# LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation

**Siyuan Chen**[1], **Mengyue Wu**[2]*, **Kenny Q. Zhu**[3]*,
**Kunyao Lan**[4], **Zhiling Zhang**[5], **Lyuchun Cui**[6]
[1,2,3,4,5]Shanghai Jiao Tong University, Shanghai, China
[6]Shanghai Mental Health Center, Shanghai, China
{[1]chensiyuan925, [2]mengyuewu, [4]lankunyao, [5]blmoistawinde}@sjtu.edu.cn,
[3]kzhu@cs.sjtu.edu.cn, [6]cuilvchun@outlook.com

## Abstract

Empowering chatbots in the field of mental health is receiving increasing amount of attention, while there still lacks exploration in developing and evaluating chatbots in psychiatric outpatient scenarios. In this work, we focus on exploring the potential of ChatGPT in powering chatbots for psychiatrist and patient simulation. We collaborate with psychiatrists to identify objectives and iteratively develop the dialogue system to closely align with real-world scenarios. In the evaluation experiments, we recruit real psychiatrists and patients to engage in diagnostic conversations with the chatbots, collecting their ratings for assessment. Our findings demonstrate the feasibility of using ChatGPT-powered chatbots in psychiatric scenarios and explore the impact of prompt designs on chatbot behavior and user experience.

## 1 Introduction

Conversational Agents (i.e., chatbots) are becoming increasingly popular in the mental health domain (Sabour et al., 2022). Applications designed for mental health therapy or coaching in daily life, such as Woebot[1] and Wysa[2], are gaining widespread attention for their ability to reduce users' negative emotions (Grové, 2021) and promote a healthy lifestyle (Fadhil et al., 2019). Another notable application is chatbot-based symptom checkers (You et al., 2023), which emulate human-like conversations while assessing users' symptoms, resembling interactive questionnaires.

However, there is still limited exploration in developing and evaluating chatbots that can (i) conduct diagnosis conversations like a psychiatrist or (ii) simulate patients in the psychiatric outpatient scenarios, though they have significant real-world applications. Doctor chatbots can be effective tools for mental disorder screening (Pacheco-Lorenzo et al., 2021) in lieu of official medical diagnosis. Patient chatbots can serve as Standard Patients (SP) in medical education, making the process more efficient and cost-effective (Torous et al., 2021).

Developing and evaluating such chatbots is particularly challenging due to the unique nature of mental health issues, including (i) the difficulty in obtaining data because of privacy concerns; (ii) the inherent ambiguity and subjectivity of mental disease symptoms. Moreover, relying solely on scales (e.g., PHQ-9) for mental disorder screening cannot provide trustworthy diagnosis, because in real outpatient scenario, patients often feel ashamed or afraid of disclosing their true conditions and difficult to describe their mental state objectively (Salaheddin and Mason, 2016). Thus, even experienced psychiatrists struggle to obtain the meaningful response from patients.

Consequently, the design goals and conversational styles of these chatbots are different from the chatbots for mental health therapy and symptom checking. We argue that chatbots for psychiatric diagnosis can not achieve satisfying performance by simply collecting symptoms like questionnaires. Instead, they should be equipped with various professional skills, such as emotional support, to complete the diagnosis task effectively. What's more, patient chatbots should aim to resemble real patients more closely, rather than precisely and robotically reporting their symptoms without any emotional fluctuations.

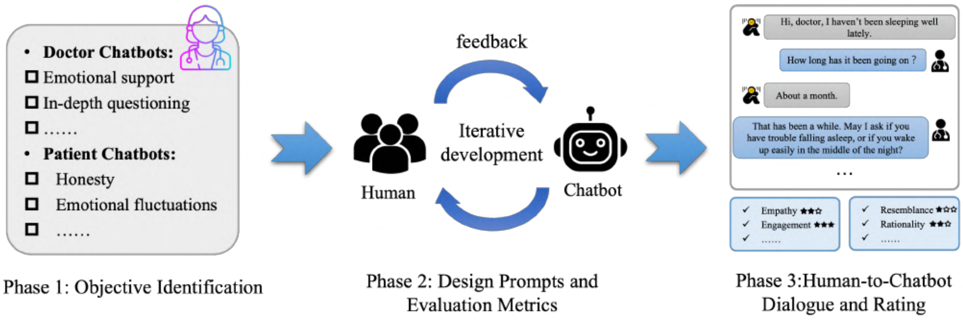
Achieving these goals is quite difficult for conventional rule-based (Medeiros and Bosse, 2018; Jaiswal et al., 2019) or data-based (Yao et al., 2022; Fansi Tchango et al., 2022; Lin et al., 2021) methods. Fortunately, recent advancements in large language models (LLMs), especially with the emergence of ChatGPT[3], provide a new way to develop chatbots that can convincingly portray spe-

*Corresponding Authors.

[1] https://woebothealth.com

[2] https://www.wysa.com

[3] https://chat.openai.com/



Hi, doctor, I haven't been sleeping well lately.

How long has it been going on ?

About a month.

