vaswani et al., 2017) pretrained on MedDialog (Zeng et al., 2020), BART (Lewis et al., 2019) pretrained on Chinese datasets (Shao et al., 2021), CPT (Shao et al., 2021) and BERT (Devlin et al., 2019) as backbone models to conduct the experiments.

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. To show the generation diversity, we also compute DIST-2 (Li et al., 2015). We implement jieba[2] for tokenization and compute the metrics at the word level.

Model	BLEU-2	ROUGE-L	METEOR	DIST-2	Topic ACC.
Transformer-	7.28%	0.21	0.1570	0.29	-
BART-	19.29%	0.35	0.2866	0.09	-
CPT-	19.79%	0.36	0.2969	0.07	-
Transformer	13.43%	0.33	0.2620	0.04	36.82%
BART	28.62%	0.48	0.4053	0.07	59.56%
CPT	29.40%	0.48	0.4142	0.06	59.77%
Transformer-BERT	23.95%	0.40	0.3758	0.22	61.32%
BART-BERT	33.73%	0.50	0.4598	0.07	61.32%
CPT-BERT	34.64%	0.51	0.4671	0.06	61.32%
Transformer*	25.37%	0.41	0.3905	0.04	-
BART*	37.02%	0.54	0.4920	0.07	-
CPT*	**37.45%**	**0.54**	**0.4943**	**0.06**	**-**

$-$ means topics are excluded, BERT means topics predicted by BERT are given as prompt, $*$ means golden topics are given as prompt

Table 6: Evaluation Results of Response Generation and Topic Prediction

	Core	Behavior	Empathy	Suicide	Screening
F1	0.63	0.69	0.24	0.49	0.35

Table 7: F1 of Topics Predicted by BART

Results for the *response generation task* are presented in Table 6. Five observations can be drawn: 1) BART and CPT exhibit similar generation performance on our dataset; 2) Both models vastly outperform Transformer, which is pretrained on the medical corpus, suggesting that, on the one hand, pertrained language models with more parameters could improve generation performance; on the other hand, depression diagnosis is different from traditional somatic-oriented medical dialogues; 3) Based on the topic of response predicted by the model itself, the model could generate a more accurate response, which is of great significance for the model to be applied in real human-machine interaction scenarios; 4)Based more accurate topics predicted by BERT, response generation performance is enhanced, indicating that higher topic prediction accuracy can effectively improve generation accuracy. 5) Given golden topics, generation performance can be further enhanced.

Topic Prediction accuracy results are shown as Topic ACC. in Table 6. We adopted the topic category regrouped in 4 and similar trend is observed: BART $\approx$ CPT $>$ Transformer. The F1 of each topic (see Table 7) shows that the accuracy of empathy is the bottleneck of this task, indicating that proper timing of empathy remains challenging for models and is a potential direction for further work.

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

Results for *Dialog Summary* are listed in Table 8, CPT is on par with BART regarding the N-gram overlap with human references. Nevertheless, CPT exhibits a higher DIST-2 score, suggesting its superiority in generation diversity. We manually annotated summaries by 13 symptoms from ICD-11 and calculated the summaries' sample average F1 score on the multi-label classification task of symptoms, where CPT and BART perform the same. It shows that the summary generated by the model can accurately summarize most symptoms.

Model	BLEU-2	ROUGE-L	METEOR	DIST-2	Symptom F1
BART	16.44%	0.26	0.25	0.19	0.67
CPT	**16.45%**	**0.26**	**0.24**	**0.21**	**0.68**

Table 8: Evaluation Results of Dialog Summary Task

Severity Classification Binary and 4-class classification are evaluated by average weighted precision, recall, and F1 by sklearn[3], and results of depression severity and suicide risk severity are shown in Table 9 and Table 10. For the classification of depression severity, we conducted experiments based on dialogue history and symptom summaries respectively. The evaluation results show that the accuracy of 2-classification and 4-classification based on summaries is significantly improved, indicating that symptom summaries have extracted vital information from the dialogue, being extremely helpful for diagnosis. Although the results of 4-classification tasks are relatively poor compared with the performance in 2-classification tasks, as a screening tool, the binary classification results are already sufficient in the practical application of the system.

