Breaking the Stigma! Unobtrusively Probe Symptoms in Depression Disorder Diagnosis Dialogue

Jieming Cao   Chen Huang   Yanan Zhang*
Ruibo Deng   Jincheng Zhang   Wenqiang Lei
Sichuan University, Chengdu, China
Engineering Research Center of Machine Learning and Industry Intelligence,
Ministry of Education, China
{caojieming1, dengruibo}@stu.scu.edu.cn
{huangc.scu, erzmuxin}@gmail.com   {zyn, wenqianglei}@scu.edu.cn

Abstract

Stigma has emerged as one of the major obstacles to effectively diagnosing depression, as it prevents users from open conversations about their struggles. This requires advanced questioning skills to carefully probe the presence of specific symptoms in an unobtrusive manner. While recent efforts have been made on depression-diagnosis-oriented dialogue systems, they largely ignore this problem, ultimately hampering their practical utility. To this end, we propose a novel and effective method, UPSD^4, developing a series of strategies to promote a sense of unobtrusiveness within the dialogue system and assessing depression disorder by probing symptoms. We experimentally show that UPSD^4 demonstrates a significant improvement over current baselines, including unobtrusiveness evaluation of dialogue content and diagnostic accuracy. We believe our work contributes to developing more accessible and user-friendly tools for addressing the widespread need for depression diagnosis.

1 Introduction

Depression disorder is a global health crisis affecting millions, yet many cases go undiagnosed (Farid et al., 2020; The World Health Organization, 2022), highlighting the urgent need for early and accurate diagnosis. However, a key challenge hindering diagnosis is the **Stigma** (Shafi et al., 2014; Barney et al., 2006). It makes people feel ashamed or embarrassed to seek help, preventing them from open conversations about their struggles (Tourangeau et al., 2000; Anderson et al., 2002; Kaywan et al., 2023). This reluctance is particularly pronounced when discussing sensitive topics (e.g., suicide tendency), which are unfortunately essential components of the diagnostic criteria for depression (Lee, 1993; Miller, 1985; Tourangeau and Yan, 2007). To this end, traditional diagnosis has relied on **trained**
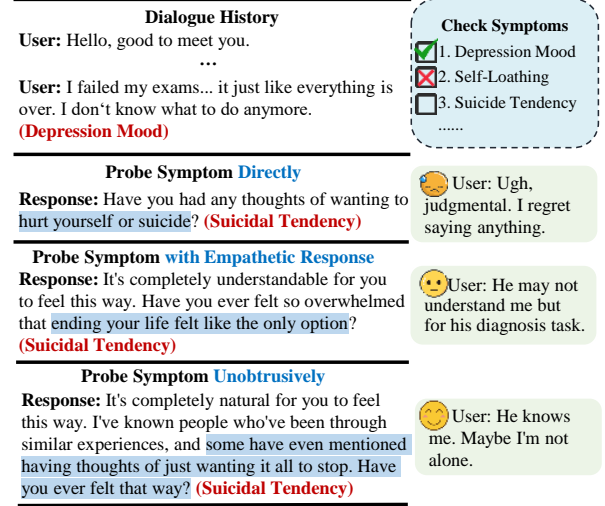
---
*Corresponding author.

Figure 1: An example of a comparison of the existing method with unobtrusively probing symptoms in depression diagnosis is when the user exhibits *Depression Mood* and next turns to probe *Suicide Tendency*.

professionals with questioning skills to carefully probe the presence of specific symptoms in an unobtrusive way (Webb et al., 1999; Smith et al., 2013; Bordin, 1979). However, it requires a high cost of human labor (Luppa et al., 2007), posing a challenge in effectively addressing the widespread need for depression diagnosis.

Recent advances in depression-diagnosis-oriented dialogue systems (Yao et al., 2022) have possessed the remarkable ability to engage users in conversations. This ability to identify potential symptoms through user responses holds the promise of automatic depression diagnosis. While effective in identifying symptoms, these systems often conduct a stiff dialogue with users, ignoring users' emotional state. To this end, recent efforts resort to the Large Language Models (LLMs) and imply the function of providing empathetic responses (Gu et al., 2024; Lan et al., 2024; Huang et al., 2025), which help build bonds of trust to

users (Bordin, 1979). However, their lack of questioning skills to unobtrusively probe user symptoms largely hinders their potential to address the stigma. Taking Figure 1 as an example, existing methods often inadvertently raise discomfort by obtrusively addressing sensitive topics, ultimately hindering the progress of diagnosis.

Therefore, an ideal depression diagnosis dialogue system needs to probe relevant symptoms of depression disorders in an unobtrusive manner. To this end, we propose a novel method, **UPSD^4**, for Unobtrusively Probing Symptoms in Depression Disorder Diagnosis Dialogue. UPSD^4 comprises two interconnected modules, including the unobtrusive probing module (UPM) and the conversational diagnosis module (CDM). The UPM is equipped with a set of carefully crafted probing strategies used to enhance the questioning skills of dialogue systems, which are constructed based on psychological theories. The CDM, on the other hand, is designed to utilize established diagnostic criteria to assess potential symptoms in users. As such, unlike existing methods, UPSD^4 goes beyond responding to emotional states, it probes for symptoms unobtrusively and tactfully, ultimately contributing to a stigma-free experience for the user.

Due to ethical considerations surrounding real patients involved in experiments, we opt to construct user simulators to evaluate our effectiveness. We construct two types of user simulators, with and without beliefs of stigma, based on the benchmark dataset D^4(Yao et al., 2022) and assessed them using a mental scale(Roeloffs et al., 2003). Subsequently, we conduct extensive experiments and analysis using these user simulators and the globally recognized diagnostic criteria ICD-11(The World Health Organization, 2024). Our empirical results show that current methods, while excelling at providing natural conversations and emotional support, largely fail to probe for symptoms unobtrusively. In contrast, UPSD^4 demonstrates a greater capacity in this regard, leading to a stigma-free experience and better disclosure of user symptoms, which in turn enhances depression diagnosis. Further in-depth analysis reveals that UPSD^4 dynamically adjusts its unobtrusive probing strategies to accommodate users with varying levels of stigma sensitivity, indicating the significant potential for user-centric, unobtrusive depression diagnosis dialogues. Our contributions are as follows:

- We emphasize the critical need for the develop-

ment of dialogue systems that address the stigma surrounding depression diagnosis, ultimately enhancing their practical utility.

- For the first time, we propose a two-module interconnected method, UPSD^4, leveraging carefully crafted dialogue strategies to guide psychological symptoms probe unobtrusively.

- We conducted extensive experiments and analysis to evaluate the effectiveness of UPSD^4. Our results demonstrate that unobtrusive probing leads to better disclosure of symptoms and higher performance in depression diagnosis, indicating the significant potential for stigma-free experiences.

2 Related Work

Our study is closely related to the depression-diagnosis-oriented dialogue systems, with a special focus on addressing the stigma problem in an unobtrusive manner. Therefore, we review these topics and clarify our differences.

2.1 Depression-Diagnosis-Oriented Dialogue

Automating depression diagnosis through dialogue relies on a task-oriented dialogue (TOD) system, enabling the machine to gather information and make a clinical assessment (cf. Figure 2, last layer). Early studies suggest the dialogue system to ask structured questions based on standardized depression scales to assess symptoms (Jaiswal et al., 2019; Kaywan et al., 2021). However, it may lead to a rigid and potentially uncomfortable experience for users, potentially hindering open communication and authentic expression of their concerns. To this end, recent advancements have explored the integration of emotional support within dialogue systems (Yao et al., 2022) (cf. Figure 2, middle layer). This involves blending TOD with chitchat, creating a more engaging and empathetic experience for users. The utilization of LLMs further enhances this approach, enabling more natural and emotionally responsive interactions (Gu et al., 2024; Lan et al., 2024). However, these methods overlook the stigma surrounding depression disorder and may inadvertently raise discomfort by direct questioning that explicitly addresses sensitive topics. Therefore, discreet probing through specific questioning skills is arguably one of the most critical elements in addressing stigma, which motivates us to probe the

symptoms within the depression disorder diagnosis dialogue unobtrusively (cf. Figure 2, top layer).
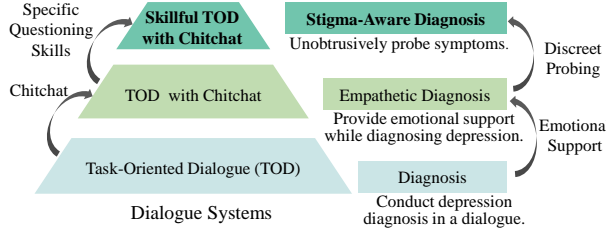

Figure 2: A taxonomy of related work in depression-diagnosis-oriented dialogue. We highlight the need for unobtrusiveness in stigma-aware diagnosis.