That has been a while. May I ask if you have trouble falling asleep, or if you wake up easily in the middle of the night?

...

Phase 1: Objective Identification

Phase 2: Design Prompts and Evaluation Metrics

Phase 3:Human-to-Chatbot Dialogue and Rating

Figure 1: The overview of the psychiatrist-guided three-phase study.

cific roles. Equipped with comprehensive training data and knowledge, LLMs can generate diverse tones and symptom descriptions with appropriate prompts rather than fine-tuning on extensive domain data.

Therefore, in this work, we aim to (i) respectively investigate the potential of ChatGPT in simulating psychiatrists and mental disordered patients in a clinical diagnosis scenario[4], as well as (ii) build a comprehensive evaluation framework for these chatbots, answering the question about what constitutes an exceptional psychiatrist chatbot and a truly patient-like chatbot. To develop and evaluate a system that truly satisfies users' expectations, we followed a human-centered design methodology. The study consists of three phases (See Figure 1). We first collaborated with psychiatrists to identify a set of objectives for doctor and patient chatbots (**Phase 1**). Based on these objectives, we conducted an experimental study (**Phase 2**) to design appropriate prompts for ChatGPT-based chatbots and establish an evaluation framework that incorporates both human evaluation and automatic metrics aligned with the objectives from Phase 1. Importantly, the design of prompts and metrics was iterated based on human feedback, with each version evaluated and improved with input from psychiatrists.

Further, to better evaluate the performance of these chatbots with varying prompt designs, we recruit real psychiatrists and patients to engage in diagnostic conversations with the simulated patient and doctor chatbots, respectively, and collect their ratings after conversation (**Phase 3**). We also conduct a comparison between the behavior of real and simulated psychiatrists based on the dialogue history, which yields some interesting findings. The main contributions of this work are:

- We formalize the task of developing doctor and patient chatbot for diagnostic purposes in a psychiatric outpatient setting.
- We conduct a user-centered investigation into the design and evaluation of these chatbots. Through an iterative development process, we actively sought feedback from both patients and psychiatrists, allowing us to establish a more solid and applicable chatbot system and evaluation framework.
- Through detailed prompt engineering and experiments, we demonstrate the feasibility of utilizing ChatGPT-powered chatbots in professional domains that demand specialized skills or unique language style. We also use interactive human evaluation to explore how different prompt designs influence user experience.

## 2 Objectives

In phase 1, we consulted with 7 psychiatrists and worked together to establish the objectives we hope the doctor and patient chatbots can achieve, which will guide us in the following stage.

Since the diagnosis standards of different mental disorders vary greatly, psychiatrists recommend concentrating on depressive disorders for this study, while leaving the scaling to include other disorders as future work.

### 2.1 Doctor Chatbot

As a doctor chatbot, the primary task is to conduct a professional diagnostic process for the patient and provide an accurate diagnosis. To achieve this, and to offer patients a superior healthcare experience, a good doctor chatbot should possess the following three capacities:

- **Comprehensiveness:** Inquire about the key symptoms of depression, including sleep,

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

mood, diet, and other relevant aspects that are required for diagnosis.

- **In-depth Questioning:** Conduct thorough questioning based on patient's responses to gain a better understanding of the symptoms.

- **Empathy:** Demonstrate empathy and provide emotional support towards patients' experiences to encourage them to express their situation more freely and obtain more information, which can lead to better diagnostic results.

### 2.2 Patient Chatbot

After establishing objectives for doctor chatbots, we encountered difficulties when defining the requirements for chatbots that resemble real patients. This is due to the fact that individuals with the same disorder can exhibit significant variations in their manifestations. Moreover, psychiatrists, though experienced, have no firsthand chatting experience with a "non-patient-like" chatbot, making it challenging for them to generalize the requirements for a "patient-like" chatbot.

To address this issue, we decide to develop an initial version of the chatbot first. This allows psychiatrists to interact with "non-patient-like" examples, which can help them better define the characteristics and behaviors that constitute a "patient-like" chatbot. Based on their feedback, we then iterate and update the chatbot accordingly. At this phase, we only establish one fundamental requirement for a patient chatbot.

- **Honesty:** Provide an accurate and rational description of symptoms in its user profile, without reporting any nonexistent symptoms.

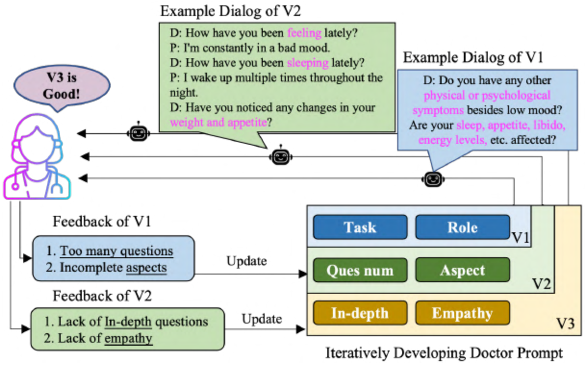
## 3 Prompt Design

We describe the iterative methodology of designing prompts with users' feedback, which will be listed in bullet point in this section.