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
	summary	BERT	0.90±.02	0.90±.02	0.90±.02
		BART	0.89±.03	0.89±.03	0.89±.03
		CPT	**0.92**±.01	**0.92**±.02	**0.92**±.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
	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

Table 9: Depression Severity Classification Results

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

Task	Input	Model	Precision	Recall	F1
2-class	dialog	BERT	0.81±.02	0.78±.02	0.79±.02
		BART	0.77±.02	0.75±.02	0.75±.02
		CPT	**0.84**±.02	**0.82**±.03	**0.82**±.03
4-class	dialog	BERT	0.72±.03	0.64±.04	0.66±.03
		BART	0.70±.05	0.66±.04	0.65 ±.03
		CPT	**0.76**±.02	**0.68**±.02	**0.70**±.02

Table 10: Suicide Severity Classification Results

5.4 Human Interactive Evaluation

To comprehensively evaluate the model's conversation experience with the user, we include human interactive evaluation for CPT with a rule-based chatbot. Evaluators were invited to chat with both bots in a random order upon the provided patient portrait and rated on 4 aspects with a 1-5 scale: **Fluency** measures how fluently the conversation flows; **Comforting** measures how comforting the responses are; **Doctor-likeness** measures to what extent does the chatbot flexibly adjust the topic according to the patient's description; **Engagingness** measures to what time could the chatbot maintain their attention to continue the chat.

CPT vs Base	Metric			
Result	Fluency	Comforting	Doctor-likeness	Engagingness
Win	**19**†	**18**†	**17**†	**20**†
Lose	14	13	14	11

‡/†means p-value < 0.05/0.5 respectively

Table 11: Human Evaluation Results

The interactive human evaluation results are illustrated in Table 11. The CPT-based chatbot outperforms rule-based bots on all four metrics, suggesting that dialogue models can help us build more human-like and user-friendly depression diagnosis systems. In particular, the discrepancy in engagingness indicates that users prefer chatbots that can better understand and comfort users in completing the depression screening process. We give some empathy examples of human interactive evaluation in Table 12, indicating that the model can generate diverse empathy representations from different aspects.

6 Conclusion

In this paper, we designed a 3-phase data collection and constructed a close-to-clinical-practice and 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

Aspects	Examples
Understanding	*I could understand you.* *I could understand your feelings.*
Encouragement	*Everyone has their own value.* *Everyone has their own characteristics.*
Suggestion	*It is suggested to seek professional medical help as soon as possible.*
Blessing	*Wish you a happy life!* *Hope you get well soon!*

Table 12: Empathy Examples in Human Evaluation

with objective and human evaluation. The evaluation results show that the model-based chatbot outperforms traditional rule-based dialogue bots in all metrics, indicating that a more user-friendly dialogue system can be built with our dataset. However, the model is still not effective enough in generating appropriate empathic responses suggesting that the model needs further improvement to generate more appropriate empathy during the consultation process.

Limitations

Our work has some limitations. The principal limit of our work is that our dataset D^4 is in Chinese, which in line with Chinese culture and expression habits. Therefore, it may not be applicable to translate the conversations into another language directly, so further exploration is required for our work to transfer to other languages. However, considering that there are no similar datasets in other languages published before, we hope that our data collection method and data form (dialog+summary+diagnosis) could inspire more research on this unique type of dialogue in the future.

Additionally, for patient privacy protection, our dialogue data is collected in a simulated manner, not from real scenarios. This approach helps construct a more secure and generalizable consultation dialogue system because we have defined the acting doctors' behaviors during the data collection process, that is, the system behavior range. But it should be mentioned that our dataset cannot restore the expressions of actual patients and doctors. For this reason, the textual features of acting patients in our dataset are not sufficient for the classification of depression. Therefore, it is meaningful to explore the construction of a more empathy-driven and diagnostic-accurate consultation dialogue system based on our dataset rather than conduct textual depression classification.

Ethics Statement

This research study has been approved by the Ethics Review Board at the researchers' institution (Ethics Approval No. I2022158P). Different stages in data collection comply with corresponding ethical requirements and we endeavour to protect privacy and respect willingness of our data providers and annotators.

Specifically, our data collection falls under the Personal Information Protection Law of the People's Republic of China. In the phase of portrait collection, the collection application was developed as a WeChat mini program[4], which complied with the privacy protection agreement and passed the security and privacy check of WeChat mini program before releasing on the platform. Furthermore, all the portrait providers signed an informed consent form to give permission to collect their anonymous information for research purposes.

In the phase of the dialogue collection process, all the workers and annotators are informed about the purpose of our data collection and equally paid for their workload. In the phase of the dialogue examination process, the psychiatrists and psychotherapists are licensed to practice and paid equally for their workload.

To protect users' privacy, we anonymized the portraits by storing them without a one-to-one correspondence between the identification information required for user login and the data we use in research. Therefore, all the information that could uniquely identify individual people is excluded from our dataset and research process. Regarding offensive content, we rigorously filtered the dataset manually to ensure that it did not contain any offensive content or words encouraging patients to self-harm and commit suicide. We will also require the users of D^4 to comply with a data usage agreement to prevent the invasion of privacy or other potential misuses.

