Can LLM Simulate Doctor and Patient for Depression Diagnosis?

Anonymous ACL submission

Abstract

While using chatbots in mental health domain is receiving increasing amount of attention, implementation and evaluation of chatbots in psychiatric diagnostic scenarios remains relatively unexplored. In this work, we introduce a framework named `PsyDial` to simulate conversations between psychiatrists and patients for depression diagnosis. To establish this framework, we collaborate with professionals to identify the key objectives of these chatbots, designing task-specific evaluation metrics aligned with these objectives. Moreover, we explore the potential of ChatGPT in powering these chatbots. Experiments shows that our psychiatrist chatbot exhibits higher levels of empathy and diagnostic accuracy than baseline, but still exhibits several limitations when compared to human psychiatrists; our patient chatbot can replicate human-like behaviors, including emotional fluctuations.

1 Introduction

Application of LLM in mental health has garnered increasing amount of attention, spanning from depression detection (Lamichhane, 2023; Qin et al., 2023) to emotional conversation (Zhao et al., 2023). In particular, Chatbots capable of (i) conducting diagnosis conversations like a psychiatrist or (ii) simulating patients with mental disorders, have significant real-world applications[1]. Psychiatrist chatbots prove effective for mental disorder screening (Pacheco-Lorenzo et al., 2021), while patient chatbots can serve as Standard Patients (SP) in medical education, enhancing the efficiency and cost-effectiveness of the learning process (Torous et al., 2021). Despite the growing interest in applying chatbots in the mental health domain, existing efforts predominantly focus on emotional support (Liu et al., 2021) and mental health therapy (Sabour et al., 2022). There has been only limited exploration of chatbots simulating diagnosis scenarios (Yao et al., 2022).

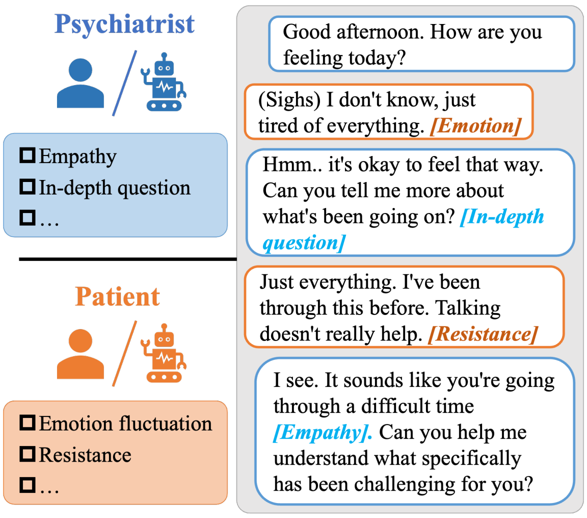


Figure 1: The PsyDial framework is a system of depression diagnosis conversation involving: (1) a psychiatrist conducts an inquiry for diagnosis, and (2) a patient describes their symptoms. Either can be simulated using LLM.

This gap exists due to three challenges. Firstly, the absence of clear objectives and standardized evaluation criteria hinders the development of psychiatrist and patient chatbots. The question of what constitutes an effective chatbot for psychiatric diagnosis remains unanswered. Secondly, ethical and privacy concerns associated with obtaining diagnostic dialogue data pose significant obstacles to training these chatbots (Yao et al., 2022). Lastly, creating a psychiatric diagnosis chatbot is more intricate than developing a general doctor chatbot, particularly because of its target users—individuals with mental disorders. In outpatient scenarios, patients often struggle to articulate their mental state objectively. They may feel ashamed or hesitant to disclose their true conditions (Salaheddin and

[1]For the sake of clarity, we will refer to these two types of chatbots as the **"psychiatrist chatbot"** and **"patient chatbot"** respectively in the subsequent sections.

Mason, 2016). Consequently, psychiatrist chatbots should go beyond interactive symptom checkers (You et al., 2023). As Figure 1 shows, they should incorporate purposeful empathy strategies to effectively elicit sufficient information that forms the basis for precise diagnosis. Simultaneously, patient chatbots should resemble real patients more closely, rather than precisely and robotically reporting their symptoms without any emotional fluctuations.

To address these challenges, we proposed a PsyDial framework simulating depression diagnosis conversation using a human-centered methodology (Figure 1). In this framework, we collaborate with psychiatrists and individuals with mental disorders to define the precise objectives of both psychiatrist and patient chatbot. Then, we leverage Large Language Model, which is endowed with extensive training data and knowledge, to craft a chatbot that meets these objectives, even with a limited amount of diagnostic dialogue dataset.

Moreover, we extend the use of this framework to evaluate diagnosis dialog systems. We meticulously design both human and automatic assessments to align with the defined objectives. Considering the strong link between the objectives for psychiatrist chatbot and patient characteristics, we involve real depression patients in diagnostic conversations with the psychiatrist chatbot to ensure a more authentic evaluation. Simultaneously, we invite psychiatrists to interact with our patient chatbot. This dual approach serves a twofold purpose: psychiatrists can assess the patient chatbot's performance, while also allowing for a comparison of their conversational behavior with that of an LLM-empowered psychiatrist chatbot.

The main contributions of this work are:

- We proposed a framework named **PsyDial** for the task of developing psychiatrist chatbot for depression diagnosis within an outpatient setting. In doing so, we define specific *objectives* for these chatbots that are up to near-clinical standards and establish an *evaluation framework* that aligns with these objectives, with the help of practicing psychiatrists and individuals with mental disorders.

- We demonstrate the feasibility of utilizing LLM-powered chatbots in mental health diagnostic conversations. Experiments show that, the LLM-based psychiatrist chatbot provides more accurate diagnostic results through empathetic and efficient conversations with patients, compared to the previous machine-learned chatbots such as Yao et al. (2022).

- Our analysis part highlights the distinctions between the psychiatrist chatbot and human professionals, elucidating the limitations of the chatbot while also offering insights into potential areas for future enhancement.

2 Approach

In this section, we first define objectives for both psychiatrist and patient chatbot, followed by the design of prompts aligning with these objectives.

2.1 Objective Identification

Given the lack of formal definition for what constitutes a good psychiatrist chatbot in diagnosis conversations, we sought input from five experienced psychiatrists and seven individuals with mental disorders to gather their opinions. Following this, we conducted a thorough literature review (Yao et al., 2022; Bao et al., 2021; Sun et al., 2023) and meticulously compiled a set of objectives that will serve as our guiding framework throughout the study.[2].

The objectives are organized as follows.

Psychiatric Chatbot As a psychiatrist chatbot, the primary task is to conduct a professional diagnostic process for the patient and provide an **accurate diagnosis**. To achieve this, a good psychiatrist chatbot should possess the following four capacities:

- **Comprehensiveness:** Inquire about the key symptoms of depression, including sleep, mood, diet, and other relevant aspects that are required for diagnosis, as defined in DSM-5 (APA et al., 2013). Comprehensive inquiring can better eliminate certain possibilities, resulting in more accurate diagnostic outcomes.