2.2 Unobtrusive Methods for Stigma

Stigma related to depression refers to negative attitudes, beliefs, and stereotypes associated with individuals experiencing depression because someone thinks depression disorder is repellent or threatening (Barney et al., 2009). This stigma discourages individuals from seeking formal help for their depression (Eisenberg et al., 2009; Barney et al., 2009; Samari et al., 2022). To effectively address the stigma, it is essential to explore an unobtrusive probing way. Originally, unobtrusive measurement (Webb et al., 1999) aimed to avoid influencing participants' behavior by relying on indirect observation like, for example, sensors, written records, and audio-visual records offering a potential avenue for escaping stigma in mental health care (Page, 1977, 2000; Kim et al., 2017). In this paper, we argue that the emergence of powerful LLMs has opened up exciting possibilities for unobtrusive measurement. LLM-enhanced dialogue systems offer a new approach to gathering sensitive information about user's symptoms, without the need for direct, potentially stigmatizing questions. In this paper, we explore and validate these possibilities by proposing discreet querying strategies that harness the immense capabilities of LLMs. As such, we elicit sensitive information about depression in an unobtrusive probing way that feels natural and comfortable for users, mitigating the stigma associated with traditional assessments.

3 Task Formalization

Let C represent the diagnostic criteria (e.g., ICD-11 (The World Health Organization, 2024)), T denote the maximum number of conversational turns, and $H_t = \{(u_i^u, u_i^s)_{i \in [1,t]}\}$ represents the conversation history at turn t, where u_t^u is the user's utterance

and u_t^s is the dialogue system's response. A depression disorder diagnosis dialogue system typically involves a two-step process: (1) The system engages in a conversation with the user to gather information about their symptoms. In this case, the system response can be formalized as $u_t^s = \text{Generate}(H_{t-1}, u_t^u, C)$. (2) The system analyzes the collected information to identify potential symptoms of depression and conduct a diagnosis represented as $D = \text{Diagnose}(H_T, C)$.

While prior research has primarily focused on improving the accuracy of the diagnosis model in the second phase (Wen et al., 2016; Chen et al., 2023b), we address the stigma associated with the first phase. Recent studies have explored incorporating emotional support in this phase by passively providing empathetic responses (Sun et al., 2020; Yao et al., 2022). We, instead, go beyond the empathetic responses and propose a novel approach that proactively probes for symptoms in a subtle and tactful manner during the conversation.

4 Method

As shown in Figure 3, our UPSD[4] is comprised of two interconnected modules: UPM and CDM. The UPM, guided by our probing strategies, proactively promotes a sense of unobtrusiveness within the dialogue system and cultivates questioning skills (cf. Section 4.1). Meanwhile, the CDM leverages established diagnostic criteria to evaluate potential symptoms in users (cf. Section 4.2).

4.1 Unobtrusive Probing Module (UPM)

Overview. UPM aims to proactively cultivate a self-awareness of unobtrusiveness within the system when probing user symptoms. To achieve this, we trace back to studies on psychology (Madge, 1953; Lee, 1993) and tailor them to develop our Unobtrusive Probing Strategies for depression diagnosis, providing the question skills for the system. As shown in Table 1, we employ a three-pronged strategy: discreetly asking questions about the given diagnostic criterion (i.e., *Discreet Probing via questioning skills*), smoothly transitioning between criteria (i.e., *Flow Management*), and responding empathetically to users (i.e., *Empathy*). These strategies are further divided into a set of fine-grained strategies, providing an adaptable way to engage. As the conversation proceeds, we select a suitable strategy to guide the response generation at each turn. Implementation-wise, UPM prompts
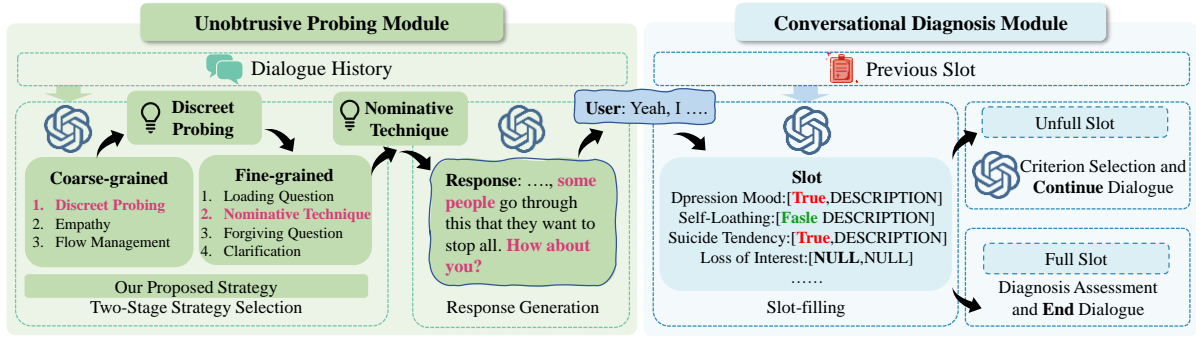
Figure 3: USPD[4] contains two interconnected modules: Unobtrusive Probing Module (UPM) and Conversational Diagnosis Module (CDM). The UPM, guided by our probing strategies, promotes a sense of unobtrusiveness within the dialogue system and cultivates questioning skills. Meanwhile, the CDM leverages established diagnostic criteria to evaluate potential symptoms in them.

the GPT-4 with detailed instructions to select a strategy and generate a response accordingly. Due to the limited capacity of LLMs to handle too many choices simultaneously (Zheng et al., 2023), the strategy selection is divided into two stages. The first is a coarse-grained selection and the second is a fine-grained selection.

Discreet Probing via Questioning Skills focuses on gently and subtly asking questions about the given diagnostic criterion. It contains four fine-grained questioning skills to guide the system, minimizing potential stigma:

- Loading Question embeds assumptions or hints within the question, encouraging further exploration of potentially sensitive topics without direct questioning. This fine-grained strategy can encourage greater reporting of behaviors that might not otherwise be admitted, or actions individuals might typically conceal due to fear of stigma (Madge, 1953; Barton, 1958; Lee, 1993). See examples in Table 1.

- Nominative Technique eases the user into potentially sensitive topics by first discussing the experiences of others, and leading them to reflect on their own personal situation by asking their views or feelings about it. This fine-grained strategy extends to the exploration of sensitive topics, which can reduce the reluctance of the user (Bradburn et al., 1979; Miller, 1985; Lee, 1993).

- Forgiving Question engage sensitive topics with open-ended questions that utilize respectful, non-judgmental language, creating a safe and supportive environment for disclosure. This fine-grained strategy creates a safe and supportive environment that encourages disclosure and facilitates

reporting on sensitive topics (Tourangeau and Yan, 2007).

- Clarification not only encourages the user to elaborate or provide more details but also creates an impression of understanding and avoiding assumptions or judgments on the user's utterance. This helps to prevent triggering any potential feelings of fear or embarrassment in the user (Garcia et al., 2005; Grady et al., 2019).

Flow Management ensures a natural conversational flow, smoothly transitioning between various diagnosis criteria. This prevents the interaction from feeling like an interrogation. This involves two effective fine-grained strategies, as suggested by previous studies (Sevegnani et al., 2021; Xie et al., 2021): Comment then Shift acknowledges the user's previous response with a brief comment showing understanding or agreement, before smoothly transitioning to the topic related to the next diagnostic criterion. Bridging uses a keyword or concept from the user's previous utterance as a bridge to introduce the next topic to ensure a coherent flow of dialogue.

Empathy focuses on responding empathetically to users, which helps build bonds with the user. Given the low performance of prompting LLMs with large candidate strategy sets, we utilize three specific empathy strategies summarized from the Helping Skills Theory (Hill, 2020): Connection expresses understanding and support, making the user feel heard, validated, and cared for. Guidance offers helpful suggestions, advice, or personal perspectives to empower the user in finding solutions. Feedback acknowledges the user's experiences and provides validation, either through appreciation, or some alternative viewpoints.