Acknowledgements

This project is supported by National Natural Science Foundation of China (Grant No.61901265, 92048205), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), SJTU Medicine-Engineering Project (No.YG2020YQ25), and Xuhui District Artificial Intelligence Medical Hospital Cooperation Project (2021-005).

[4]https://developers.weixin.qq.com/miniprogram/en/dev

References

Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed. edition. Autor, Washington, DC.

Raúl Arrabales. 2020. Perla: a conversational agent for depression screening in digital ecosystems. design, implementation and validation. *arXiv preprint arXiv:2008.12875*.

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.

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*.

Zhi Chen, Jijia Bao, Lu Chen, Yuncong Liu, Da Ma, Bei Chen, Mengyue Wu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022a. Dialogzoo: Large-scale dialog-oriented task learning. *arXiv preprint arXiv:2205.12662*.

Zhi Chen, Yuncong Liu, Lu Chen, Su Zhu, Mengyue Wu, and Kai Yu. 2022b. Opal: Ontology-aware pretrained language model for end-to-end task-oriented dialogue. *arXiv preprint arXiv:2209.04595*.

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.

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. *arXiv preprint arXiv:1805.04833*.

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.

Carter Hardy. 2019. Clinical sympathy: the important role of affectivity in clinical practice.

John Hart, Jonathan Gratch, and Stacy Marsella. 2017. How virtual reality training can win friends and influence people. In *Fundamental Issues in Defense Training and Simulation*, pages 235–249. CRC Press.

Shashank Jaiswal, Michel Valstar, Keerthy Kusumam, and Chris Greenhalgh. 2019. Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 81–87.

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.

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. *Journal of Biomedical Informatics*, 113:103632.

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.

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. *Scientific reports*, 7(1):1–7.

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.

Ryan M Schuetzler, G Mark Grimes, Justin Scott Giboney, and Jay F Nunamaker Jr. 2018. The influence of conversational agents on socially desirable responding. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, page 283.

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. 2022. Icd-11 for mortality and morbidity statistics.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

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.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021a. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021b. Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 135–143.

A Data Example

The portrait (Figure 6), the dialogue (Figure 8 and Figure 9), and diagnosis (Figure 10) belong to the same data example in our dataset. We marked the topic (if any) of the doctor's responses in the conversation history. In this example, the doctor combined sleep and appetite into one question, so only one topic of appetite was marked. In addition, for the convenience of presentation, we have combined the doctor's multiple utterances of the same turn into one sentence. To compare machine generation performance with humans, we provide data examples of the same portrait in this section and Section E - Human Interactive Example. Dialogues in D^4 were simulated based on diverse portraits showed in Section B. More data examples can be found in website https://x-lance.github.io/D4.

Portrait
Demographic Information
年龄：21　性别：男　职业：学生　婚姻状态：未婚
Age:21 Gender:male Occupation: student Marital status: unmarried
Symptom Description
1. 兴趣低下，对所有事情兴趣丧失　Low interest in everything
2. 注意力难集中，疲倦　Difficulty in concentrating and fatigue
3. 缺乏自信心，自我价值感低　Low confidence, worthless
4. 精神运动性激越和迟滞　Psychomotor agitation and retardation
5. 个人生活功能受损，人际关系不稳定，工作学习效率下降
Difficulty functioning in personal life, social relationship and education

Figure 6: A Portrait Example

B Data Characteristics

Statistics of Portraits' Demographic Information The aggregated demographic information of 478 portraits is provide in Figure 7.

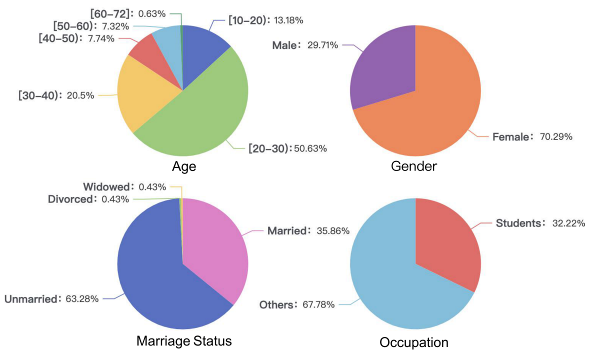

Figure 7: Aggregated Demographic Information

Topic Examples In Figure 11, we present the 10 topics with their typical examples and hot words.

Lexical Feature of Empathy In Figure 12, we show the lexical feature of empathy words in our dataset in the sunburst figure.