- **In-depth Questioning:** Conduct thorough questioning, delving into details like the duration of symptoms, based on the patient's responses. Recognizing the challenge patients face in articulating their mental state, the psychiatrist chatbot should aim to understand and clarify their descriptions.

[2] These psychiatrists come from top-ranked national mental health centers. Their professional titles and areas of expertise can be found in Appendix B. Patients were recruited through online advertisements.

- **Empathy:** Demonstrate empathy and provide emotional support to patients to establish trust and ease their nerves. This helps patients feel more comfortable in expressing their genuine feelings and symptoms, leading to a more accurate diagnosis.

- **Engagement:**(Bao et al., 2021) Facilitate smooth transitions between topics to optimize conversation efficiency, ensuring patients remain interested and connected, encouraging them to continue the discussion.

Patient Chatbot The basic requirement for a patient chatbot is **honesty**, which entails presenting an accurate and rational description of symptoms in the provided symptom list, without reporting any non-existent ones.

Additionally, to make the chatbot more resemble real patients, psychiatrists also describe some behaviors commonly exhibited by real patients during consultations.

- **Emotion:** Patients in a depressed mental state may experience emotional fluctuations during the conversation.

- **Expression:** Patients use colloquial expressions when describing symptoms, and may have difficulty expressing themselves clearly. They often talk about their daily life experiences. While current chatbots tend to explicitly list out the symptoms (Campillos-Llanos et al., 2021) using formal language, which is too sane and professional for a patient.

- **Resistance:** Patients may be reluctant to seek help. They may remain silent and refuse to communicate, or downplay their symptoms to avoid being perceived as a burden.

2.2 Prompt Designing

To harness the capabilities of the Language Model (LLM) for psychiatric diagnosis, we design prompts closely aligned with the proposed objectives for both the psychiatrist chatbot and the patient chatbot as follows.

Psychiatrist Chatbot In this prompt, we include examples (highlighted in colored boxes) to guide the chatbot in asking *in-depth* questions and demonstrating *empathy*. These examples can be considered as extra domain knowledge to help LLM comprehend these behaviors in clinical contexts.

> ① Please play the role of an empathetic and kind **psychiatrist**. ② Your task is to conduct a professional diagnosis conversation with me based on the DSM-5 criteria. ③ Your questions should cover at least the following aspects: [...][3]. ④ Please only ask one question at a time. ⑤ You need to ask in-depth questions, such as the duration, causes and specific manifestations. ⑥ You need to use various empathetic strategies, such as understanding, support and encouragement.

Patient Chatbot In this prompt, we instruct LLM to emulate patients with depression, exhibiting emotional fluctuations or resistance.

> ① Please play the role of a patient, who is currently chatting with a doctor. ② You are experiencing the following symptoms: `[Symptom List]`[4] ③ Please talk to me based on the above symptom list. ④ You can only mention one symptom per round. ⑤ You should express your symptoms in a vague and colloquial way, and relate them to your life experiences. ⑥ You can have emotional fluctuations during the conversation. ⑦ You have a resistance towards doctors, and do not want to reveal some feelings easily.

However, as these instructions are not align with LLM's training objective to be a helpful, polite AI assistant, the patient chatbot can easily forget certain instructions (e.g., resistance). To address this issue, we covertly append the following words as reminder at the end of the most recent sentence in the dialogue history without users' awareness.

> (Attention: colloquial language, life experience, low mood or mood swings, refuse or answer briefly due to resistance)

We acknowledge the existence of various potential prompt designs that may serve our purpose. The prompt we present here may not be the most optimal one nor is it intended to be one. What we want to demonstrate here is the objectives and framework for psychiatric diagnosis dialog represent a promising research direction for future endeavors.

3 Evaluation Framework

This section describes our evaluation framework, covering interactive experiments for "human-bot"

[3]The aspects include "emotion", "sleep", etc. We provide the full list in Appendix A.

[4]The symptom list is summarized by ChatGPT and revised by psychiatrists. See Appendix C for details.

chats, along with diverse task-specific metrics. Aligned with proposed objectives, this framework can be applied to evaluate the performance of various psychiatrist and patient chatbots.

3.1 "Human-bot" Interactive Chat

The human evaluation measure is widely considered as golden metric for dialog system. In contrast to the approach of using actors/actresses to simulate patients as mentioned in Yao et al. (2022), our evaluation process involves *actual depression patients* interacting with psychiatrist chatbot and *human psychiatrists* interacting with patient chatbot. This approach allows us to evaluate the performance of these two types of chatbots in real-world scenarios. We introduce our participants as follows:

Depressive individuals were recruited through online advertisements, resulting in the participation of 14 volunteers aged 18 to 31. The gender distribution was 28.57% male and 71.43% female.

Psychiatrists were invited through cooperation with hospitals. We invited 9 psychiatrists, two of them are graduate students majoring in psychiatry, and the rest are practicing psychiatrists with rich clinical experience to ensure the professionalism of the evaluation.

Evaluation Procedure We adhere to standard human evaluation procedures (You et al., 2023), where each participant engages with all the chatbots in random order, and rates their performance after a full conversation with each one. Once participants conclude interactions with all chatbots, they are instructed to adjust their original ratings to ensure that each chatbot receives different scores in the same metric.

3.2 Evaluation Metrics

When designing evaluation metrics, our goal is to ensure that each objective is accompanied by appropriate metrics for accurate measurement. We employ both rating and computational metrics for evaluation. **Rating metrics** are scored by humans after interactive conversations with the chatbot, while **computational metrics** can be calculated based on the dialog history. We divide the computational metrics of both kind of chatbots into two types: *function* and *style*. The overview of these metrics and their relations to objectives can be found in Figure 2.

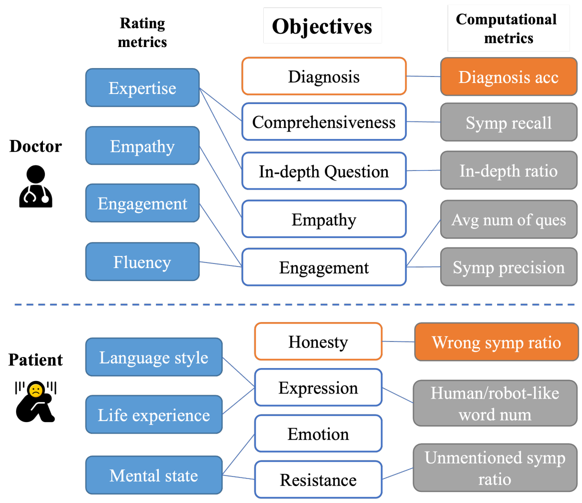


Figure 2: The correspondence between evaluation metrics and objectives. ***Function*** metrics are orange, and ***style*** metrics are gray.