### 3.1 Doctor Chatbot

**Version 1** The original version of the prompt for doctor chatbot is as follows. We simply describe the task without providing any other information.

> 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.



D: How have you been feeling lately?
P: I'm constantly in a bad mood.
D: How have you been sleeping lately?
P: I wake up multiple times throughout the night.
D: Have you noticed any changes in your weight and appetite?

D: Do you have any other physical or psychological symptoms besides low mood? Are your sleep, appetite, libido, energy levels, etc. affected?

Figure 2: The iterative development process of the prompt of doctor chatbots. Psychiatrists will identify the limitations of the current version, and we will address these issues in the subsequent version.

- Although the chatbot's questions are in line with DSM-5(APA et al., 2013), it asks about almost all symptoms in one go (See Example V1 in Figure 2), which can be overwhelming for the patient.

- Additionally, the chatbot draws a diagnosis conclusion in less than five turns without gathering sufficient information, which may result in unreliable diagnostic outcomes and fail to achieve the objective of "comprehensiveness".

**Version 2** Therefore, we developed a new version to address these problems by adding sentence ③ and ④ (i.e., "ques num" and "aspect" in Figure 2) to the original prompt.

> ③ Your questions should cover at least the following aspects: [...][5]. ④ Please only ask one question at a time.

- After the modification, the doctor chatbot does satisfy the requirements in the prompt, but it appears indifferent to the patient's input and mechanically transits to the next aspect without offering empathy or support (see example V2 in Figure 2).

- Moreover, psychiatrists also emphasize the importance for the doctor chatbot to ask in-depth questions. For example, if a patient expresses feeling down, the chatbot should ask the follow-up question, "How long have you been experiencing this mood?" to gain a better understanding of the symptom, rather than inquiring next symptom immediately.

[5] We will provide the full list in the Appendix.

**Version 3** Therefore, we focus on empathy and in-depth questioning in the upcoming version.

We first modify sentence ① as "Please play a role of an empathetic and kind psychiatrist". Then, we add sentences ⑤⑥ to the previous version.

> 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.

We include examples (highlighted in colored boxes) in the prompt to guide the doctor chatbot in asking in-depth questions and demonstrating empathy. These examples are crucial because, without them, the chatbot tends to ask superficial questions and rely on a generic phrase like "thank you very much for your answer" to show empathy. This arises from ChatGPT's limited comprehension of "in-depth questioning" and "empathy" in clinical contexts. Consequently, providing examples can be a promising approach to help ChatGPT grasp certain specialized skills within professional domains.

Since this version fulfills the three requirements for doctor chatbots, we deem it the final iteration[6].

### 3.2 Patient Chatbot

**Version 1** The original version of the prompt for patient chatbot is as follows.

> ① Please play the role of a patient, who is currently chatting with a doctor. ② You are experiencing the following symptoms: `[Symptom List]`[7] ③ Please talk to me based on the above symptom list. ④ You can only mention one symptom per round.

Similarly, we simply describe the task, provide a symptom list in the prompt, and add sentence ④ to avoid listing all the symptoms in one turn.

Through experimentation, we observed that the chatbot can fulfill the basic "honesty" requirement in most cases. However, the psychiatrists generally found the chatbot does not resemble patients, and highlighted numerous behaviors commonly exhibited by real patients during consultations that differed significantly from the chatbot's responses.

- **Emotion:** Patients in a depressed mental state may experience emotional fluctuations during the conversation, while the chatbot's presentation of symptoms is too calm and polite.

- **Expression:** Patients use colloquial expressions when describing symptoms, and may have difficulty expressing themselves clearly. They often talk about their daily life experiences. However, the chatbot tends to use formal language similar to the official diagnostic criteria (DSM-5).

- **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. In contrast, the chatbot is overly cooperative, readily acknowledging and providing detailed descriptions of its symptoms.

**Version 2** According to the above suggestions provided by the psychiatrists, we revised the prompt by adding the following instructions:

> 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.

After adding these instructions, we found that the language of the patient chatbot became more natural. Sometimes it even expressed reluctance to seek help and made some emotional statements. It becomes more human-like and less like "a polite and calm AI". We provide several utterances of the patient chatbot in Table 13 (Appendix E).

However, we found that only the first few rounds of conversation could clearly reveal the effect of adding sentences ⑤⑥⑦ to the prompt, indicating that the patient chatbot is prone to forgetting some of the instructions given at the beginning.

To address the issue of forgetting, we insert new reminders during the conversation. Inspired by the fact that the latter part of the prompt has the greatest impact on the responses generated by ChatGPT, our method is straightforward yet effective. Without the users' awareness, we subtly append the following words at the end of the most recent sentence in the dialogue history.

[6] The final version is in Table 8 in the Appendix.

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

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

We aim to use simple phrases or words as reminders during the conversation to ensure that the sentences are not overly long. Moreover, these reminders are only temporarily attached to the most recent round, and will not persist in the dialogue history for subsequent rounds. With these reminders, the patient chatbot can maintain a colloquial language style consistently and exhibit resistance even in the latter part of the conversation, so we consider this version as the final one[8].