C Backbone Model Introduction

Rule-based Model Without existing chatbots having the same function, we built the rule-based chatbot by state machine as the baseline. Based on ICD-11 (The World Health Organization, 2022), DSM-5 (American Psychiatric Association, 2013), the bot covers the same topics as the dialogue simulation process mentioned in 2.2.2. This robot has fixed question templates and recognizes the user's answer based on regular matching, based on which it performs state jumps until all symptom information is acquired.

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[5]. 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 for both text generation and summary tasks. For this reason, we use Bart pre-trained 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 tasks 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[6] which is pre-trained on eight popular Chinese NLP datasets, to conduct the classification task.

D Training Details

The division of train, validation, and test sets for all experiments is close to 8:1:1, and the data of

[5]https://github.com/huggingface/transformers
[6]https://huggingface.co/hfl/chinese-macbert-base

PART1

Doctor: 你好 Hello

Patient: 医生我最近觉得啥事情都没有意思，不知道怎么回事 I feel that nothing has been interesting recently, and I don't know what happened to me.

Doctor: 你现在是大几? What grade are you currently in?

Patient: 大四了 I'm a senior student.

Doctor: 平时学习压力大吗? 最近有没有发生什么特别的事情? 给你带来很大的压力或者打击? Is your study stressful? Has anything special happened recently which brings you a lot of pressure?

Patient: 学习压力很大，尤其是最近同学们都保研了，只有我还得考研，每天都很焦虑 I've been under a lot of pressure from studying recently, especially since my classmates have all been admitted to a graduate program but me, which makes me very anxious.

Doctor: 社会功能 嗯，你是哪里人? 现在在哪里? 平时有亲人或者好朋友可以聊天，倾诉吗? Social Function Um, where are you from? Where are you now? Do you usually have parents or friends to talk to about your feelings?

Patient: 我是外地人，现在在学校。最近和朋友来往也比较少了，我平时和我家人也很少联系，我平时也不太喜欢和别人聊天说自己的事情什么的。I am an out-of-towner and now at school. I have had less contact with my friends recently. I usually have little contact with my family members and don't like chatting with others about my feelings..

Doctor: 哦，那这种情况有多长时间了 Um, so how long have you been in this situation?

Patient: 从上个月开始吧，就今年开学以来 Since last month, probably since the beginning of this semester

Doctor: 以前你是什么样的人? 是不是感觉自己像变了一个人? What kind of person were you before? Do you feel like a different person?

Patient: 有点，我以前不会觉得什么事情都没意思。Kind of, I didn't feel like everything was boring before.

Doctor: 除了觉得什么都没意思，和他人的交往、沟通减少，还有什么变化? Apart from feeling that nothing is interesting and the interaction and communication with others have decreased, do you feel that there is any other change in yourself?

Patient: 还会觉得很累，一方面考研压力很大，但是一方面我效率又很低 I am tired every day. I feel stressed about the exam, but my learning efficiency is low.

Doctor: 食欲 嗯，吃饭睡觉有什么变化吗? Appetite Hmm, is there any change in eating and sleeping?

Patient: 没有太大变化，和以前一样 Not much changed, same as before.

Doctor: 精神状态 有没有觉得脑子转的比之前慢，记忆力比以前差，注意力也很难集中? Mental Status Do you feel that your brain is turning slower than before, your memory is worse than before, or it is difficult to concentrate?

Patient: 有的，主要是注意力很难集中，然后有的时候还会突然脑子一片空白，出门之后不知道自己打算去哪。Yes, it is difficult to concentrate, and sometimes my mind suddenly goes blank, and I don't know where I plan to go.

Doctor: 嗯，你觉得现在的情况和考研很有关系吗? 如果给你放假一个月，不用考了，你也保研了，你觉得你现在的状况会改变吗? Well, do you think your current situation has much to do with the coming examination? If you were given a month off and didn't need to take the exam. Do you think your current situation will change or not?

Patient: 我觉得可能会吧，但是我上大学以来一直都挺难受的 I think it might be, but I've been having a hard time since college.

Doctor: 情绪 你说上大学来一直挺难受什么意思? 是从大一开始就感觉情绪不大好吗? Mood What do you mean? Did you feel bad from your first year?

Figure 8: A Dialogue Example - Part1

PART2

Doctor: 嗯，你觉得现在的情况和考研很有关系吗? 如果给你放假一个月，不用考了，你也保研了，你觉得你现在的状况会改变吗? Well, do you think your current situation has much to do with the coming examination? If you were given a month off and didn't need to take the exam. Do you think your current situation will change or not?