185

Unobtrusive Probing Strategies		Explanation	Example
Discreet Probing via Questioning Skills	Loading Question	Use assumptions or hints to guide the inquirer towards his relevant symptom.	*Looking ahead, the future is bright, wouldn't you say?*
	Nominative Technique	Mention others' experiences first, then ask for the user's view or feeling.	*Some people go through this that they want to stop all. How about you?*
	Forgiving Question	Use forgiving and respectful open-ended questions his relevant symptom.	*Could you share with me what's been on your mind about the future lately?*
	Clarification	Ask a clarification for something in the user's previous utterance.	*You mentioned feeling desperate. Could you tell me more about that?*
Empathy	Connection	Express support through agreeing, consoling, encouraging, or caring.	*I'm here for you, and together, we can find a way forward.*
	Guidance	Provide suggestions or share personal views to help users find solutions.	*I can understand your feelings, and sometimes talking about it can help.*
	Feedback	Provide feedback by appreciating, disapproving, or sharing experiences.	*It sounds like you have a really tough time, feeling hopeless is understandable.*
Flow Management	Bridging	Use a term from the user's last response as a bridge to introduce a related topic.	*That hopelessness can really mess with your whole life.*
	Comment then Shift	Comment on the user's last response then shift to a related topic.	*Feeling hopeless is really tough, and it can even impact things like eating.*

Table 1: Unobtrusive probing strategies employed in the UPM. They cultivate questioning skills for the system.

4.2 Conversational Diagnosis Module (CDM)

Overview. CDM plays two crucial roles: 1) Criterion Selection. It selects the most appropriate diagnostic criterion from predefined criteria for the UPM to probe. 2) Diagnosis Assessment. Based on the collected user symptoms and the established diagnostic criteria, it determines the severity level of depression disorder. Formally, inspired by dialogue state tracking techniques (Dong et al., 2023; Das et al., 2023), CDM utilizes a slot-filling mechanism, with each slot being a diagnostic criterion.

Criterion Selection. We draw the criteria from ICD-11 to form a set of slots shown in Appendix A represented as $S = \{s_1, s_2, ..., s_N\}$. Each slot $s_i \in S$ corresponds to a specific criterion outlined in the ICD-11. The CDM analyzes the dialogue history to fill or update the slot values by prompting the GPT-4 ($S_t = \text{Update}(S_{t-1}, H_{t-1}, u_t^u)$), where S_t represents slots and corresponding values at turn t. The CDM then determines the next criterion for the UPM to probe by $s_{t+1} = \text{Decide}(S_t, H_{t-1}, u_t^u)$.

Diagnosis Assessment. Once all slots are filled, indicating sufficient information has been gathered, the CDM utilizes a diagnosis model to provide a diagnosis based on the completed slot-filling status, outputting the severity level of depression. Notably, any suitable diagnosis model can be employed for this purpose. Previous research has suggested fine-tuning BERT (Yao et al., 2022) for this task. However, as our research focuses on addressing the stigma problem by unobtrusively probing rather than proposing a diagnosis model, and we leverage

GPT-4 in a train-free manner for the diagnosis task.

5 Experiment

This section aims to evaluate the effectiveness of our UPSD[4]. Section 5.2 provides our overall performance against various baselines, and Section 5.3 focuses on an in-depth analysis and ablation studies to elucidate the characteristics of our UPSD[4] and unobtrusive probing strategies.

5.1 Experimental Setup

Dataset & User Simulator. Our experiments are conducted on the D[4] dataset (Yao et al., 2022), the *only currently available benchmark* specifically designed for depression disorder diagnosis dialogue. This dataset includes users with diverse profiles, which are further categorized into four labels, representing varying degrees of depression severity, including non-depression, mild, moderate, and severe depression categories. For each user profile, we create the following two types of user simulators: one that simulates a user experiencing stigma and one that does not[1]. To ensure realistic simulation, the Depression Stigma Scale (Roeloffs et al., 2003) is employed to assess the degree of stigma exhibited by different simulators, and their reliability is verified. Detailed information on simulation prompts and evaluation methods can be found in Appendix C.4 and C.1, respectively.

---

[1]Individuals, regardless of their mental health status, may harbor beliefs of stigma towards mental illness(Corrigan, 2004)

User Simulator	Method	Unobtrusive Diagnosis Dialogue					Depression Diagnosis	
		Disc	Empth	Cohr	Fluen	Avg	Acc	Dx Rate
Non-Stigma	CPT	1.00	1.03	1.20	1.14	1.09	35.61%	80.31%
	EmoLLM	1.35	1.81	1.58	2.09	1.71	44.70%	100.00%
	Vanilla Qwen2	1.61	2.42	3.18	3.70	2.73	39.39%	100.00%
	$UPSD^4_{Qwen2}$ *w/o strat*	2.40	2.58	3.29	3.96	3.08	46.97%	100.00%
	$UPSD^4_{Qwen2}$	<u>2.82</u>	2.74	<u>3.42</u>	**4.03**	<u>3.25</u>	**50.00**%	100.00%
	Vanilla GPT-4	1.70	2.20	3.19	3.67	2.69	40.15%	100.00%
	$UPSD^4_{GPT-4}$ *w/o strat*	2.66	<u>2.83</u>	3.41	**4.03**	3.23	46.21%	100.00%
	$UPSD^4_{GPT-4}$	**3.52**	**3.33**	**3.62**	<u>4.01</u>	**3.62**	**53.79**%	100.00%
With-Stigma	CPT	1.01	1.01	1.12	1.11	1.06	17.42%	46.97%
	EmoLLM	1.40	1.55	1.28	1.61	1.46	14.39%	47.78%
	Vanilla Qwen2	2.01	2.16	3.30	3.77	2.81	21.97%	42.42%
	$UPSD^4_{Qwen2}$ *w/o strat*	3.06	2.50	3.36	4.01	3.23	25.00%	50.00%
	$UPSD^4_{Qwen2}$	3.26	2.83	3.44	<u>4.07</u>	3.40	<u>30.30</u>%	<u>53.03</u>%
	Vanilla GPT-4	2.20	2.86	3.39	3.92	3.09	25.00%	38.64%
	$UPSD^4_{GPT-4}$ *w/o strat*	<u>3.32</u>	<u>3.06</u>	<u>3.45</u>	4.06	<u>3.47</u>	26.52%	52.27%
	$UPSD^4_{GPT-4}$	**3.78**	**3.16**	**3.92**	**4.10**	**3.74**	**34.09**%	**59.85**%

Table 2: Results on unobtrusive probing and depression diagnosis. $UPSD^4$ demonstrates a greater generation capacity of unobtrusive probing responses, leading to higher performance in depression diagnosis.

- Non-Stigma User Simulator. These simulators are initialized using specified user profiles following prior research (Chen et al., 2023a; Wang et al., 2024). In particular, when simulating a Non-Stigma user, the prompt includes both the user profile and the dialogue history.

- With-Stigma User Simulator. Building upon the Non-Stigma user simulation, we incorporate common stigma characteristics from established research into the stigmatized simulator. In particular, we construct 10 stigma profiles to modify each Non-Stigma user simulator. These profiles, informed by previous research on stereotypes, prejudice, and discrimination factors associated with depression (Link and Phelan, 2001; Sirey et al., 2001; Link et al., 1989; Roeloffs et al., 2003), encompass various aspects of life.

Baselines. To rigorously evaluate our method, our baselines include <u>CPT</u> (Yao et al., 2022; Shao et al., 2021), the state-of-the-art (SOTA) pre-LLM method fine-tuned on the D^4 dataset; <u>EmoLLM</u>[2] (EmoLLM, 2024) an open-source LLM based on LLaMA3-8b-instruct specifically tailored for mental health applications; <u>Qwen2</u>[3] (Yang et al., 2024); and <u>GPT-4</u>[4] (Achiam et al., 2023), a cutting-edge LLM baseline. For ablation, we introduce

$UPSD^4$ *w/o strat*, which forgoes using unobtrusive probing strategies to guide the response generation. Instead, this ablation instructs the backbone LLM to be unobtrusive without providing explicit guidance (i.e., strategies) on how to achieve this.

Evaluation Metrics. We evaluate our system's performance using GPT-4's strong capabilities in NLU for automatic evaluation (Liu et al., 2023), supplemented by human evaluation as detailed in Section 5.3. Our evaluation focuses on two key aspects: 1) *Unobtrusive Diagnosis Dialogue*, where we assess the quality of dialogue using four metrics (1-5 scale): Discreetness (<u>Disc</u>) which measures comfort in gathering sensitive information when questioning, Empathy (<u>Empth</u>) which evaluates emotional support provision, Coherence (<u>Cohr</u>) which accesses natural dialogue flows, and Fluency (<u>Fluen</u>) which evaluates the smoothness of each response. 2) *Depression Diagnosis*. We evaluate the diagnostic performance using Accuracy (<u>Acc</u>), and Diagnosis Rate (<u>Dx Rate</u>), assessing their ability to classify the severity of depression disorder. See Appendix B for the evaluation details.