## 4 Evaluation Framework

To assess the performance of dialogue systems, it is crucial to employ both human evaluation and automatic metrics, especially in mental health domain. Since there is little previous work on how to evaluate simulated psychiatrists and patients, we design several task-specific metrics and interactive experiments for human evaluation. This section provides a detailed discussion of these metrics and experiment design.

### 4.1 Human Evaluation

We first implemented a website to host our chatbots, making it easier for participants to interact with them and rate their performance. The details of the website can be found in Appendix D.2.

#### 4.1.1 Participants

To evaluate the performance of different chatbots in real-world scenarios, we recruited real depression patients and psychiatrists.

Depression patients are recruited through online advertisements. A total of 14 volunteers completed the entire process, with ages ranging from 18 to 31, and male and female participants accounted for 28.57% and 71.43% respectively. Notably, we have a balanced distribution of healthy, mild, moderate and severe depression subjects.

We invited 11 psychiatrists who are not involved in the prompt design, through cooperation with hospitals. 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.

#### 4.1.2 Human Evaluation Process

Due to the complexity and high time cost of human evaluation, we select several representative prompt versions for comparison, and discuss the evaluation process of doctor and patient chatbots respectively.

**Doctor Chatbot** First, patients are asked to complete the Beck Depression Inventory (Beck et al., 1996) to assess the severity of their depression, serving as the ground truth of diagnosis. The severity distribution, presented in Table 11 in Appendix D.1, is balanced among the participants.

Next, each patient will have a conversation with four different doctor chatbots in a random order, and then rate them on four human evaluation metrics, which will be introduced in Section 4.1.3, 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, D4, is a representative deep learning chatbot trained on domain-specific data (Yao et al., 2022) using CPT model (Shao et al., 2021).

**Patient Chatbot** Each psychiatrist participants needs to have a conversation with two different patient chatbots, and then rate their performance with 1-4 scale. The two patient chatbots are P1 and P2, aligning with the two prompt versions in Section 3.2. A brief description of these chatbots is in Table 1.

| | Chatbot | Description |
|---|---|---|
| Doctor | D1 | use the full doctor prompt |
| | D2 | remove empathy parts in prompt |
| | D3 | remove aspect part in prompt |
| | D4 | CPT model trained on domain data |
| Patient | P1 | use version 1 patient prompt |
| | P2 | use version 2 (full) patient prompt |

Table 1: Brief description of the chatbots for comparison. Detailed description and prompt is in Appendix B.

To ensure the quality of the dialogue data and evaluation, we also utilize a series of quality control strategies, which can be found in Appendix D.3.

#### 4.1.3 Human Evaluation Metrics

**Doctor Chatbot** In most cases, patients do not have specialized knowledge in psychiatry, making it difficult for them to assess a doctor's professional skills precisely. Therefore, when designing human evaluation metrics for doctor chatbots, we focus mainly on the user experience and referred to some evaluation metrics for conversational agents in the

[8] The final version is in Table 9 in Appendix B.

previous works (Yao et al., 2022). The proposed human evaluation metrics are shown in Table 2.

| 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 2: Human evaluation metrics of doctor chatbot.

**Patient Chatbot** 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 resamble patients enough and whether they can describe their symptoms in a reasonable way, so the main metrics for human evaluation are **Resamblance** and **Rationality**.

What's more, we divide the Resamblance metric into three aspects in Table 3, according to the psychiatrists' suggestions in Section 3.2.

| 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 3: Three aspects of the "Resamblance" metric.

### 4.2 Automatic Metrics

We can divide the automatic metrics of both kind of chatbots into two types: **functionality** and **style**.

**Doctor Chatbot** Different from human evaluation metrics, we mainly measure the expertise of the doctor chatbot using automatic metrics. The *functional* requirements for doctor chatbot is to decide the next question based on the patient's current description, completely collect the patient's symptom-related information, and provide an accurate diagnosis in the end. Consequently, we characterized functional performance using "diagnosis accuracy" and "symptom recall".

The *style* part concerns the doctor chatbot's professional skills, such as asking in-depth questions and conducting diagnosis in an efficient way. A higher level of professionalism can enhance the patient's diagnostic experience and enable the collection of more comprehensive information for diagnosis. There are three metrics in this part, including "in-depth ratio", "avg question num", and "symptom precision".

**Patient Chatbot** The *functional* requirement of patient chatbot is "Honesty", meaning it should accurately report its symptoms without fabricating nonexistent ones. To assess this, we can calculate "wrong symptom ratio" by comparing the patient's persona with the symptoms it reported.

Then, we evaluate the patient chatbots' *style* using some linguistic features, like "Distinct-1", "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. Detailed explanation of these automatic metrics for doctor and patient chatbot is provided in Appendix D.1.

## 5 Experiments

In this section, we will introduce the evaluation results of doctor chatbot and patient chatbot.