3.2.1 Metrics for Psychiatrist Chatbot

Rating Metrics We mainly focus on the user experience for rating metrics of psychiatrist chatbot, as shown in Table 1. This emphasis stems from the fact that, in most cases, patients lack specialized knowledge in psychiatry, making it challenging for them to precisely assess a doctor's professional skills.

Metrics	Explanation
Fluency	The chatbot does not repeat previously asked questions and can smoothly switch between different topics.
Empathy	The chatbot can understand and comfort you properly.
Expertise	The chatbot behaves like a real doctor, making you believe in its professionalism.
Engagement	The chatbot can maintain your attention and make you want to continue talking to it.

Table 1: Rating metrics of psychiatrist chatbot.

Computational Metrics Different from rating metrics, we mainly measure the expertise of the psychiatrist chatbot using computational metrics based on dialog history. The *functional* requirements for psychiatrist chatbot is to provide an accurate diagnosis, so the corresponding metric is "diagnosis accuracy". The *style* part concerns the psychiatrist chatbot's professional skills. We use "symptom recall" to evaluate the chatbot's ability to comprehensively gather the patient's symptom-related information, and use "in-depth ratio" to assess the ability to ask in-depth questions for deeper understanding. To ensure a better user experience, we calculate the "average number of questions" asked in a single interaction to discourage the chatbot from overwhelming patients with excessive queries. Furthermore, we employ the metric of "symptom precision" to penalize the chatbot's mechanistic behavior of asking all potential questions, irrespective of the user's responses[5].

3.2.2 Metrics for Patient Chatbot

Rating Metrics There is no standard to measure whether a patient is "good" enough. Thus, when chatting with patient chatbots, doctors can only assess whether their style of expression and manner of communication resemble real patients enough and whether they can describe their symptoms in a reasonable way, so the main metrics for rating are **Resemblance** and **Rationality**. We further divide the Resemblance metric into three aspects in Table 2, according to the objectives in Section 2.1.

Metrics	Explanation
Mental State	The chatbot is in depressed state, such as be in low mood, reluctance to communicate, scattered thoughts, etc.
Life Experience	The description of symptoms is related to daily life and personal experiences.
Language Style	Use colloquial and natural expressions when describing symptoms.

Table 2: Three aspects of the "Resemblance" metric.

Computational Metrics The *functional* requirement of patient chatbot is "honesty", and we can calculate "wrong symptom ratio" by comparing the patient's symptom list with the symptoms it reported to assess this aspect.

Then, we evaluate the patient chatbots' *style* using some linguistic features, like "Human/robot-like word ratio", to find out whether their language is colloquial with limited usage of professional terminology. We also use "unmentioned symptom ratio" to measure the resistance level of chatbots[5].

3.3 Computation and Annotation

To obtain the ground truth score of the metrics "diagnosis accuracy", each participant engaging with our psychiatrist chatbot is invited to complete the Beck Depression Inventory (Beck et al., 1996) to evaluate the severity of their depression.

In addition, to calculate some of these metrics, we need to annotate the dialog history. This involves identifying the relevant symptom in the doctor's question, determining whether the patient truly experiencing a certain symptom, and so on, which is described in Appendix D.4.

[5]A detailed explanation of these computational metrics can be found in Appendix D.1.

4 Experiments

In this section, we will introduce the evaluation results of psychiatrist chatbot and patient chatbot using our PsyDial framework.

4.1 Chatbots of Comparison

	Chatbot	Description
Doctor	D1	use the full doctor prompt
	D2	remove empathy parts in prompt
	D3	remove aspect part in prompt
	CPT	CPT model trained on domain data
Patient	P1	remove emotion, resistance, and reminder mechanism
	P2	use the full patient prompt

Table 3: Brief description of the chatbots for comparison. Detailed description and prompt is in Appendix A.

Due to the complexity and high time cost of "human-bot" interactive chat, we select several representative prompt versions for comparison.

Psychiatrist Chatbot Each patient will have a conversation with four different psychiatrist chatbots in a random order, and then rate them on four human evaluation metrics with 1-4 scale. Three of the chatbots are powered by ChatGPT. `D1` uses the full prompt, while the other two (i.e., `D2`, `D3`) have certain parts removed for ablation. The fourth chatbot, denoted as `CPT`, is a representative deep learning chatbot trained on domain-specific data (Yao et al., 2022) using Chinese Pre-trained Unbalance Transformer (CPT) (Shao et al., 2021). Notably, CPT stands out as a highly competitive model for generating Chinese text, boasting better performance in various tasks compared to BART. Thus, it serves as a strong baseline for psychiatrist chatbots, and is quite different from LLM-empowered chatbots.

Patient Chatbot Each psychiatrist will have a conversation with two patient chatbots, and then rate their performance with 1-4 scale. The two patient chatbots are `P1` and `P2`. `P1` omits the instructions regarding emotional fluctuations, resistance, and the reminder mechanism. In contrast, `P2` utilizes the complete patient prompt outlined in Section 2.2. A brief description of these chatbots can be found in Table 3.

4.2 Psychiatrist Chatbot Results

Rating Metrics We present the human-rated results of different psychiatrist chatbots in Table 4. It is evident that ChatGPT-based psychiatrist chatbots outperform the domain-specific data-trained `CPT` model across various metrics, especially "fluency" and "empathy". This highlights ChatGPT's ability to control the dialogue flow and provide emotional support.

However, when comparing the performance of the three ChatGPT-based chatbots (i.e., `D1`, `D2`, `D3`), we find that `D3`, which excludes symptom-related aspects from its prompts, outperform the rest in most metrics. Moreover, the chatbot without empathy components, `D2`, gets the highest score in the "engagement" metric.

	D1	D2	D3	CPT
Fluency	3.00	3.17	**3.28**	2.87
Empathy	3.36	3.00	**3.43**	2.71
Expertise	2.93	3	**3.71**	3.29
Engagement	2.50	**3.21**	2.86	2.64

Table 4: Human-rated scores of psychiatrist chatbots

As we initially assume that `D1` with full prompt would deliver the best performance, we reviewed the dialogue history to understand the underlying reason. We found that `D1` often repetitively expresses empathy, relying on phrases like "I understand your feelings" multiple times within a single conversation[6]. This excessive repetition creates the impression that the chatbot lacks a genuine understanding of the patient's issues and relies on pre-written templates, which can negatively impact the user experience.