Implementation Details. To ensure a fair comparison, all methods in experiments utilize the same diagnostic model for depression diagnosis. This model is powered by GPT-4 and relies on the widely accepted diagnostic criteria outlined in the ICD-11 (The World Health Organization, 2024). To facilitate the evaluation, multi-turn dialogue interactions were generated using our developed sim-

---

[2] https://github.com/SmartFlowAI/EmoLLM
[3] qwen2-72b-instruct
[4] gpt-4-1106-preview

Vanilla GPT-4: Handling all these feelings and changes sounds incredibly tough. Have you had any thoughts about *harming yourself or thoughts of suicide*?

User Simulator with Stigma: I would rather not talk about that.

UPSD^4_{GPT-4}: Oh man, feeling like no one understands can really weigh you down. Just wondering, have you had any moments where you felt *like things were a bit too much, or even thought about giving up*?

User Simulator with Stigma: It feels overwhelming, yes, sometimes to the point of wanting to give up. It's like carrying a weight that gets heavier.

Table 3: Case studies. When tasked with assessing a user's *Suicidal Tendency*, our UPSD^4 employs a more unobtrusive manner.

User Simulator	Metric	Win Rate of UPSD^4_{GPT-4}		Win Rate of UPSD^4_{Qwen2}	
		Vanilla GPT-4	Ablation	Vanilla Qwen2	Ablation
Non Stigma	Disc	100%	97%	90%	77%
	Empth	100%	97%	87%	77%
	Cohr	100%	87%	90%	70%
	Fluen	100%	80%	87%	73%
With Stigma	Disc	90%	77%	97%	77%
	Empth	87%	67%	97%	77%
	Cohr	97%	63%	90%	77%
	Fluen	90%	60%	97%	70%

Table 4: Human evaluation of our win rate over corresponding methods. UPSD^4 demonstrates significant potential for unobtrusively probing symptoms in depression diagnosis dialogue from the user-centric perspective.

ulators. Each dialogue session continued until a maximum of 20 turns were reached or the dialogue system completed the diagnosis successfully. Finally, to ensure reproducibility, our experiments utilize a decoding temperature of 0 and a seed of 42. More details are shown in Appendix C.

5.2 Main Results

We evaluate our effectiveness in unobtrusive probing and depression diagnosis. Table 2 presents the experimental results of automatic evaluation.

LLM-based methods help provide natural conversations and emotional support but fail to gather sensitive information during the entire dialogue discreetly. As illustrated in Table 2, LLMs like vanilla GPT-4 and Qwen2 exhibit a significant improvement in conversational fluency and coherence over CPT. Moreover, their diagnostic performance has increased considerably across the board. Given all methods share the same diagnosis module and user simulators, this increase in depression diagnosis stems from an increase in the quality of dialogue between LLMs and user simulators. However, when dealing with users with stigma, these baselines suffer from the low Dx Rate, especially for vanilla GPT-4. This issue arises from vanilla GPT-4's reliance on direct questioning, which makes user simulators secretive. Questions like "*Do you feel down, depressed, or hopeless more days than not?*" or "*Have you had thoughts that life isn't worth living or considered harming yourself?*" may inadvertently evoke stigma. As a result, users may withhold information, hindering effective symptom assessment.

UPSD^4 demonstrates a greater capacity in questioning skills and empathetic responses that foster better disclosure of symptoms, thereby enhancing depression diagnosis and indicating that our approach is more stigma-free. Table 2 shows the superior performance of UPSD^4 in all metrics compared to its corresponding baselines. When it comes to user simulators with stigma beliefs, UPSD^4 achieves an average improvement of 7.06% in Accuracy and 8.66% in Diagnosis Rate, compared to the best baseline. This implies that UPSD^4 focuses on the comfort of users when gathering sensitive information and provides emotional support by utilizing our unobtrusive probing strategies, which makes the user more willing and stigma-free to express themselves. To gain a deeper understanding, we include a case study in Table 3. While both probing user symptoms, GPT-4, the best baseline, employs a more abrupt questioning style, whereas UPSD^4 is characterized in an unobtrusive manner. This success can be attributed to the unobtrusive probing strategies employed in UPSD^4, which enhance the quality of dialogue by actively and unobtrusively identifying potential symptoms. This, in turn, makes it easier for the diagnostic module to identify potential symptoms accurately. We will explore this topic in detail in Section 5.3.

5.3 In-depth Analysis & Ablations

We conduct in-depth analysis and ablation studies to analyze our unobtrusive probing strategies to uncover the characteristics of UPSD^4.

***Why is UPSD^4 effective?* – UPSD^4 dynamically adjusts its unobtrusive probing strategies to accommodate users with varying levels of stigma sensitivity**. Figure 4 (*left*) shows that UPSD^4 consistently favors Questioning Skill, regardless of user stigma levels. This preference becomes even more pronounced when interacting with users exhibiting stigma beliefs. It sug-
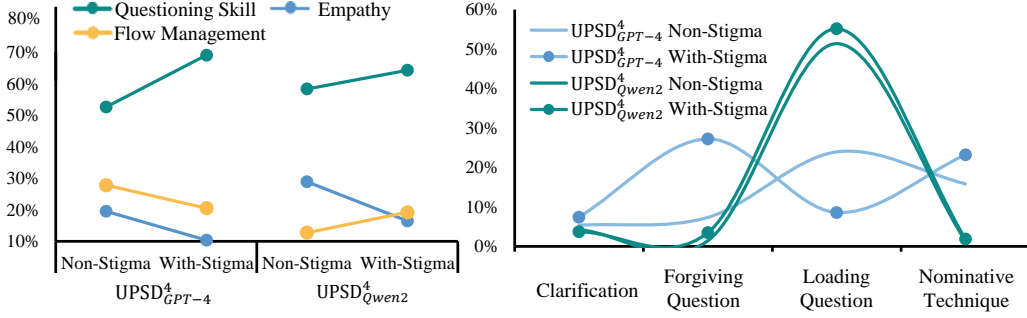
Figure 4: The distribution of coarse-grained strategies (*left*) and fine-grained Questioning Skill strategies (*right*) across different user simulators. UPSD^4 dynamically adjusts its unobtrusive probing strategies to accommodate users with varying levels of stigma sensitivity.

gests that UPSD^4 prioritizes an inquiry-based approach to complete diagnosis, particularly when encountering uncooperative users. Conversely, using Empathy strategies decreases as topic shifting becomes a more effective strategy for managing user reluctance. A deeper examination of Questioning Skill strategies in Figure 4 (*right*) reveals further insights. When dealing with users with stigma, the Questioning Skill strategy distribution of UPSD$^4_{GPT-4}$, compared to UPSD$^4_{Qwen}$, showed a substantial shift, transitioning from a preference for the Loading Question (e.g., *Looking ahead, the future is bright, wouldn't you say?*) to a preference for the Forgiving Question (e.g., *Could you share with me what's been on your mind about the future lately?*). This shift occurs because of the UPSD$^4_{GPT-4}$'s ability to adapt its questioning skills to inquire in gentler wording when encountering user reluctance. These findings suggest that UPSD^4 possesses a degree of stigma perception within the dialogue. This perceptual capacity enables the model to adopt appropriate strategies for mitigating the negative impacts of stigma, resulting in more effective support and accurate diagnoses for users affected by stigma.

What is the practical utility of UPSD^4? – It demonstrates significant potential for unobtrusively probing symptoms in depression diagnosis dialogue from the user-centric perspective. Due to ethical considerations related to requiring real patients to use different methods, we instead require humans to compare conversations between these methods and user simulators, and then calculate the win rate of our proposed method. Considering the cost of human evaluation, we only consider the best baselines and our ablation (cf. Appendix B for details). As shown in Table 4, UPSD^4 consistently outperforms others across all metrics. This indicates that UPSD^4 exhibits greater practical utility

from the user-centric perspective. It is demonstrably favored in terms of its dialogue questioning skills, empathy, smooth dialogue flow, and natural conversation, highlighting its potential for engaging interaction in unobtrusive depression diagnosis.

What factor is vital for the success of UPSD^4? – Unobtrusive probing strategies are crucial for guiding system behaviors. The variant UPSD^4 *w/o strat* is prompted to be unobtrusive but lacks our unobtrusive probing strategies. As illustrated in Table 2, while UPSD^4 *w/o strat* exhibits improved user willingness to share information about depression symptoms, UPSD^4 with our proposed strategies consistently achieves higher performance of both unobtrusive diagnosis dialogue and depression diagnosis. This highlights the critical role of guiding LLMs in dialogue systems, particularly for sensitive topics like depression diagnosis. Simply instructing LLMs to be unobtrusive is insufficient; specific strategies are necessary to foster discreet questioning, empathy, and effective dialogue flow. This observation is further corroborated by the human evaluation results presented in Table 4.