### 5.1 Doctor Chatbot Results

**Human Evaluation** We present the human evaluation results of different doctor chatbots in Table 4. Chatbots utilizing prompts with empathy components (i.e., D1 and D3) are scored higher in "Empathy" metrics than other chatbots. Surprisingly, 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 | D4 |
|---|---|---|---|---|
| Fluency | 3.00 | 3.07 | **3.28** | 3.14 |
| 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 evaluation scores of doctor 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. It became evident that D1 often repetitively expresses empathy, relying on phrases like "I understand your feelings" multiple times within a single conversation. 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.

**Automatic Evaluation** Then, we continue to explore the dialogue history and calculate some automatic metrics, hoping to find out more reasons of the unexpected human evaluation results. The results of automatic metrics are in Table 5.

We can find that D3 has the fewest average number of dialogue turns and the least amount of text per turn among all the ChatGPT-based bots. Additionally, it tends to ask more in-depth questions while asking fewer questions each turn, both of which indicate higher professional skills as a psychiatrist. Furthermore, the symptom precision metric is the highest, suggesting that the chatbot's questions are highly efficient, with few "no" responses. However, as the required aspects are not explicitly stated in the prompt, the symptom recall metric of this chatbot is relatively low, indicating that its diagnosis may not be comprehensive enough. Nevertheless, the chatbot's questions are more flexible and free-flowing, precisely because there are no predetermined aspects to ask. As a result, patients feel more understood, leading to a better experience overall.

What's more, D2 received the longest responses from patients, which is consistent with the human evaluation metric "Engagement", suggesting that patients are more willing to converse with this chatbot. It also achieves the highest symptom recall among all the chatbots, even surpassing D1 which also includes aspects in the prompt. This could because D1 contains too many instructions regarding empathy and other factors, which may have hindered its ability to thoroughly inquire about all the required symptoms.

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| - Statistics | | | | |
| avg turns | 25.64 | 24 | **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 |
| - Functionality | | | | |
| diagnose acc | 36.36% | 18.18% | **55.56**% | - |
| symp recall | 58.93% | **66.07**% | 38.10% | 61.90% |
| - Style | | | | |
| avg # of ques | 1.6 | 1.9 | **1.22** | 0.92 |
| in-depth ratio | 25.08% | 27.64% | **32.64**% | 41.39% |
| symp precision | 72.40% | 71.93% | **92.24**% | 49.61% |

Table 5: Automatic evaluation scores of doctor chatbots

## 5.2 Human vs. Doctor Chatbots

As we invited human doctors to engage in conversations with patient chatbots, we are able to analyze the behaviors of human doctors and chatbots, and establish an upper bound for our doctor chatbots. Therefore, we first annotate the question topics, dialogue acts (i.e., empathy behaviors and in-depth questions) in the dialogue history, which is described in Appendix D.4.

**Topic Proportion** Accordingly, we calculated the average proportion of question topics of different doctor chatbots, as well as human doctors. Figure 3 displays the outcomes.

Most doctor 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, with relatively greater emphasis on emotion, somatic symptoms, and social function. Moreover, human doctors often do "screening" to rule out other possible conditions (see Example 2 in Appendix E), while chatbots rarely exhibit such behavior, indicating the possible limitations in multi-disease scenarios (Zhang et al., 2022).

**Empathy Behaviors** Then we calculated the average number of empathetic strategies utilized by doctors in the dialogue history, as illustrated in Figure 4a. The figure shows that D4 does not exhibit any empathetic behaviors. Conversely, when prompted with empathy instructions, D1 and D3 are capable of utilizing a range of empathetic strategies, while D2 only offers suggestions to patients. Moreover, though human doctors use all the strategies, their usage is less frequent than that of chatbots. When asked for the reasons behind this, doctors attributed it to the limited inquiry time in real outpatient scenarios and the bias resulting from the difference in interaction with chatbots compared to real people (Yun et al., 2021).

**In-depth Questions** Further, we also calculated the various ways of in-depth questioning, and obtained Figure 4b. Our analysis revealed 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.

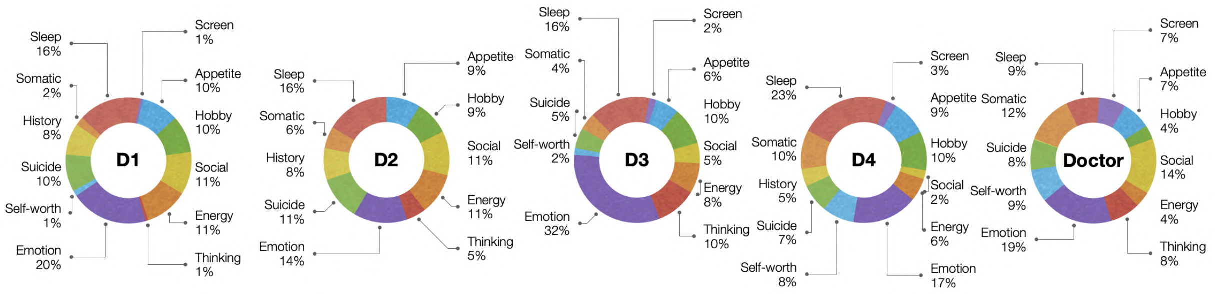
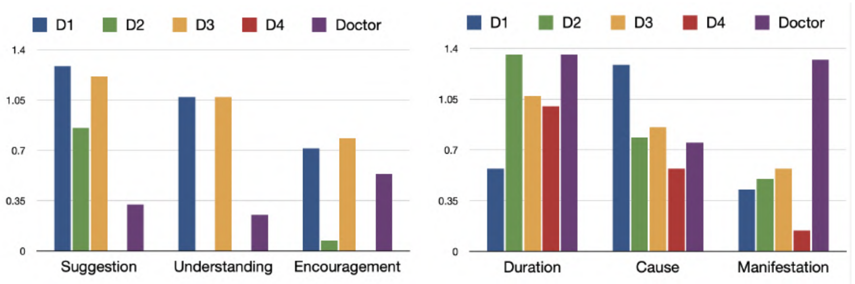