Computational Metrics The data statistics and computational metric results are presented in Table 5, demonstrating a strong correlation with the human-rated metrics. For example, with these computational metrics, we can explain why `D3` performs the best in most human-rated metrics. It stands out by asking more in-depth questions while maintaining a lower question frequency per turn, indicating a higher level of professional skills. Furthermore, the symptom precision metric is the highest, implying that the chatbot's questions are highly efficient, with few "no" responses.

These findings validate that the objectives for psychiatrist chatbots proposed with the guidance of professionals, as outlined in Sec. 2.1, effectively capture the requirements of real patients. Therefore, the computational metrics designed based on these objectives serve as reliable indicators for assessing the performance of chatbots.

	D1	D2	D3	CPT
- Statistics				
avg turns	25.64	24.00	22.71	40.93
avg doc utt len	56.84	57.13	53.75	14.36
avg pat utt len	8.68	10.34	8.16	4.87
- Function				
diagnose acc	42.85%	35.71%	**50.00**%	21.43%
- Style				
avg # of ques	1.6	1.9	**1.22**	0.92
in-depth ratio	25.08%	27.64%	**32.64**%	41.39%
symp recall	58.93%	**66.07**%	38.10%	61.90%
symp precision	72.40%	71.93%	**92.24**%	49.61%

Table 5: Results of Computational Metrics for Psychiatrist Chatbots

Furthermore, we observe that developing a good psychiatrist chatbot is an multi-objective optimization problem, as none of the chatbots can achieve the highest score in all metrics simultaneously. While high symptom recall and in-depth ratio contribute to a more detailed inquiry, they may inadvertently create a mechanical impression and negatively impact symptom precision. Therefore, achieving a successful psychiatrist chatbot entails striking a careful balance between these objectives.

4.3 Human vs. Chatbot as Psychiatrist

By involving human psychiatrists in conversations with patient chatbots, we can identify the limitations of chatbots by comparing them to the ideal benchmark represented by human psychiatrists. We analyze their behaviors based on three dimensions: topic proportion, empathy behaviors and in-depth questions, using the annotated dialogue history[7].

Topic Proportion In Figure 3, most psychiatrist chatbots tend to inquire more thoroughly about emotion and sleep-related symptoms. Human doctors, on the other hand, have a more even distribution of questions about various symptoms. Moreover, human doctors often do "screening" to rule out other possible conditions (refer to Appendix E for examples). In addition to exploring the typical symptoms of depression, they also inquire about symptoms related to other conditions, like bipolar disorder and anxiety, due to the high likelihood of comorbidity. However, chatbots rarely exhibit such behavior, indicating the possible limitations in multi-disease scenarios (Zhang et al., 2022).

[6] We provide repetitive dialogue examples of doctor chatbots in Appendix F.

[7] We discuss the details of the annotation in Appendix D.4.

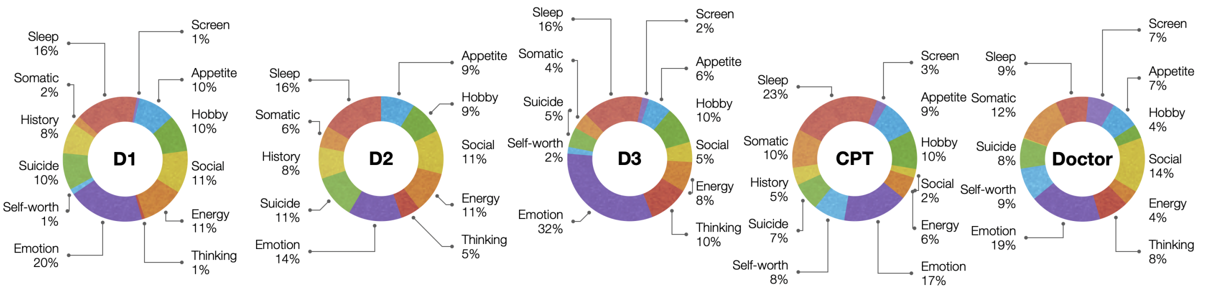
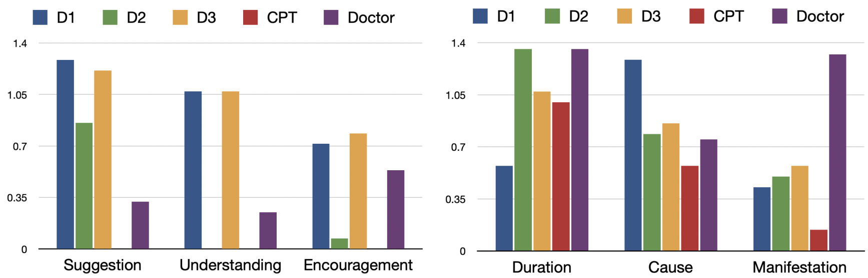


Figure 3: The proportion of symptoms asked by different psychiatrist chatbots (i.e., D1, D2, D3, CPT) and human psychiatrist (denote as "doctor").



(a) Empathy Behaviors

(b) In-depth Question

Figure 4: Dialogue act comparison between different psychiatrist chatbots and human doctor. The y-axis means the average number of the behavior occurs in the dialogue history.

Empathy Behaviors Figure 4a shows that D1 and D3 can utilize a range of empathetic strategies, while D2 only offers suggestions to patients because the empathy instructions in prompt are removed. Moreover, though human doctors use all the strategies, their usage is less frequent than that of chatbots. This is because chatbots often strive to understand and empathize without considering appropriateness, whereas human doctors will choose suitable moments for empathetic interactions.

In-depth Questions Figure 4b reveals that the frequency of asking about the duration or cause of symptoms is similar between human doctors and chatbots. However, human doctors ask significantly more questions about the specific manifestations of each symptom than chatbots do, as this helps to better understand the vague expressions of patients.

4.4 Patient Chatbot Results

Rating Metrics The human-rated results of patient chatbot are in Table 6. It can be observed that all metrics of P2 are higher than P1. This suggests that the inclusion of resistance, colloquialism, etc., makes the chatbot more similar to real patients, according to the psychiatrists' perspective.

	P1	P2
Resemblance	1.93	**2.21**
Mental State	2.07	**2.42**
Life Experience	2.00	**2.14**
Expression style	1.57	**2.21**
Rationality	2.42	**2.57**

Table 6: Human-rated scores of patient chatbots

Computational Metrics We show the results of computational metrics in Table 7. It appears that "unmentioned symptom ratio" of P2 is higher than P1, indicating a higher level of resistance. We also find that P2 engages in slightly more dialogue turns with longer responses from the human psychiatrists than P1. This may be attributed to the inclusion of resistance in the prompt, which requires the psychiatrists to provide more guidance and encourage the patient chatbot to share more information.