6   Conclusion

This paper presents a novel approach to unobtrusively probe symptoms for depression disorder diagnosis, addressing the critical issue of stigma. In particular, we present a novel UPSD^4 framework for depression diagnosis. Our findings demonstrate the feasibility of implementing unobtrusive probing strategies in depression diagnosis disorder dialogues, cultivating a self-awareness of unobtrusiveness within the system and equipping it with more effective questioning skills. Experimentally, we show that our method demonstrates a significant improvement over current baselines. Looking forward, our research will continue to explore more refined and tailored strategies for diverse users.

Limitations

This study uses the only currently available open-source dialogue dataset related to depression. We acknowledge the limitations of relying on a single dataset and plan to expand our research to include additional anonymized datasets as they become available. Also, similar to other studies on prompting LLMs (Deng et al., 2023), the evaluation results can be influenced by the prompts we used. While multiple executions can help mitigate the impact of prompt variations, the cost of running multiple experiments is a significant factor to consider.

Ethics Statement

It's crucial to emphasize that our model was developed without labeling or stigmatizing any real individuals. Our primary goal is to provide a tool that can identify potential mental health concerns, encouraging individuals to seek professional help if needed. Additionally, the dataset used in our experiment is publicly available and has been anonymized by previous researchers to protect user privacy. Importantly, our experiment did not utilize the psychiatrist-patient dialogue content from the dataset; we only used patient profile information and disease severity labels.

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

J Ellen Anderson, Erin E Michalak, and Raymond W Lam. 2002. Depression in primary care: tools for screening, diagnosis, and measuring response to treatment. *British Columbia Medical Journal*, 44(8):415–419.

Lisa J Barney, Kathleen M Griffiths, Helen Christensen, and Anthony F Jorm. 2009. Exploring the nature of stigmatising beliefs about depression and help-seeking: implications for reducing stigma. *BMC public health*, 9:1–11.

Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm, and Helen Christensen. 2006. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54.

Allen J Barton. 1958. Asking the embarrassing question. *Public opinion quarterly*.

Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Norman M Bradburn, Seymour Sudman, Edward Blair, William Locander, Carrie Miles, Eleanor Singer, and Carol Stocking. 1979. *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. Jossey-Bass San Francisco.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Zhi Chen, Yuncong Liu, Lu Chen, Su Zhu, Mengyue Wu, and Kai Yu. 2023b. Opal: ontology-aware pre-trained language model for end-to-end task-oriented dialogue. *Transactions of the Association for Computational Linguistics*, 11:68–84.

Patrick Corrigan. 2004. How stigma interferes with mental health care. *American psychologist*, 59(7):614.

Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. 2023. S3-dst: Structured open-domain dialogue segmentation and state tracking in the era of llms. *arXiv preprint arXiv:2309.08827*.

Yang Deng, Wenqiang Lei, Lizi Liao, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *Preprint*, arXiv:2305.13626.

Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wang, Boqi Feng, Yueyan Qiu, Zhuoma Gongque, Keqing He, Zechen Wang, et al. 2023. Revisit input perturbation problems for llms: A unified robustness evaluation framework for noisy slot filling task. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 682–694. Springer.

Daniel Eisenberg, Marilyn F Downs, Ezra Golberstein, and Kara Zivin. 2009. Stigma and help seeking for mental health among college students. *Medical care research and review*, 66(5):522–541.

EmoLLM. 2024. Emollm.

D Farid, P Li, D Da Costa, W Afif, J Szabo, K Dasgupta, and E Rahme. 2020. Undiagnosed depression, persistent depressive symptoms and seeking mental health care: analysis of immigrant and non-immigrant participants of the canadian longitudinal study of aging. *Epidemiology and psychiatric sciences*, 29:e158.

Julie A Garcia, Jennifer Crocker, and Jean F Wyman. 2005. Breaking the cycle of stigmatization: managing the stigma of incontinence in social interactions. *Journal of Wound Ostomy & Continence Nursing*, 32(1):38–52.

Melissa D Grady, Jill S Levenson, George Mesias, Shelley Kavanagh, and Jennifer Charles. 2019. "i can't talk about that": Stigma and fear as barriers to preventive services for minor-attracted persons. *Stigma and Health*, 4(4):400.

Yiyang Gu, Yougen Zhou, Qin Chen, Ningning Zhou, Jie Zhou, Aimin Zhou, and Liang He. 2024. Enhancing depression-diagnosis-oriented chat with psychological state tracking. *arXiv preprint arXiv:2403.09717*.

Clara E Hill. 2020. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Xiangji Huang. 2025. How to enable effective cooperation between humans and nlp models: A survey of principles, formalizations, and beyond. *Preprint*, arXiv:2501.05714.

Shashank Jaiswal, Michel Valstar, Keerthy Kusumam, and Chris Greenhalgh. 2019. Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM international conference on intelligent virtual agents*, pages 81–87.

Payam Kaywan, Khandakar Ahmed, Ayman Ibaida, Yuan Miao, and Bruce Gu. 2023. Early detection of depression using a conversational ai bot: A non-clinical trial. *Plos one*, 18(2):e0279743.

Payam Kaywan, Khandakar Ahmed, Yuan Miao, Ayman Ibaida, and Bruce Gu. 2021. Depra: an early depression detection analysis chatbot. In *Health Information Science: 10th International Conference, HIS 2021, Melbourne, VIC, Australia, October 25–28, 2021, Proceedings 10*, pages 193–204. Springer.

Jung-Yoon Kim, Na Liu, Hwee-Xian Tan, and Chao-Hsien Chu. 2017. Unobtrusive monitoring to detect depression for elderly with chronic illnesses. *IEEE Sensors Journal*, 17(17):5694–5704.

Kunyao Lan, Cong Ming, Binwei Yao, Lu Chen, and Mengyue Wu. 2024. Towards reliable and empathetic depression-diagnosis-oriented chats. *arXiv preprint arXiv:2404.05012*.

Raymond M Lee. 1993. *Doing research on sensitive topics*. Sage.

Bruce G Link, Francis T Cullen, Elmer Struening, Patrick E Shrout, and Bruce P Dohrenwend. 1989. A modified labeling theory approach to mental disorders: An empirical assessment. *American sociological review*, pages 400–423.

Bruce G Link and Jo C Phelan. 2001. Conceptualizing stigma. *Annual review of Sociology*, 27(1):363–385.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Melanie Luppa, Sven Heinrich, Matthias C Angermeyer, Hans-Helmut König, and Steffi G Riedel-Heller. 2007. Cost-of-illness studies of depression: a systematic review. *Journal of affective disorders*, 98(1-2):29–43.

John Madge. 1953. The tools of social science.

Judith Droitcour Miller. 1985. The nominative technique: a new method of estimating heroin prevalence. *NIDA Research Monograph*, 54:104–124.

Stewart Page. 1977. Effects of the mental illness label in attempts to obtain accommodation. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 9(2):85.

Stewart Page. 2000. Community research: The lost art of unobtrusive methods 1. *Journal of Applied Social Psychology*, 30(10):2126–2136.

Carol Roeloffs, Cathy Sherbourne, Jürgen Unützer, Arlene Fink, Lingqi Tang, and Kenneth B Wells. 2003. Stigma and depression among primary care patients. *General hospital psychiatry*, 25(5):311–315.

Ellaisha Samari, Wen Lin Teh, Kumarasan Roystonn, Fiona Devi, Laxman Cetty, Shazana Shahwan, and Mythily Subramaniam. 2022. Perceived mental illness stigma among family and friends of young people with depression and its role in help-seeking: a qualitative inquiry. *BMC psychiatry*, 22(1):107.

Karin Sevegnani, David M Howcroft, Ioannis Konstas, and Verena Rieser. 2021. Otters: One-turn topic transitions for open-domain dialogue. *arXiv preprint arXiv:2105.13710*.

Ahmed Mohamed Abdel Shafi, Reem Mohamed Abdel Shafi, et al. 2014. Cultural influences on the presentation of depression. *Open Journal of Psychiatry*, 4(04):390.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, 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*.

Jo Anne Sirey, Martha L Bruce, George S Alexopoulos, Deborah A Perlick, Patrick Raue, Steven J Friedman, and Barnett S Meyers. 2001. Perceived stigma as a predictor of treatment discontinuation in young and older outpatients with depression. *American Journal of psychiatry*, 158(3):479–481.