Figure 3: The Proportion of Symptoms Asked by Different Doctor Chatbots and Human Doctor.



(a) Empathy Behaviors

(b) In-depth Question

Figure 4: Dialogue Act Comparison between Different Doctor Chatbots and Human doctor.

### 5.3 Patient Chatbot Results

**Human Evaluation** The human evaluation results of patient chatbot are in Table 6. It can be observed that all metrics of P2 are higher than P1, especially in terms of "Mental state" and "Expression style". This suggests that the inclusion of resistance, colloquialism, etc., makes the chatbot more similar to real patients, according to the doctors' perspective.

| | P1 | P2 |
|---|---|---|
| Realistic | 1.93 | **2.21** |
| Mental State | 2.07 | **2.42** |
| Life Experience | 2 | **2.14** |
| Expression style | 1.57 | **2.21** |
| Rationality | 2.42 | **2.57** |

Table 6: Human evaluation scores of patient chatbot

**Automatic Evaluation** We show the results of automatic 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 doctor 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.

In informal conversations, people often rely on a smaller set of familiar words and phrases. This limited range of vocabulary contributes to less diversity in spoken language compared to written language. Therefore, the lower Distinct-1 in P2 indicates that its language style is more colloquial compared to P1. Additionally, P2 also has more human-like words and fewer robot-like words, which supports the higher human evaluation score in the dimension of "expression style".

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. One possible reason for this could be the excessive focus on language style and resistance in the prompt, which might cause ChatGPT to "forget" the actual symptoms of the patient.

| | 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 |
| - Functionality | | |
| wrong symp ratio | **15.07%** | 18.38% |
| - Style | | |
| Distinct-1 | 42.6% | **37.3%** |
| 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: Automatic evaluation scores of patient chatbot

## 6 Conclusion

In this work, we investigated the capacity of ChatGPT to serve as the underlying technology for developing chatbots that can emulate psychiatrists and patients with mental disorders, respectively. To ensure the validity of our approach, we collaborated with 7 professional psychiatrists who provided their expertise and insights throughout the study. With their guidance, we developed a comprehensive evaluation framework that takes into account the distinctive characteristics of diagnostic conversations within the mental health domain. We then evaluated the performance of different chatbots, each utilizing distinct prompts, and observed how varying designs can influence chatbot behavior. This provides valuable insights for future studies in this area.

## 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.

## 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.

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.

Ahmed Fadhil, Yunlong Wang, and Harald Reiterer. 2019. Assistive conversational agent for health coaching: A validation study. Methods of Information in Medicine, 58:009 – 023.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. In Advances in Neural Information Processing Systems, volume 35, pages 31306–31318. Curran Associates, Inc.

Chris Gillette, Robert B. Stanton, Nicole Rockich-Winston, Michael Rudolph, and Jr. H. Glenn Anderson. 2017. Cost-effectiveness of using standardized patients to assess student-pharmacist communication skills. American Journal of Pharmaceutical Education, 81(10).

Christine Grové. 2021. Co-developing a mental health and wellbeing chatbot with and for young people. Frontiers in Psychiatry, 11.

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, IVA '19, page 81–87, New York, NY, USA. Association for Computing Machinery.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. Journal of general internal medicine.

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

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. Proceedings of the AAAI Conference on Artificial Intelligence, 35(15):13362–13370.

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.

Lenin Medeiros and Tibor Bosse. 2018. Using crowdsourcing for the development of online emotional support agents. In Practical Applications of Agents and Multi-Agent Systems.

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 Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2022. Chatbots for mental health support: Exploring the impact of emohaa on reducing mental distress in china.

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.

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.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis.

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.

Jin Ho Yun, Eun-Ju Lee, and Dong Hyun Kim. 2021. Behavioral and neural evidence on consumer responses to human doctors and medical artificial intelligence. Psychology & Marketing, 38:610–625.

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.