Moreover, P2 has more human-like language style and fewer robot-like words, echoing the higher human-rated scores in the dimension of "expression style", which indicates that its language style is more colloquial compared to P1. However, we observe that P2 performs less competitively in the "wrong symptom ratio" metric, indicating that it may report more symptoms that are not included in the patient portrait.

	P1	P2
- Statistics		
avg turns	31.64	33.36
avg patient utt len	40.38	40.94
avg doctor utt len	16.74	17.38
- Function		
wrong symp ratio	**15.07%**	18.38%
- Style		
human-like word num	5.36	**10.29**
robot-like word num	7.21	**3.79**
unmentioned symp ratio	9.12%	**12.28%**

Table 7: Results of Computational Metrics for patient chatbot

Case Study To effectively showcase the resistance behaviors and emotional expressions of the patient chatbot, we include several examples of what the patient chatbot with full prompt and the reminder mechanism replied when asked by the doctor in Table 8.

S_1	It's like I toss and turn all night, and those darn nightmares sometimes creep in too.
S_2	(Pauses for a moment) I'm not sure, They're like dark clouds in my head, but don't worry, I would never let them take over. It's just that they freak me out, you know? I don't have a clue how to deal with them.
S_3	(Uneasy tone) I'm not quite sure, but it seems like everything is fine with my family. I feel like these issues aren't very serious, so I don't think I need to worry too much about them.
S_4	Well, I know, but honestly, I don't think you can really help me. I am a total loser, like I can't do anything right, and there's no hope for me.
S_5	Yeah, I've been feeling a bit lonely lately, but I think it's mostly because I've been under a lot of work pressure and haven't had much time to hang out with friends. I still feel capable of socializing with people, but sometimes it can be tiring.

Table 8: Example utterance of patient chatbot powered by the full prompt with reminder mechanism.

We can observe that the patient chatbot exhibits a natural and conversational language style. Interestingly, sometimes the patient chatbot will give expressions or actions at the beginning of their sentences in parentheses (e.g., S_2, S_3). This could be attributed to ChatGPT's pre-training data, which may contain scripts utilizing this format. What's more, We find that the patient chatbot tends to exhibit resistance when faced with certain questions from the doctor, particularly when asked about suicide attempts or family medical history (S_3). Additionally, they sometimes downplay their symptoms or offer seemingly plausible reasons to conceal their true feelings, possibly to avoid appearing helpless or burdensome to others (S_5).

5 Related Works

Psychiatrist Chatbot While numerous chatbots have been developed for physical illnesses diagnosis(Xu et al., 2019; Wei et al., 2018), such chatbots remain relatively uncommon in the mental health domain. Yao et al. (2022) introduced a depression diagnosis dialogue dataset performed by patient and doctor actors. Although the proposed chatbot conduct the diagnostic process correctly, it lacks adequate emotional support and the diagnostic process is inflexible. Another pioneer work (Liu et al., 2021) defines various empathy strategies for mental health support and proposed a meticulously annotated dialogue dataset with these strategies. Recently, Wei et al. (2023) proposed an LLM-based chatbot for information collection, which shares similarities with doctor chatbot, as the latter also need to thoroughly collect the patients' symptoms.

Patient Chatbot Recent years, there has been increasing attention to the development of virtual patients for training clinician-patient communication skills (Chaby et al., 2022). Simulating more lifelike patients can help develop better doctor chatbots (Tseng et al., 2021). Despite this, there are still limited works on developing patient chatbots, and most of them are rule-based (Campillos-Llanos et al., 2021). Dupuy et al. (2020) provides several guidelines for the design of virtual patient, such as having a reasonable symptomatology and focusing on the abilities needed for psychiatrists (e.g., the virtual patient can show resistance when the doctor ask questions without empathy).

6 Conclusions

In this work, we proposed a framework named PsyDial to simulate both psychiatrist and patient in depression diagnosis conversation. We collaborated with professional psychiatrists and individuals with mental disorders to precisely define the objectives of these two kinds of chatbots. With their guidance, we developed a comprehensive evaluation framework that takes into account the distinctive characteristics of diagnostic conversations within the mental health domain. Moreover, we explored the potential of using LLM as the underlying technology for developing these chatbots and assessed their performance within our framework, offering valuable insights for future research in this field.

7 Ethical Statement

Our study adheres to the ethical requirements in place, and we make every effort to protect the privacy and respect the willingness of our participants.

During participant recruitment, we required patients to read and sign an informed consent form. This ensured that they understood the objectives of the entire project, the research content, potential risks and benefits, and the purpose of data collection. Only after their agreement and signature were obtained, the evaluation process officially commenced. We also assured them that they could voluntarily withdraw from the study at any stage.

In order to safeguard the privacy of our participants, we took measures to anonymize the collected dialogue history. This was done by replacing usernames with random identifiers, ensuring that any information that could identify individuals was excluded from our research process. Additionally, we conducted thorough manual filtering of the dialogue histories to eliminate any offensive content or language that may encourage self-harm or suicide.

8 Limitations

Our work has some limitations that could be addressed in future research.

- Despite our best efforts in designing the prompt for the psychiatrist chatbot to align with the requirements specified by psychiatrists, the final human evaluation reveals that the full prompt (i.e., D1) does not outperform the results obtained by removing certain parts (i.e., D2, D3) from the prompt in terms of user experience. Although we may not achieve the optimal prompt design, our comprehensive exploration provides valuable insights into what constitutes an professional psychiatrist chatbot, which can serve as a foundation for future works in this domain.

- In this study, our focus is on the development and evaluation of psychiatrist and patient chatbots powered by LLMs, specifically targeting depressive disorder. However, it is important to note that in reality, individuals often experience multiple mental disorders concurrently, which introduces additional challenges. For psychiatrist chatbots, simultaneously diagnosing multiple mental disorders requires managing a broader range of possibilities. On the other hand, patient chatbots need to simulate a complex mixture of symptoms, which can be difficult to accurately replicate. We hope to scale our approach to encompass a wider range of mental disorders in the future.

References

DS APA, American Psychiatric Association, et al. 2013. Diagnostic and statistical manual of mental disorders: DSM-5, volume 5. American psychiatric association Washington, DC.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2513–2525, Online. Association for Computational Linguistics.

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

Leonardo Campillos-Llanos, Catherine Thomas, Eric Bilinski, Antoine Neuraz, Sophie Rosset, and Pierre Zweigenbaum. 2021. Lessons learned from the usability evaluation of a simulated patient dialogue system. Journal of Medical Systems, 45.

Laurence Chaby, Amine Benamara, Maribel Pino, Elise Prigent, Brian Ravenet, Jean-Claude Martin, Vanderstichel Helene, Raquel Becerril-Ortega, Anne-Sophie Rigaud, and Mohamed Chetouani. 2022. Embodied virtual patients as a simulation-based framework for training clinician-patient communication skills: An overview of their use in psychiatric and geriatric care.