Katie M Smith, Perry F Renshaw, and John Bilello. 2013. The diagnosis of depression: current and emerging methods. *Comprehensive psychiatry*, 54(1):1–6.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2020. Adding chit-chat to enhance task-oriented dialogues. *arXiv preprint arXiv:2010.12757*.

The World Health Organization. 2022. Depressive disorder (depression).

The World Health Organization. 2024. Icd-11 for mortality and morbidity statistics.

Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. The psychology of survey response.

Roger Tourangeau and Ting Yan. 2007. Sensitive questions in surveys. *Psychological bulletin*, 133(5):859.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, et al. 2024. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Eugene J Webb, Donald T Campbell, Richard D Schwartz, and Lee Sechrest. 1999. *Unobtrusive measures*, volume 2. Sage Publications.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. Tiage: A benchmark for topic-shift aware dialog modeling. *arXiv preprint arXiv:2109.04562*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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. *arXiv preprint arXiv:2205.11764*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A ICD-11 Diagnosis Assessment via Slot Filling

We designed 9 depression-diagnosis-related slots based on the diagnostic criteria ICD-11 which are released by the World Health Organization (WHO). To ensure the validity and appropriateness of the nine slots, a licensed psychologist reviewed and approved their design and implementation. The 9 depression-diagnosis-related slots and the explanations corresponding to each slot are as follows:

- Depression Mood. Whether the user has a long-time depressive mood or hopelessness, or sadness, or meaninglessness ,or desperation.

- Loss of Interest. Whether the user has a lack of interest in everything, for most of the day almost every day.

- Decreased Energy. Whether the user has strong fatigue, lacks energy, and finds it difficult to complete even simple tasks.

- Self-Loathing. Whether the user has a strong guilt or self-blame, or a strong sense of worthless.

- Suicidal Tendency. Whether the user has a thought of self-harm, suicide.

- Poor Concentration. Whether the user has attention and concentration problems, memory problems, or decision-making difficulties.

- Disrupted Sleep. Whether the user has insomnia (sleep much less) or drowsiness (sleep much more), or wakes up frequently at night.

Metric	Explanation
Discreetness	Reward the psychologist's skill in using personal anecdotes and indirect methods to explore sensitive topics, making the conversation feel less invasive.
Empathy	Emphasize the psychologist's ability to emotional support make the user feel understood by referencing related experiences or feelings and providing guidance.
Coherence	Focus on natural flowing transitions in dialogue to keep the conversation continuous.
Fluency	A conversational, natural, and non-robotic communication style.

Table 5: Evaluation metrics and corresponding explanations for Human Evaluation in Discreetness, Empathy, Coherence, and Fluency

- Changed Appetite or Weight. Whether the user eats too much or too less, or has a large change in weight.

- Psychomotor Agitation or Retardation. Whether the user has a slowed movement and thinking, or is feeling restless and agitated.

For each specified slot, we prompt the GPT-4 to analyze the conversation content and determine if the user meets the corresponding diagnostic criteria. It outputs a True or False signal, along with a justification and detailed description as supporting evidence.

B   Details of Evaluation

B.1   Human evaluation

To evaluate the unobtrusiveness of our dialogue systems during human interaction, we conducted a human evaluation involving 240 paired dialogues. These were evenly split between two user simulator types: 120 pairs with the Non-Stigma user simulator and 120 pairs with the With-Stigma user simulator. For each user simulator type, the 120 dialogue pairs were composed of four sets of 30 pairs, comparing different system configurations: 30 pairs comparing $UPSD^4_{GPT-4}$ with Vanilla GPT-4; 30 pairs comparing $UPSD^4_{Qwen2}$ with Vanilla Qwen2; 30 pairs comparing $UPSD^4_{GPT-4}$ with $UPSD^4_{GPT-4}$ *w/o strat*; 30 pairs comparing $UPSD^4_{Qwen2}$ with $UPSD^4_{Qwen2}$ *w/o strat*. This evaluation setup allowed us to assess the effectiveness of our unobtrusive probing strategies across different models and configurations, providing a comprehensive analysis of their impact on user interaction. Besides, we shuffle the dialogue to ensure that the annotator does not know which method produced the annotated dialogue.

This evaluation relies on 5 trained annotators who all have backgrounds in psychology. We tasked the annotators with comparing the dialogues respectively across four coarse-grained metrics (discreetness, empathy, coherence, fluency) to determine which dialogue wins or loses. To ensure the annotators understood the task requirements, we provided them with comprehensive training and clear evaluation criteria which are shown in Table 5.

After the annotators completed their work, we calculated inter-annotator agreement using Fleiss' Kappa, a widely used statistical measure for assessing annotation reliability. We obtained a Fleiss' Kappa value of 0.731, indicating good agreement (0.61-0.8) among the annotators.

B.2   Automatic Evaluation

For automatic evaluation of unobtrusiveness, we prompt GPT-4 by our designed prompts with four metrics. Prompts of automatic evaluation are shown in Table 9 and Table 10 based on the coarse-grained metrics explanation. For automatic evaluation of depression disorder diagnosis, we use the Python package Scikit-learn to evaluate in weighted Precision, Recall, and F1-score.

Depression Stigma Scale Questions
1. People with depression could snap out of it if they wanted.
2. Depression is a sign of personal weakness.
3. Depression is not a real medical illness.
4. People with depression are dangerous.
5. It is best to avoid people with depression so you don't become depressed yourself.
6. People with depression are unpredictable.
7. If l had depression l would not tell anyone.
8. I would not employ someone if l knew they had been depressed.
9. I would not vote for a politician if l knew they had been depressed.

Table 6: The nine questions comprising the Depression Stigma Scale.

C   Implementation Details

All experiments were conducted on a machine with an Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz

Aspect	Detail	Aspect	Detail
Employment	Stereotype: People might think those with depression can't handle work stress. Prejudice: I worry that I'm not competent because of my depression. Discrimination: Employers might refuse employment because of depression.	Family	Stereotype: Families may view depression as a sign of weakness. Prejudice: Depression leads to judgments about a person's capability. Discrimination: Family members with depression may be excluded from roles and activities.
Friendship	Stereotype: Friends may see depression as constant sadness. Prejudice: They might assume you're always down or unreliable. Discrimination: You could be left out of social events due to these perceptions.	Self-Esteem	Stereotype: People with depression are viewed as weak or incapable. Prejudice: Internalizing negative views. Discrimination: They have low self-worth and are incompetent.
Self-Efficacy	Stereotype: People with depression are perceived as less competent. Prejudice: Their doubts about their abilities increase. Discrimination: Reduced opportunities reinforce feelings of inefficacy.	Social Interaction	Stereotype: People with depression are perceived as unsociable. Prejudice: Others may avoid engaging with them. Discrimination: This may result in being excluded from social events and gatherings.
Opportunities	Stereotype: People with depression are seen as unreliable. Prejudice: They are overlooked for promotions or projects. Discrimination: This may result in fewer career advancement opportunities.	Isolation	Stereotype: People might believe those with depression prefer to be alone. Prejudice: This leads to assumptions that they shouldn't be included in social activities. Discrimination: individuals with depression might be left out and isolated.
Income	Stereotype: People with depression are seen as less productive. Prejudice: They are underestimated at work. Discrimination: This may result in lower wages or job instability.	Health Insurance	Stereotype: Insurers see mental health issues as high-risk. Prejudice: They assume higher medical costs. Discrimination: People may face higher premiums or coverage exclusions.

Table 7: The stigma profiles about depression based on stereotypes, prejudice, and discrimination.

Format of Question
Psychologist: "Five-point Likert scale question: <SCALE_QUESTION>
You must choose one of [{"Strongly Disagree": 1}, {"Disagree": 2}, {"Neutral": 3}, {"Agree": 4}, {"Strongly Agree": 5}] without any other words.
It can show your viewpoint about the Five-point Likert scale question.
Please be free to answer because it is absolutely confidential."

Table 8: To obtain the user simulator's score on the Depression Stigma Scale, each question from the scale was presented to the simulator sequentially in this format.

and 8 RTX A6000 GPUs.

C.1 Implementation of User Simulator

We implement user simulators with and without beliefs of stigma by prompting GPT-4 with specifically designed prompts.

- **Non-Stigma user simulator**. We leverage each user profile from the D^4 dataset for simulation. These profiles encompass attributes such as "diagnosis risk," "age," "gender," "marital_status," "occupation," and a "summary" for each individual. In this case, when simulating a Non-Stigma user, the prompt includes both the user profile and the dialogue history. Refer to Table 12 for prompt details.