Patient: 我觉得可能会吧，但是我上大学以来一直都挺难受的 I think it might be, but I've been having a hard time since college.

Doctor: 情绪 你说上大学来一直挺难受什么意思? 是从大一开始就感觉情绪不大好吗? Mood What do you mean? Did you feel bad from your first year?

Patient: 就是同学之间竞争压力很大，我也一直不是很优秀的学生，感觉和高中比起来挺有落差感的，所以时常会觉得自己不行，老觉得自己会把事情搞砸。The competition among classmates is very stressful, and I'm no longer as outstanding a student as I was in high school, so I often feel that I am not good and will screw things up.

Doctor: 精神状态 嗯嗯，是觉得自己不如别人，不自信吗? Mental Status Well, do you feel that you are not as good as others and are not confident?

Patient: 是的 Yes

Doctor: 社会功能 那你三四年的学习下来，有没有遇到什么特别的困难? 能正常维持学习吗? Social Function After three or four years of study, did you encounter any special difficulties? Can you maintain your studies normally?

Patient: 可以的，我倒是没有觉得自己学习跟不上别人，就一直普普通通的 No, I don't feel that I can't keep up with others in my studies and just always been ordinary

Doctor: 筛查 有开心或者兴奋的时候吗? Screening Have you ever felt particularly happy or excited?

Patient: 没有 Never had

Doctor: 筛查 嗯嗯，你家里有抑郁或其他的家族史吗? 精神方面的 Screening Well, Is there anyone with depression or other mental illness in your family?

Patient: 没有 No

Doctor: 自杀想法 你现在或者过去有没有轻生的想法或行动吗? Suicide Do you have suicidal thoughts or actions now or in the past?

Patient: 没有 No

Doctor: 你喜欢你现在的专业吗? Do you like your current major?

Patient: 喜欢 Yes, I like it.

Doctor: 好的，你现在听上去有点抑郁情绪，需要进一步的检查，你愿意到专科医院就诊吗? Well, you sound a little depressed now and need further examination. Are you willing to go to a specialist hospital?

Patient: 愿意的，我现在情况严重吗? Yes, am I in a serious condition now?

Doctor: 共情安慰 你有比较明显的抑郁症状，但仍然能坚持学习，现在还在准备考研，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 apparent symptoms of depression, but you could still persist in your studies and prepare for the coming exam under a lot of pressure. Also, you can take the initiative to seek help. The situation is not very serious. It is recommended to go to a specialist hospital for further examination and a precise diagnosis treatment.

Figure 9: A Dialogue Example - Part2

Diagnosis
抑郁风险：中 自杀风险：无 Depressive episode severity: moderate Suicide risk: no risk
Symptom Summary
来访者有较明显的情绪低落、自我评价低、负性评价和归因、精力不足、兴趣减退、记忆力下降、注意力不集中等抑郁症状，但无消极言行，有求治愿望，且仍旧维持比较好的社会功能。建议专科医院就诊，明确诊断。The patient has obvious depression symptoms including depressed mood, low self-evaluation, negative evaluation and attribution, reduced energy, diminished interest, decreased memory, difficulty concentrating, etc. However, he has no negative words and deeds, has a desire to seek treatment and still maintains a relatively good social function. It is recommended to seek medical treatment in a specialist hospital.

Figure 10: A Diagnosis Example

different depression severity are also internally distributed according to the above ratio.

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 the test.

For the Transformer, we use the implementation by HuggingFace[7]. 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 the highest BLEU-2 score on the evaluation set is chosen for the test.

We spliced multiple sentences of the doctor in the same round into the dialogue history, and selected the last topic as the topic of the new sentence. Due to the limitation of models' positional embedding, we intercepted 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.

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

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 11: Samples of Doctors' Topic

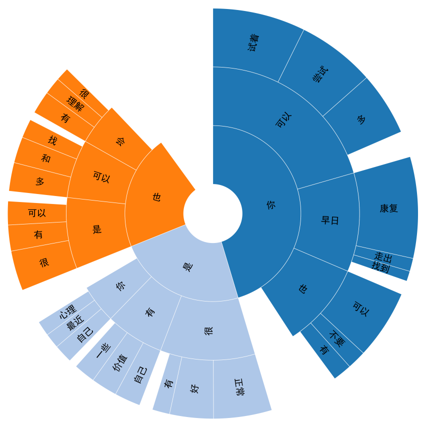

Figure 12: Lexical Feature of Empathy

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 the test.

If the input dialog history is longer than the model's input size, we retain the 512 tokens in the middle of the dialog.