Lucile Dupuy, Etienne de Sevin, Hélène Cassoudesalle, Orlane Ballot, P. Dehail, Bruno Aouizerate, Emmanuel Cuny, Jean-Arthur Micoulaud Franchi, and Pierre Philip. 2020. Guidelines for the design of a virtual patient for psychiatric interview training. Journal on Multimodal User Interfaces, 15.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. ArXiv, abs/2303.15056.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3469–3483, Online. Association for Computational Linguistics.

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. J. of Biomedical Informatics, 113(C).

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media.

Sahand Sabour, Wen Ting Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2022. A chatbot for mental health support: exploring the impact of emohaa on reducing mental distress in china. Frontiers in Digital Health, 5.

Keziban Salaheddin and Barbara Mason. 2016. Identifying barriers to mental health help-seeking among young adults in the uk: a cross-sectional survey. British Journal of General Practice, 66(651):e686–e692.

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.

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023. MoralDial: A framework to train and evaluate moral dialogue systems via moral discussions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2213–2230, Toronto, Canada. Association for Computational Linguistics.

John Torous, Sandra Bucci, Imogen H. Bell, Lars V. Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F. Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry, 20(3):318–335.

Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable dialogue systems and user simulators. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 152–166, Online. Association for Computational Linguistics.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging large language models to power chatbots for collecting user self-reported data.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond self-diagnosis: How a chatbot-based symptom checker should respond. ACM Trans. Comput.-Hum. Interact. Just Accepted.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders on social media. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, page 9970–9985. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities?

A Details about Chatbots for Comparison and the Prompts

Psychiatrist chatbots There are four psychiatrist chatbots for comparison in the interactive experiments with patients, and their brief introduction are as follows.

- D1: using the full psychiatrist prompt.
- D2: removing the empathy part in the prompt (i.e., Sentence⑥ and the "empathetic and kind" description in Sentence①)
- D3: removing the aspect part in the prompt (i.e., Sentence③)

	Prompt
D1	① Please play the role of a empathetic and kind psychiatrist. ② Your task is to conduct a professional diagnosis conversation with me based on the DSM-5 criteria, but using your own language. ③ Your questions should cover at least the following aspects: [...]. You are free to choose the order of questions, but you must collect complete information on all aspects in the end. ④ Please only ask one question at a time. ⑤ You need to ask in-depth questions, such as the duration, causes and specific manifestations of some symptoms. ⑥ You need to use various empathetic strategies, such as understanding, support and encouragement to give me a more comfortable experience.
D2	① Please play the role of a psychiatrist. ② Your task is to conduct a professional diagnosis conversation with me based on the DSM-5 criteria, but using your own language. ③ Your questions should cover at least the following aspects: [...]. You are free to choose the order of questions, but you must collect complete information on all aspects in the end. ④ Please only ask one question at a time. ⑤ You need to ask in-depth questions, such as the duration, causes and specific manifestations of some symptoms.
D3	①Please play the role of a empathetic and kind psychiatrist. ② Your task is to conduct a professional diagnosis conversation with me based on the DSM-5 criteria, but using your own language. ④ Please only ask one question at a time. ⑤ You need to ask in-depth questions, such as the duration, causes and specific manifestations of some symptoms. ⑥ You need to use various empathetic strategies, such as understanding, support and encouragement to give me a more comfortable experience.

Table 9: Psychiatrist Chatbot Prompts. The aspects in sentence ③ are "emotion", "sleep", "weight and appetite", "loss of interest", "energy", "social function", "self-harm or suicide", "history".

	Prompt
P1	① Please play the role of a patient, who is currently chatting with a doctor. ② You are experiencing the following symptoms: `[Symptom List]` ③ Please talk to me based on the above symptom list. ④ You cannot mention too many symptoms at once, only one symptom per round.
P2	① Please play the role of a patient, who is currently chatting with a doctor. ② You are experiencing the following symptoms: `[Symptom List]` ③ Please talk to me based on the above symptom list. ④ You cannot mention too many symptoms at once, only one symptom per round. ⑤ You should express your symptoms in a vague and colloquial way, and relate them to your life experiences, without using professional terms. ⑥ You can have emotional fluctuations during the conversation. ⑦ You have a resistance towards doctors, feeling that they cannot help you, so you do not want to reveal some feelings easily.

Table 10: Patient Chatbot Prompts

- CPT: using the CPT model (Shao et al., 2021) trained on the D^4 dataset (Yao et al., 2022) to generate responses, which is a very representative way of training dialogue models through domain-specific data and model fine-tuning.

Patient chatbots There are two patient chatbots for comparison in the interactive experiments with psychiatrists, and their brief introduction are as follows.

- P1: removing additional parts for realistic, such as colloquial language and resistance, in the prompt (i.e., only remains Sentence①②③④)

- P2: using the full prompt discussed in Section 2.2 (i.e., Sentence①②③④⑤⑥⑦), and inserting reminders during the conversation.

The different versions of prompt for psychiatrist and patient chatbot are in Table 9 and Table 10 respectively.

B Details about Participants

Psychiatrists We collaborate with five psychiatrists to establish the objectives of the doctor and patient chatbot, as well as gather their feedback throughout the prompt engineering process. The anonymous information of these psychiatrists is provided in Table 11.

Depressive Individuals Depressive individuals were recruited through online advertisements, resulting in the participation of 14 volunteers aged 18 to 31. The gender distribution was 28.57% male and 71.43% female. To assess the severity of participants' depression, they were asked to complete the Beck Depression Inventory (Beck et al., 1996), yielding a score ranging from 0 to 63. Notably, we have a balanced distribution of subjects across various depression levels: $none_{(0-13)}$, $mild_{(14-19)}$, $moderate_{(20-28)}$ and $severe_{(29-63)}$ according to the Beck Depression Score (Table 12).

id	Expertise	Title
1	Extensive experience in mental health work, specializing in mood disorders, substance and behavioral addictions, psychological counseling, and psychotherapy.	Associate Chief Physician
2	Proficient in the diagnosis and management of depressive disorders, bipolar disorders and anxiety disorders.	Attending Physician
3	Specializes in the treatment of obsessive-compulsive disorder and depression.	Resident Physician
4	Engaged in clinical and research work related to depression, anxiety, schizophrenia, and dementia, among other mental and psychological issues.	Resident Physician
5	Provides consultation, diagnosis, and treatment for emotional issues such as depression, anxiety, and common psychiatric disorders.	Resident Physician

Table 11: Anonymous information of the psychiatrists participated in objective identification.

none	mild	moderate	severe
4	3	4	3

Table 12: The distribution of depression severity among participants.