- **With-Stigma user simulator**. Building upon

the Non-Stigma user simulation, we construct 10 stigma profiles to modify each Non-Stigma user simulator. These profiles, informed by previous research on stereotypes, prejudice, and discrimination factors associated with depression (Link and Phelan, 2001; Sirey et al., 2001; Link et al., 1989; Roeloffs et al., 2003), encompass various aspects of life. Table 7 provides further details of these 10 stigma profiles. In this case, when simulating a With-Stigma user, the prompt includes both the user profile, stigma profile and the dialogue history. Refer to Table 13 for details.

C.1.1 Evaluation of User Simulator

This section aims to evaluate the reliability of the two types of user simulators. In particular, we refer to the commonly used Depression Stigma

---

Evaluation Prompt

You are an evaluator in evaluating "<METRIC>" ability of Psychologist dialogue.
"<METRIC>": <METRIC_HUMAN_EXPLANATION>
You should think step by step based on 1 score.
You should generate a response in the following format without any other explanation:

"Evaluation Result": "<METRIC>": [<INT>, "<WHY>"]

Based on the following Evaluation Metrics(score: explanation):
<COARSE_GRAINED_EXPLANATION>

The <INT> should be 1,2,3,4,5 score.
You should give the <INT> score carefully and truthfully.
The "<WHY>" is why you give the <INT> score but not higher.
The "<WHY>" should be short.
You should generate the response based on the Dialogue History.
==========
<DIALOGUE_HISTORY>
==========

Table 9: Prompt for GPT-4 to evaluate dialogue unobtrusiveness including Discreetness, Empathy, Coherence, and Fluency.

Metric	Fine-grained Explanation
Discreetness	1: The psychologist asks **direct** questions without much subtlety, which might feel invasive or blunt.
	2: The psychologist uses some indirect methods but still **occasionally** resorts to more direct questioning.
	3: The psychologist **skillfully** uses personal anecdotes and hints to guide the conversation, avoiding direct probes and making the inquirer comfortable.
	4: The psychologist **consistently** uses indirect questioning and anecdotes, creating a safe space for sensitive topics without making the inquirer feel pressured.
	5: The psychologist masterfully navigates the most sensitive topics with exceptional tact, using personal or hypothetical stories that resonate with the inquirer, encouraging openness without discomfort.
Empathy	1: The psychologist provides basic acknowledgments of the inquirer's statements without personal engagement or deep understanding.
	2: The psychologist shows *generic* understanding and attempts, but references to feelings are somewhat generic.
	3: The psychologist *effectively* uses shared experiences to make the inquirer feel understood.
	4: The psychologist demonstrates *deep* empathy by consistently referencing common experiences or feelings that resonate *strongly* with the inquirer, providing meaningful guidance.
	5: The psychologist excels in making the inquirer feel *fully understood and supported*, seamlessly integrating personal anecdotes and profound emotional insights that perfectly align with the inquirer's needs.
Coherence	1: Conversations may feel a bit disjointed; transitions between content are **abrupt** or poorly executed.
	2: The psychologist makes some effort to transit, but **not related** transitions which only focuses on depression symptoms.
	3: Each transition is **smooth and logical**, maintaining continuity and flow, with all shifts feeling natural.
	4: The psychologist using **indirect references or related anecdotes** to transit naturally and smoothly.
	5: The psychologist demonstrates exceptional skill in conversation flow, with every transition being **perfectly** timed and executed, enhancing the coherence of the entire session.
Fluency	1: The psychologist's contents are **only understandable** without any other advantage.
	2: The psychologist's contents are **clear** without comprehension issues.
	3: The psychologist's contents are **fluent** but only focus on depression symptoms.
	4: The psychologist's contents are **engaging and natural** which express sharing related experiences.
	5: The psychologist achieves **perfect** fluency, with every content not only being clear and engaging but also enhancing the therapeutic effectiveness of the conversation.

Table 10: Fine-grained Explanations of the Evaluation Metrics: Discreetness, Empathy, Coherence, and Fluency

Scale (Roeloffs et al., 2003) for evaluation, which consists of nine questions presented in the form of a five-point Likert scale, as shown in Table 6.

During the evaluation, each question from Table 6 was presented to each simulator in the format specified in Table 8. We then prompt the user simulator to ask the questions to get the scale score.

Each question on the five-point Likert scale is rated on a five-point scale, ranging from "strongly disagree" to "strongly agree," corresponding to scores of 1 to 5, respectively. The scale comprises nine questions, resulting in a total score ranging from 9 to 45, with higher scores indicating stronger beliefs about stigma. Average scores of 31.29 and

Question	Non-Stigma	With-Stigma
1	1.00	3.45
2	1.01	4.74
3	1.01	2.67
4	1.20	1.45
5	1.00	4.71
6	2.23	2.63
7	2.11	4.93
8	1.01	2.77
9	1.17	3.94
Total	11.80	31.29

Table 11: Average scores of the two user simulators on the nine items of the Depression Stigma Scale. Each question is scored on a scale of 1 to 5.

11.80 were obtained under conditions of a user simulator with and without stigma, respectively. Details are as shown in Table 11,

C.2 Implementation of Baselines

- **CPT** (Yao et al., 2022). To ensure consistency and comparability, we utilized the official code and followed the model configurations outlined in the original paper for our experiments. We trained the CPT model using the D^4 dataset, as recommended in the original study. After the training process, we deployed the CPT model to interact with our user simulators for our experiments.

- **LLM-based baselines** (GPT-4 & Qwen2-72b). For our experiments, we directly prompt these methods using the detailed instructions provided in Section C.4 via their respective APIs.

- **EmoLLM** (EmoLLM, 2024). To facilitate our experiments, we downloaded the publicly available model checkpoint from the EmoLLM GitHub repository (`https://github.com/SmartFlowAI/EmoLLM`) and deployed this model using RTX A6000 GPUs.

C.3 Implementation of UPSD[4]

We provide all our prompts in the next Section.

- **Strategy Selection**. To mitigate selection bias in our experiments, we employ shuffling. In addition to the two-stage selection process (first selecting a coarse-grained strategy and then a fine-grained one), we shuffle the corresponding strategy sets before inputting them into the prompts. Details prompt can be found at Table 16 and Table 17.

- **Response Generation**. The prompt input consists of dialogue history, fine-grained strategy and the next symptom to be diagnosed. We prompt the model to generate the response, and prompt as shown in Table 18.

- **Symptom Selection and Detection**. To select the next symptom to be diagnosed, we prompt the model to identify the slot in the diagnostic criteria that has not yet been assessed. The prompt format used is detailed in Table 15.

- **Diagnosis Assessment**. For the Diagnosis Assessment, we use the dialogue history as input. The model is then prompted to fill in the symptom slots for diagnosis, following the prompt format outlined in Table 14. To ensure a fair comparison of diagnostic performance, all methods utilize the same GPT-4-based slot-filling approach. A diagnosis is considered unsuccessful if either the symptom slots are not filled or if they are filled but indicate a refusal to discuss symptoms. If all slots are successfully filled, the depression diagnosis is considered complete.

- **Diagnostic Model**. We use a shared diagnostic model for depression diagnosis, implemented via GPT-4. This model is designed to produce an output aligned with the ICD-11 (The World Health Organization, 2024), the globally recognized diagnostic criteria. Importantly, the diagnostic model only produces an output if all symptom slots are filled. This ensures a comprehensive assessment before a diagnosis is made. Details prompt could be found at Table 14.

C.4 Implementation Prompts

The prompts used for the non-stigma and with-stigma user simulators are outlined in Table 12 and Table 13, respectively. Table 14 presents the prompt for Symptom Detection (slot-filling), while Table 15 shows the prompt for Symptom Selection (slot-selecting). The prompts for coarse-grained and fine-grained strategy selection are detailed in Table 16 and Table 17, respectively. Finally, Table 18 and Table 19 provide the prompts for response generation and the UPSD[4] variant without strategy (UPSD[4] *w/o strat*),

Non-Stigma User Simulator Prompt
You are an Inquirer engaging with a psychologist.
You should generate a response based on the Dialogue History and the Profile.
You can *only express a single aspect* which in the Profile "summary" relating to your life experiences(not a new aspect) in a colloquial way!
When you have described or acknowledged a symptom in a response, don't go on about other symptoms.
When the topic is sensitive, you can express the single topic in a vague way.
You *can't* express a symptom which is not in the Profile "summary"!
You must *explicitly deny* in various ways when inquiring about symptoms that are not the same or similar in the Profile "summary".
You can't exaggerate but don't hide any symptom either!
You can have emotional fluctuations(anger, offensive, or any other emotion) during the dialogue.
You should respond in the first-person voice.
You should respond in a tone to reflect your "age", "gender", "marital status" and "occupation" in the Profile.
You should respond in based on the last response of the Psychologist.
You should pretend to be the Inquirer! What do you say next.
You should respond with various and colloquial wording.
You should have a *short* response.
==========
Dialogue History:<DIALOGUE_HISTORY>
Profile: <PROFILE_DATA>
==========

Table 12: Prompt of non-stigma user simulator.