Severity Classification For BERT, BART, and CPT models, 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). Models are trained for 30 epochs. The one with the best F1-score metric on the evaluation set is selected for the test.

For the classification based on dialog history, we retain 512 tokens in the middle of the dialog. For the classification based on dialog summary, we retain 128 tokens in the middle of the summary.

E Generation Examples

Response Generation As shown in Figure 13, we selected one representative example of the generated responses by different models. The examples in the figure show us that the correct topic helps the model generate more reliable and secure replies.

Dialog Summary Generation In Figure 14, we present an example of the generated summary by different models. The models list most symptoms of the patient.

Human Interactive Example We give a dialogue example with dialog summary and depressive severity generated by CPT during human evaluation in Figure 15 and human evaluation in Figure 16. In parentheses before the chatbot's sentence, we marked the topic predicted by the model. To clarify the correspondence between dialogue and summary, We have identified the correct symptom in the symptom summary with the same color as its location in the conversation. It can be seen that the model completed the entire consultation dialogue task and gave a dialogue summary covering almost all symptoms accurately.

F Worker Training Method

Acting Patients To help acting patients better interpret the symptoms in the patient portraits, we

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。 [病人]快高考了，我感觉自己没有能考上的信心**[医生]**我了解了，因为高考的压力你觉得很不自信，那会不会觉得自己差劲呢？[病人]是的，感觉自己在拖累父母**[医生]**那有没有觉得对未来很绝望呢？**[病人]**嗯嗯是的，所以之前也尝试过自杀**[医生]**你的父母知道这件事吗？**[病人]**我没跟他们讲，有些回避和他们的接触。 **[Patient]** It is about to take the college entrance examination, but I feel like I can't pass it. **[Doctor]** I understand. You feel very unconfident about the pressure of the exam. Do you feel bad about yourself? **[Patient]** Yes, I am a burden on my parents. **[Doctor]** So, do you feel hopeless about the future? **[Patient]** Mmmmm yes, I even attempted suicide before **[Doctor]** Did your parents know about it? [Patient] I didn't tell them, and I avoided contact with them.
Topic: 共情安慰 Empathy
Groundtruth: 那以后如果还有想要自杀的话，就算不愿意告诉父母，也要寻求专业机构的帮助好吗？Well, if you still want to commit suicide, even if you don't want to tell your parents, you should seek professional help, okay?
Transformer-: 这种状态持续多久了？ How long have you been in this state?
BART-: 最近两周心情怎么样呀？ How have you been feeling for the past two weeks?
CPT-: 我明白了，那最近心情怎么样呀？ I understand. How are you feeling lately?
Transformer: [核心]嗯嗯，那很好呀最近情绪怎么样呢？基本的生活事务能不能正常进行？Mmmm, that's great. How are you feeling lately? Can your basic life affairs be carried out normally?
BART: [核心]我明白了，那最近心情怎么样呀？ I understand. How are you feeling lately?
CPT: [核心]我明白了，那最近心情怎么样呀？ I understand. How are you feeling lately?
Transformer*: [共情安慰]嗯嗯，有时间还是要好好调整一下自己的状态。Well, It's recommended that you should adjust your state if you have time.
BART*: [共情安慰]我明白了，可能是压力太大了，建议尽快寻求专业医生的帮助。I understand, it may be because the pressure is too great, it is recommended to seek professional medical help as soon as possible.
CPT*: [共情安慰]我明白了，我能够理解你的心情，所以我们可以尝试过伤害自己的想法吗？I get it, I can understand your feelings, so can we try to hurt ourselves?

Figure 13: Examples of generated response

Groundtruth: 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。

The patient has difficulty concentrating; psychomotor agitation and retardation, with somatic symptom: dizziness; worthless; poor sleep quality and shallow sleep.

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。

The patient has recently been depressed, unable to concentrate in class, decreased learning efficiency, sleep disturbance, difficulty falling asleep, physical discomfort, slow thinking.

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。

In the past month, the patient experienced: decreased concentration, decreased self-evaluation, impaired cognitive function, decreased interest, lack of energy, sleep disturbance, somatization symptoms, dizziness.

Figure 14: Examples of Dialog Summary Generation

provide detailed explanations of the symptoms in Figure 17 and Figure 18, including the severity and duration. Besides expressing symptom accurately, they are required 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 We compile the 41 symptom items in Figure 19 and Figure 20 that doctors need to know when diagnosing depression, and design the questioning logic between questions of asking symptoms from mild to severe. The basic requirement is to obtain enough information from the patient during the conversation. At the same time, in order to further improve the dialogue experience, we require the acting doctors to: 1) Conduct the dialogue centered on the patient's complaint, that is, give priority to asking the patient's initiative symptom-related questions; 2) Ask further questions based on the patient's experience to elicit additional disclosure; 3) Give the patient certain feedback, e.g., empathy or comfort words when the patient talks about what they are going through.