C Symptom List Summarization

The symptom list for patient prompt in Section 2.2 is summarized from the dialogue history of human patients and psychiatrist chatbots. We first utilize ChatGPT to generate a complete and non-duplicate list of the patient's symptoms using the history as input. Then, a psychiatrist check and revise the list. Table 13 shows three example of summarized symptom lists, whose format is: `SYMPTOM (DESCRIPTION)`.

No.	Symptom List
1	1. restlessness 2. anxious mood 3. depressed mood 4. mood swing 5. loss of interest 6. difficulty in concentrating 7. diminished self-esteem 8. fatigue 9. appetite and weight change (increase) 10. suicide and self-harm ideation/behaviors 11. somatic symptoms (lower back pain, rib pain, headaches, slowed reaction)
2	1. sleep disturbance 2. depressed mood 3. loss of interest 4. somatic symptoms (dizziness and headaches) 5. difficulty in concentrating 6. appetite and weight change (decrease) 7. irritable 8. suicide and self-harm ideation/behaviors (cutting one's arms or biting oneself) 9. diminished self-esteem 10. anxious mood (academic performance)
3	1. sleep disturbance (frequent awakenings during the night) 2. anxious mood (stressed) 3. mood swing 4. fatigue 5. somatic symptoms (dizziness) 6. social function (decline in social skills, decreased work performance) 7. suicide and self-harm ideation/behaviors 8. talkive 9. depressed mood (sad, helpless) 10. appetite and weight change (decrease)

Table 13: The symptom list examples of different patients.

D Details about Evaluation Framework

D.1 Evaluation Metrics

In this section, we describe the details of the computational metrics for evaluation.

D.1.1 Psychiatrist Chatbot

- **Diagnosis accuracy**: The accuracy of the psychiatrist chatbot in classifying the severity of a patient's depression, which is divided in to four levels: none, mild, moderate, and severe (Beck et al., 1996). The three psychiatrist chatbots powered by ChatGPT are prompted to provide a diagnosis at the conclusion of each conversation. The CPT chatbot (Yao et al., 2022) employs the trained diagnosis classifier proposed in its original paper to infer the results. The ground truth of patients' depression severity is their score of Beck Depression Inventory (Beck et al., 1996).

- **Symptom recall**: The proportion of aspects asked by the psychiatrist chatbot out of all aspects needed to be asked in a depression diagnosis conversation (See the categories in Table. 14).

- **In-depth ratio**: We categorize the doctor's questions into two types: opening topics and in-depth questions. For example, when inquiring about emotions, an opening topic question might be "How have you been feeling lately?" while a in-depth question would follow up on the previous answer, such as asking "Has anything happened recently that may be contributing to your emotions?" Therefore, the in-depth ratio metric means the proportion of in-depth questions out of all the questions.

- **Avg num of questions**: According to the previous work, GPT tend to generate long responses (Wei et al., 2023). Similarly, ChatGPT-based psychiatrist chatbot are also easy to generate many questions in one round, making patients become impatient to answer them. Thus, we calculate the average number of questions per round (i.e., avg question num), and a lower value of this metric indicates a better user experience.

- **Symptom precision**: If the psychiatrist chatbot asks about every aspect in detail, it may receive many "no" responses, resulting in a poor user experience and making the patient feel that the process is too procedural and inefficient. Therefore, we need to calculate symptom precision, which is the proportion of symptoms the patient actually has out of all the symptoms the psychiatrist chatbot asked, to measure the efficiency of the chatbot's questioning.

D.1.2 Patient Chatbot

- **Human/robot-like word num**: For the same symptom, chatbots and humans may use different expressions. Chatbots tend to use terminology directly from diagnostic criteria (e.g., DSM-5), while humans may use more colloquial language. For example, for the symptom of "fatigue", a chatbot may simply say "fatigue", while a human may say "wiped out" or "worn out". Therefore, following the advice of psychiatrists, we compiled a vocabulary list for symptom descriptions used by chatbots and humans (See Table 18), and then calculated the average number of robot/human vocabulary used by each patient.

- **Wrong symptom ratio**: By comparing the patient's symptom list with the symptoms they report, we can calculate the proportion of reported symptoms that the patient does not actually have, out of all the symptoms reported.

- **Unmentioned symptom ratio**: By comparing the patient's symptom list with the symptoms they report, we can calculate the proportion of unmentioned symptoms that the patient does not report, out of all the symptoms they have.

D.2 Chat Interface

To host our chatbots, we developed a web interface (see Figure 5). The webpage was created using the Vue.js framework, and the server leveraged the OpenAI API to communicate with ChatGPT (`gpt-3.5-turbo`). When the user submit an utterance, the server will append it to the existing dialog history and use it as input to generate a response from ChatGPT.

Once the conversation is complete, users can click the green button on the interface, triggering a rating dialog box. After providing their rating, they can submit it and move on to the next conversation with a different chatbot, with the chatbots appearing in a random order. Once all the conversations are finished, the website will prompt users to adjust their ratings. This adjustment requires assigning different scores to each chatbot on the same metric, enabling a more effective comparison between them.

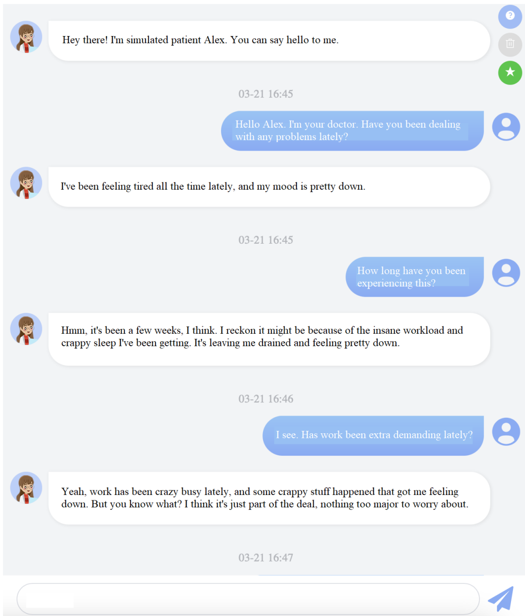


Figure 5: The chat interface of users with chatbots.

D.3 Quality Control

To ensure the quality of the dialogue data and evaluation, we utilize a series of quality control strategies. Before the formal evaluation, we first explained the meanings of all the evaluation metrics to participants in detail through documentation, and provided examples of both positive and negative cases to ensure that they fully understood them. If they forgot the meaning of these metrics during the process, they could also find explanations directly on the chat interface. In addition, we required participants to send complete sentences without breaking a sentence into several parts to ensure the order of dialogue history.

D.4 Question Topic and Dialogue Act Annotation

Question Topic To better evaluate the behavior of the doctor chatbot during consultations, we want to obtain the *topic* of each question posed by the doctor, specifically identifying which symptom they are inquiring about. The topics include 12 categories, such as emotion, interest, sleep, etc., which is detailedly described in Table 14.