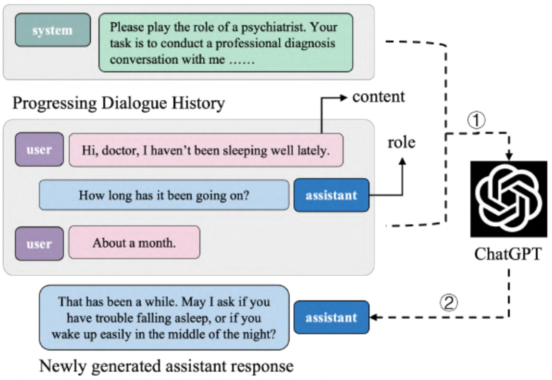
Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities?

# A Mechanism of ChatGPT-powered Chatbot

We utilize the chat model[9] developed by OpenAI to build our chatbots. This model operates by taking a sequence of *messages* as input, and returns a

[9] https://platform.openai.com/docs/guides/chat/

model-generated response. As Figure 5 shows, at each turn, we combine the system message and the ongoing conversation history with alternating user and assistant messages into a sequence and feed them into the ChatGPT language model. The resulting output is the Chatbot's response to the user's input.


system
Progressing Dialogue History
content
role
①
user
assistant
user
ChatGPT
assistant
②
Newly generated assistant response

Please play the role of a psychiatrist. Your task is to conduct a professional diagnosis conversation with me ......

Hi, doctor, I haven't been sleeping well lately.

How long has it been going on?

About a month.

That has been a while. May I ask if you have trouble falling asleep, or if you wake up easily in the middle of the night?

Figure 5: The reponse generation process of ChatGPT-based chatbots. ① means combining the system message and the dialogue histroy together as the input of ChatGPT. ② means ChatGPT generates new response according to the input.

The system message serves as an instruction for ChatGPT, providing information about the task and some specific requirements needed to generate an appropriate response. Prompt engineering, or the design of the system message, is critical to achieving better performance, as it sets the context for the large language model and guides its output.

## B Details about Chatbots for Comparison and the Prompts

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

- D1: using the full doctor 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)
- D4: using the CPT model (Shao et al., 2021) trained on the D4 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 3.2 (i.e., Sentence①②③④⑤⑥⑦), and inserting reminders during the conversation.

The different versions of prompt for doctor and patient chatbot are in Table 8 and Table 9 respectively.

## C Symptom List Summarization

The symptom list for patient prompt in Section 3.2 is summarized from the dialogue history of real patients and doctor 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 10 shows three example of summarized symptom lists, whose format is: SYMPTOM (DESCRIPTION).

## D Details about Evluation Framework

### D.1 Evaluation Metrics

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

### D.1.1 Doctor Chatbot

- **Diagnosis accuracy**: The accuracy of the doctor 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).
- **Symptom recall**: The proportion of aspects asked by the doctor chatbot out of all aspects needed to be asked in a depression diagnosis conversation (See the categories in Table. 12).
- **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

| | 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 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. |
| D3 | 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. |

Table 8: Doctor 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 9: Patient Chatbot Prompts

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 question num**: According to the previous work, GPT tend to generate long responses (Wei et al., 2023). Similarly, ChatGPT-based doctor 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 doctor 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 doctor chatbot asked, to measure the efficiency of the doctor chatbot’s questioning.

### D.1.2 Patient Chatbot

- **Distinct-1**: Distinct-1 is the total number of unique words divided by the total number of words in a given sentence, which can measure the lexical diversity.

- **Human/robot-like word ratio**: 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 17), and then calculated the proportion of robot/human vocabulary used by each patient.

| 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 10: The symptom list examples of different patients.

- **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.

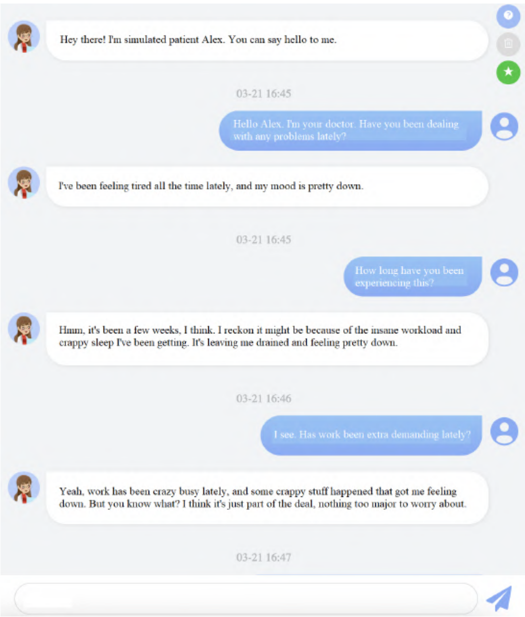
| none | mild | moderate | severe |
|---|---|---|---|
| 4 | 3 | 4 | 3 |

Table 11: The distribution of depression severity among participants.

### D.2 Chat Interface

To host our chatbots, we developed a web interface (see Figure 6). 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.



Hey there! I'm simulated patient Alex. You can say hello to me.

03-21 16:45

Hello Alex. I'm your doctor. Have you been dealing with any problems lately?

I've been feeling tired all the time lately, and my mood is pretty down.