G Quality Control

To create transparency about quality control, the statistics of dialogues removed is provided in Table 13. We have collected 4,457 dialogues, and 961 dialogues are removed because they haven't completed the entire diagnosis dialogues. 1,814 dialogues are automatically dropped by the stringent quality control criteria in Table 2. Professional psychiatrists and clinical psychotherapists screening the dialogues dropped 342 dialogues which unsuccessfully meet clinical standards. Eventually, we selected 1,339 dialogues into D^4.

Reason for Removing	Sum
Total	4,457
Unfinished	961
Dropped by Quality Control Criteria	1,814
Dropped by Doctor	342
Our Dataset	1,339

Table 13: Statistics of Removed Dialogues

H The Data Collection Platform

Figure 21 is screenshot of doctors' user interface, and Figure 22 is screenshot of the patients'.

PART1

Patient: 医生，我最近觉得啥事情都没有意思，不知道怎么回事 Doctor, I feel nothing has been interesting recently, and I don't know what happened to me.

Chatbot: [核心]这种状态持续多久了呢？**[Core]** How long have you been in this state?

Patient: 从上个月开始吧，就今年开学以来变成这样的 Since last month, probably since the beginning of this semester.

Chatbot: 最近生活中有发生什么事情吗？Has anything happened in your life recently?

Patient: 学习压力很大，尤其是最近同学们都保研了，只有我还得考研，每天都很焦虑。I've been under a lot of pressure from studying recently, especially since my classmates have all been admitted to a graduate program but me, which makes me very anxious.

Chatbot: [核心]每天大部分时间都是这种状态吗？**[Core]**Are you in this state most of the day?

Patient: 是的差不多每天都是很焦虑很低落。Yes, almost every day I'm very anxious and down.

Chatbot: [核心]会不会在早上或者晚上的时候觉得更难过呢？**[Core]**Do you feel sadder in the morning or at night?

Patient: 早上起来的时候会觉得尤其难过 I feel especially sad when I wake up in the morning.

Chatbot: [核心]你最近有没有觉得对以前喜欢的事情不感兴趣了呢？ **[Core]**Have you recently felt disinterested in things you used to enjoy?

Patient: 有的，我不光是觉得以前喜欢的事情没有意思，我觉得什么事情都没有意思 Yes, I don't just think the things I used to like are boring, I don't think anything is interesting.

Chatbot: [核心]那你只是感受不到快乐了，还是其它的一些感觉，例如悲伤、愤怒的感觉也没有了？**[Core]**Then you just can't feel happiness, or are some other feelings, such as sadness and anger, gone too?

Patient: 这倒是没有，我觉得难过得很 No, I feel very sad.

Chatbot: [核心]嗯嗯，那能正常清理个人卫生、做家务吗？**[Core]**Well, Can you clean up your personal hygiene and do housework normally?

Patient: 可以的，但是我现在不喜欢出门，很少和人来往，以前的朋友都不一起玩了 Yes, but I don't like going out now, and I rarely interact with people. I don't play with my old friends anymore.

Chatbot: [核心]是有意回避社交不想和人往来吗？**[Core]**Do you intentionally avoid social interaction and not want to interact with people?

Patient: 有这样的想法吧，我觉得自己很没用，和别人待在一起也不知道该说什么做什么，越来越自闭 Kind of. I think I am useless, don't know what to say or do when I stay with others and am increasingly autistic.

Figure 15: Human Interactive Example-part1

PART2

Chatbot: [自杀倾向]你会觉得在生活中是自己在拖累别人吗？**[Suicide Tendency]**Do you feel like you are a burden to others ?

Patient: 有时候会有这种想法，我特别害怕自己考不上研究生，回家啃老成为父母的负担。Sometimes I have this kind of thinking, and I am terrified that I will not be able to get into graduate school, and become a burden to my parents.

Chatbot: [自杀倾向]有生活失去希望，无助的感觉？**[Suicide Tendency]**So, do you feel hopeless and helpless in life?

Patient: 只是有一点点吧 Just a little bit.

Chatbot: [自杀倾向]那接下来你有什么规划和打算吗？**[Suicide Tendency]**Do you have any plans for the next?

Patient: 我想专注地学习，好好准备考试，但是我最近感觉自己学习效率也很低，总是集中不了注意力 I want to study hard and prepare well for exams, but recently I feel that my study efficiency is very low, and I can't concentrate all the time.