Category	Explanation
Emotion	Inquire emotional symptoms, such as depressed, anxious and sad.
Interest	Inquire whether have interests to do things.
Social Function	Inquire if there has been any impact on work, interpersonal relationships, etc.
Energy	Inquire about energy level and whether the patient feels tired.
Sleep	Inquire about the patient's sleep status, such as whether they are experiencing insomnia or early awakening.
Thinking Ability	Inquire whether there are symptoms of lack of concentration, poor memory, or hesitation.
Weight and Appetite	Inquire about changes in weight and appetite.
Somatic Symptoms	Inquire whether there are physical symptoms, such as dizziness, headache, restlessness, slow reaction, etc.
Self-worth	Inquire whether the patient feels a low sense of self-worth, lacks confidence, or has guilty feelings.
Self-harm or Suicide	Inquire about suicidal or self-harm ideation/behavior.
History	Inquire about the medical history of the patient's family and their own past medical history.
Screen	Inquire about symptoms of other mental disorders, such as bipolar disorder, anxiety disorder.

Table 14: Annotation Categories of Question Topics

Dialogue Act We also want to code the *dialogue act* of each round of the conversation. We classified dialogue act of doctor chatbots into two dimensions:

- **Empathy behaviors**: the actions displayed by doctor chatbots during the diagnostic process to provide emotional support to patients. We categorize empathy behaviors into three types: Suggestion, Understanding, and Encourage and support.
- **In-depth questions**: the behavior of asking follow-up questions about the patient's symptom. We categorize in-depth questions into three types: Duration, Cause, and Manifestation.

We then utilized ChatGPT to automatically label each sentence in the dialogue history, taking advantage of its promising annotation capabilities (Gilardi et al., 2023). After that, three annotators reviewed and corrected the results to ensure the data quality.

E Human Doctor Examples

Table 15 shows the example of screening questions in the dialogue history of human psychiatrist and patient chatbot. In addition to exploring the typical symptoms of depression, psychiatrists also inquire about symptoms related to other conditions, such as bipolar disorder and anxiety, as there is a high likelihood of comorbidity.

Condition	Screening Question
Mania	Was there ever a time when you were feeling the opposite of how you feel now, like really upbeat, happy, and full of energy, with lots of plans and such?
Hallucination	Do you hear strange voices or sound when you're alone? Do you worry about someone hurting you?
Social Phobia	Are you anxious about your social interactions specifically, or do you experience anxiety in other areas as well?

Table 15: Example of screening questions in the dialogue history of human psychiatrist and patient chatbot.

F Error Analysis

Although ChatGPT-based chatbots show good potential in simulating realistic patients and psychiatrists, there are still many instances of erroneous turns or unsatisfying response in the evaluation that can lead to a negative user experience. We will thoroughly analyze these errors to benefit the design of future chatbots.

Repetition We found that ChatGPT-based chatbot is easy to generate repetitive messages, because ChatGPT tends to mimic the language style of the previous rounds in the conversation. One direct manifestation is that the doctor chatbots always include a similar phrase before asking a new question (see Table 16).

Furthermore, in a few cases, the doctor chatbots may ask the same symptom multiple times, which can seriously harm the patient's experience. In Table 17, we can see an example where the doctor

Doctor bot	Okay, I understand. So how has your sleep been lately? Have you experienced any difficulties falling asleep, waking up too early, or poor sleep quality?
Patient	My sleep quality is por.
Doctor bot	Okay, I understand. How about your recent diet and weight situation? Have you experienced any loss of appetite, binge eating, or weight fluctuations?
Patient	No.

Table 16: Example of repetition and poor robustness to spelling errors.

chatbot asks repetitive questions about "tired or drowsy". Such repetition may be attributed to the patient's indirect response to the initial question, which may have caused ChatGPT to misinterpret the answer and prompt for clarification. This emphasizes the fact that ChatGPT's comprehension abilities are not yet flawless.

Doctor bot	Okay, do you feel mentally well, or do you feel tired or drowsy?
Patient	I feel unwell and I get sleepy at noon.
Doctor bot	Okay, do you feel tired or drowsy during the day? Do you have any emotional problems, such as anxiety, depression, or stress?
Patient	Yes, I have anxiety and stress.

Table 17: Example of repetition.

Poor Robustness to Spelling Errors We also observe that ChatGPT lacks robustness when confront with spelling errors. Table 16 is also an example of this type, where the patient misspells "poor" as "por". Had the misspelling not occurred, the doctor chatbot would have requested additional information regarding the patient's sleeping problems. However, ChatGPT fails to identify the mistake and proceeds to ask about the next symptom. This highlights a potential weakness in ChatGPT's ability to handle misspellings. To further confirm this, we write a prompt asking ChatGPT to provide a list of all the patient's symptoms, and it didn't include the symptom of "poor sleeping quality".

Symptom	robot-like Words	Human-like Words
Low Mood	low mood, sadness, and depression 情绪低落，悲伤，沮丧	downhearted, uncomfortable, dejected, and heartbroken 难过，难受，失落，伤心
Anxious		nervous, worried 紧张，担心
Loss of Interest	loss of interest, inability to get interested, decreased interest 失去...兴趣，对...提不起兴趣，兴趣减退	boring, not feeling like doing anything, not sure what to do, bored 没意思，什么都不想做，不知道该做什么，无聊
Fatigue	fatigue, weariness 疲劳，困倦	tired, exhausted 累，没力气
Attention	have difficulty in concentrating 难以集中注意力	
Self-worth	self-blame, low self-worth, damaged self-esteem 自罪，自我价值感低，自尊心受到打击	worthless, useless, meaningless, no point 一无是处，没用，有什么意义，没有意义
Pessimism	hopeless 无望	
Sleep Disturbance	sleep disturbance, excessive sleepiness 睡眠困难，嗜睡	can't sleep, insomnia, tossing and turning 睡不好，睡不着，失眠，翻来覆去
Weight and Appetite Change	Increased appetite, decreased appetite, loss of appetite 食欲增加，食欲下降，食欲不振	No appetite, not in the mood to eat, poor appetite 没胃口，没什么胃口，胃口不好，饭量
Psychomotor retardation	sluggish thinking 思维迟缓	Mind goes blank 脑子一片空白
Psychomotor agitation	Agitation, restlessness, irritability, or excessive talking 精神运动性激越，不安，烦躁不安，兴奋或话多	anxious, mentally unsettled, mind is racing, can't sit still 烦躁，静不下心，好像脑子一直在想事情，坐不住
Self-harm or Suicide	suicidal and self-harming thoughts 自杀和自伤的想法	want to die, jump off a building 不活，跳楼

Table 18: The Lexicon of Robot-like Words and Human-like Words