03-21 16:45

How long have you been experiencing this?

Hmm, it's been a few weeks, I think. I reckon it might be because of the insane workload and crappy sleep I've been getting. It's leaving me drained and feeling pretty down.

03-21 16:46

I see. Has work been extra demanding lately?

Yeah, work has been crazy busy lately, and some crappy stuff happened that got me feeling down. But you know what? I think it's just part of the deal, nothing too major to worry about.

03-21 16:47

Figure 6: 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 12.

| 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 12: 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. After that, three annotators reviewed and corrected the results to ensure the data quality.

## E Examples

In this section, we list all the example dialogue fragments to better showcase the distinctive characteristics of each chatbot.

**Example 1** Table 13 shows some examples of what the patient chatbot with prompt V2 replied when asked by the doctor.

| | |
|---|---|
| 1 | It's like I toss and turn all night, and those darn nightmares sometimes creep in too. |
| 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. |
| 3 | 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. |

Table 13: Example utterance of patient chatbot powered by the Version 2 prompt.

Interestingly, sometimes the patient chatbot will give expressions or actions at the beginning of their sentences in parentheses. This could be attributed to ChatGPT's pre-training data, which may contain scripts utilizing this format. These expressions and actions can make the conversation more immersive and realistic to some extent.

**Example 2** Table 14 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?<br>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 14: 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 15).

| | |
|---|---|
| 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 15: Example of repetition and poor robustness to spelling errors.

Furthermore, in rare cases, the doctor chatbots may ask the same symptom multiple times, which can seriously harm the patient's experience. In Table 16, we can see an example where the doctor 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 16: Example of repetition.

**Poor Robustness to Spelling Errors** We also observe that ChatGPT lacks robustness when confront with spelling errors. Table 15 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".

## G Related Works

### G.1 ChatGPT for Mental Health

Recently, several studies have assessed the performance of ChatGPT in tasks like depression detection (Lamichhane, 2023), emotional conversation (Zhao et al., 2023), factor detection of mental health conditions, and emotion recognition in conversations (Yang et al., 2023). However, these evaluations were performed on existing datasets using conventional metrics, and did not involve human interaction. What's more, Qin et al. (2023) developed a chat interface using ChatGPT. However, they mainly focus on more interpretable and interactive depression detection from social media, while our work focuses on outpatient scenarios, where information should be obtained from conversation, and user experience is the major concern.

### G.2 Doctor Chatbot

Automatic diagnosis by doctor chatbot has significant practical applications. It enables large-scale screening, alleviates the issue of insufficient medical resources, and provides patients with a more engaged experience than using scales like PHQ-9 (Kroenke et al., 2001).

While numerous chatbots have been developed to automatically diagnose physical illnesses (Xu et al., 2019; Wei et al., 2018), such chatbots remain relatively uncommon in the mental health domain due to the difficulty in obtaining dialogue data because of ethical concerns. Yao et al. (2022) introduced a depression diagnosis dialogue dataset performed by patient and doctor actors, and a doctor chatbot trained on it. Although the 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

| Symptom | robot-like Words | Human-like Words |
|---|---|---|
| Low Mood | low mood, sadness, and depression<br>情绪低落，悲伤，沮丧 | downhearted, uncomfortable, dejected, and heartbroken<br>难过，难受，失落，伤心 |
| Anxious | | nervous, worried<br>紧张，担心 |
| Loss of Interest | loss of interest, inability to get interested, decreased interest<br>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<br>没意思，什么都不想做，不知道该做什么，无聊 |
| Fatigue | fatigue, weariness<br>疲劳，困倦 | tired, exhausted<br>累，没力气 |
| Attention | have difficulty in concentrating<br>难以集中注意力 | |
| Self-worth | self-blame, low self-worth, damaged self-esteem<br>自罪，自我价值感低，自尊心受到打击 | worthless, useless, meaningless, no point<br>一无是处，没用，有什么意义，没有意义 |
| Pessimism | hopeless<br>无望 | |
| Sleep Disturbance | sleep disturbance, excessive sleepiness<br>睡眠困难，嗜睡 | can't sleep, insomnia, tossing and turning<br>睡不好，睡不着，失眠，翻来覆去 |
| Weight and Appetite Change | Increased appetite, decreased appetite, loss of appetite<br>食欲增加，食欲下降，食欲不振 | No appetite, not in the mood to eat, poor appetite<br>没胃口，没什么胃口，胃口不好，饭量 |
| Psychomotor retardation | sluggish thinking<br>思维迟缓 | Mind goes blank<br>脑子一片空白 |
| Psychomotor agitation | Agitation, restlessness, irritability, or excessive talking<br>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<br>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<br>自杀和自伤的想法 | want to die, jump off a building<br>不活，跳楼 |

Table 17: The Lexicon of Robot-like Words and Human-like Words

need to thoroughly collect the patients' symptoms.

### G.3 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). Additionally, patient chatbots could serve as standardized patients (SPs) in medical education, as currently, actors are hired to portray patients with mental disorders, which is both costly and time-consuming (Gillette et al., 2017).

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).