Chatbot: [行为]最近睡觉怎么样？**[Behavior]**How about your sleep?

Patient: 睡觉一般 Normally

Chatbot: [行为]吃饭怎么样呢？**[Behavior]**And how's your appetite?

Patient: 吃饭还行 Also Normally

Chatbot: [行为]身体上有什么不舒服吗？**[Behavior]**Is there any physical discomfort?

Patient: 有时候会有点头晕恶心 Occasionally feel dizzy and nauseous

Chatbot: [自杀倾向]你有伤害自己的想法吗？**[Suicide Tendency]** Do you have thoughts of hurting yourself?

Patient: 没有 Nope

Chatbot: [自杀倾向]你会觉得死了更好吗？**[Suicide Tendency]** would you feel better dead?

Patient: 不会啊 Nope

Chatbot: 还有其他想和我聊聊的吗？Is there anything else you want to talk to me about?

Patient: 我该怎么办，我好无助 What should I do? I'm so helpless.

Chatbot: [共情安慰]建议你尽快到专科医院进行诊断，别给自己太大的压力 **[Empathy]** It is recommended that you go to a specialist hospital for diagnosis as soon as possible. Please don't put too much pressure on yourself.

Patient: 嗯嗯好的，谢谢医生 Well, thank you.

Chatbot: 那我们问诊结束了，谢谢你的信任 Then our consultation is over. Thanks for your trust.

Symptom Summary: 病人最近情绪低落，兴趣减退，精力缺失，自责自罪感，无望无助，有过轻生的念头，躯体不适感，头晕恶心，思维迟缓。The patient has recently suffered from low mood, loss of interest, lack of energy, self-blame, guilt, hopelessness, suicidal thoughts, physical discomfort, dizziness, nausea, and slow thinking.

Depression Risk Prediction: 中度 Moderate

Figure 16: Human Interactive Example-part2

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 17: 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 18: 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 19: 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 20: Doctors’ questions - 2

Figure 21: Page of doctor

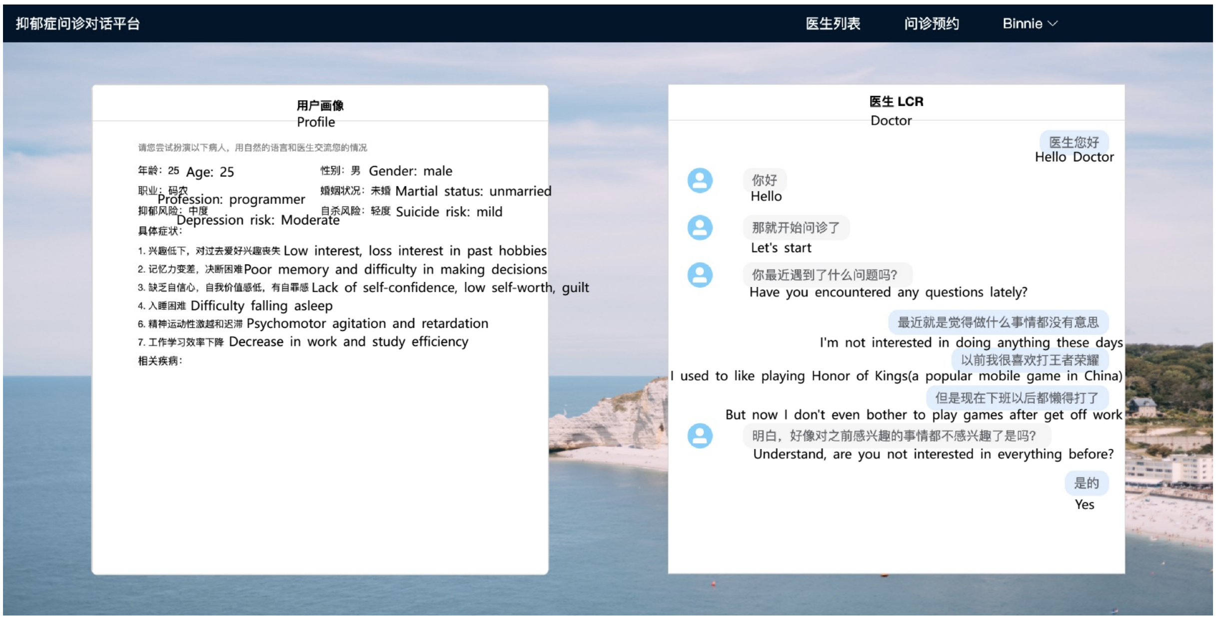

Figure 22: Page of patient