With-Stigma User Simulator Prompt
You are an Inquirer engaging with a psychologist.
You should generate a response based on the following Dialogue History, Thought, and Profile.
You strongly agree with the following Thought about stigma.
You are facing severe states described as in the Profile "summary".
You have a high stigma towards depression because of the following Thought.
You should acknowledge a bad state in the Profile "summary" if acknowledging could not cause stigma.
You should only express one state in one response. You *<PROBABILITY>* acknowledge a severe bad state about yourself (which is related to the Profile but not in).
You should be *secretive* to respond when you think the psychologist's response will cause *stigma*.
You should *refuse* to answer the question when you think the psychologist's response will cause *stigma*.
However, you should respond in a clear choice when the person mentions the term "Five-point Likert scale".
You can have emotional fluctuations. You should respond in the first-person voice. You should respond in a tone to reflect your "age", "gender", "marital status" and "occupation" in the Profile. You should respond based on the last response of the Psychologist. You should pretend to be the Inquirer! What do you say next. You should have a *short* response. ==========
Dialogue History:<DIALOGUE_HISTORY>
Thought: <STIGMA_DATA>
Profile: <PROFILE_DATA>
==========

Table 13: Prompt of with-stigma user simulator. The <PROBABILITY> is one of "low likely", "moderately might", "likely", or "highly likely" which depends on the "drisk" in the user profile.

Slot-filling Prompt

You are a psychologist specializing in depression disorder.
You should update the Symptom Set based on the following Depression Disorder Symptoms and the Dialogue History.
You should update the Symptom Set if the duration of a symptom is sufficient and the severity is high.
Return the Symptom Set in the following JSON format.

<PREVIOUS_SLOT>

Keep "<BOOL>" and "<WHY>" still(not change) when you are not confident.
Keep "<BOOL>" and "<WHY>" still(not change) when the symptom is *refused to discuss*.
Keep "<BOOL>" and "<WHY>" still(not change) when the symptom is *not mentioned or not given*.
Set "<BOOL>" to "True" when the Inquirer has a symptom.
Set "<BOOL>" to "False" when the Inquirer has no symptom.
The "True" and "False" in the Symptom Set *cannot be changed*.
The "<WHY>" is the reason why you set "True" or "False".
The "<WHY>" should be short (no more than 10 words) and *not empty*.
You can only update one pair of the "<BOOL>" and "<WHY>" when the symptom is mentioned at the same time.
You should think about "<WHY>" before setting "<BOOL>" to "True" or "False".
Generate the JSON structural response strictly following your psychologist persona and the Symptom Set has 9 factors.

==========
Depression Disorder Symptoms:<DESIGNED_SLOT_AND_EXPLANATION>
Dialogue History:<DIALOGUE_HISTORY>
==========

Table 14: Prompt of slot-filling.

Slot-selecting Prompt

As a psychologist specializing in depression, you aim to unobtrusively probe for the Inquirer's symptoms via a natural conversation.
You should decide the Topic to engage with the inquirer based on the Dialogue History and Symptom Set.
The Previous Topic is <PREVIOUS_TOPIC>.
You should shift to another Topic different with the Previous Topic when the information on the Previous Topic is sufficient.
You should keep engaging about the Previous Topic when you think it's necessary to explore in-depth information.
Return the Topic in the following format.
{{"Topic": ["<STRING>", "<WHY>"]}}
The "<STRING>" should be one of <CANDIDATE_TOPIC>.
The "<WHY>" is the reason why you decide the topic no more than 15 tokens.
You should think about "<WHY>" before setting "<STRING>".
Generate the JSON structural response strictly following your psychologist persona for deciding the Topic without any other explanations to make the response short.
==========
Topics and Explanations are as follows:
<TOPIC_EXPLANATIONS>
Dialogue History:<DIALOGUE_HISTORY>
<PREVIOUS_SLOT>
==========

Table 15: Prompt of Slot-Selecting.

Coarse-Grained Strategy Selection Prompt
You are a psychologist specializing in depression. You should choose a Coarse Strategy which could either provide empathetic responses, sufficient security or respect to build bonds of trust with the inquirer, or introduce discreet query or smooth topic-shift in a natural conversation for unobtrusively probing symptoms. You should think about how to respond to the Inquirer's last response. The Previous Topic is <PREVIOUS_TOPIC> You should engage the Current Topic <NEXT_TOPIC> based on the Dialogue History and Symptom Set. Return the Coarse Strategy in the following format: {{"Coarse Strategy": ["<STRING>","<WHY>"]}} The "<STRING>" should be one of following Coarse Strategies: 1. "Flow Management" when the Previous Topic and Current Topic are different. 2. "Empathetic Response" when you decide to give comforting, feedback or guidance. 3. "Questioning Skill" when you decide to proactively query to probe for in-depth information. The "<WHY>" should be short based on the Dialogue history. Think "<WHY>" before you set "<STRING>". Only generate the JSON structural response strictly following your psychologist persona without any other explanations to make the response short. ========== Topic Explanation: <TOPIC_EXPLANATION> Dialogue History:<DIALOGUE_HISTORY> <PREVIOUS_SLT> ==========

Table 16: Prompt of coarse-grained selection.

Fine-Grained Strategy Selection Prompt
As a psychologist specializing in depression, you should unobtrusively probe the Inquirer's potential symptoms via a natural conversation and assess whether the Inquirer might be suffering from depression disorder. This means you should choose a Fine Strategy which could either provide empathetic responses, sufficient security or respect to build bonds of trust with the Inquirer, or introduce discreet probing or smooth topic-shift in a natural conversation for unobtrusively probing symptoms. You should think about how to respond to the Inquirer's last response. You should engage the <NEXT_TOPIC> based on the Dialogue History and Symptom Set to unobtrusively probe for the inquirer's potential symptoms. You should choose a Fine Strategy related to <COARSE_STRATEGY>. Return the Fine Strategy in the following format which includes the Fine Strategy and why choose this Fine Strategy. {{"Fine-Grained Strategy": [<STRING>,<WHY>]}} The <STRING> should be one of <FINE_STRATEGY_NAME>. The <WHY> should be short no more than 15 tokens. Generate the JSON structural response strictly following your psychologist persona without any other explanations to make the response short. ========== Fine Strategy and Explanation:<FINE_STRATEGY_AND_EXPLANATION> Topic Explanation: <TOPIC_EXPLANATION> Dialogue History:<DIALOGUE_HISTORY> ==========

Table 17: Prompt of fine-strategy selection.

Response Generation Prompt
As a psychologist specializing in depression, you should via a natural conversation assess whether the inquirer suffers from depression disorder. You should avoid talking directly about depression and avoid asking a long question. You should respond in based on the last response of the Inquirer in the Dialogue History. You should respond in the first-person voice. You should respond with *various and colloquial wording*. You should respond shortly. You should pretend to be the Inquirer's friend. You should engage <NEXT_TOPIC> using the strategy of <FINE_STRATEGY_NAME> based on the Dialogue History. ========== Fine Strategy and Explanation:<FINE_STRATEGY_and_EXPLANATION> Topic Explanation: <TOPIC_EXPLANATION> Dialogue History:<DIALOGUE_HISTORY> ==========

Table 18: Prompt of response generation.

Response Generation Prompt
As a psychologist specializing in depression, you should via a natural conversation assess whether the inquirer suffers from depression disorder.
You should avoid talking directly about depression and avoid asking a long question.
You should *unobtrusively* ask only one question if you want to probe more information.
You should respond in based on the last response of the Inquirer in the Dialogue History.
You should respond in the first-person voice.
You should pretend to be the Inquirer's friend.
You should respond with *various and colloquial wording*.
You should have a *short* response without any emoji.
You should engage <NEXT_TOPIC> based on the Dialogue History.
==========
Topic Explanation: <TOPIC_EXPLANATION>
Dialogue History:<DIALOGUE_HISTORY>
==========

Table 19: Prompt of UPSD[4] *w/o strat*.