# Toward Flexible Psychiatric History-Taking and Visualization: Exploring Clinician Perspectives with Large Language Models

Yugyeong Jung
School of Computing
KAIST
Daejeon, Republic of Korea
yugyeong.jung@kaist.ac.kr

Thu Hoang Anh Vo
School of Computing/ICLab
Korea Advanced Institute of Science & Technology
Daejeon, Republic of Korea
thu.vohoanganh96@gmail.com

Hyun Seung Moon
Industrial Design/ AI Exprience Lab
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
mzes0401@kaist.ac.kr

Jae Young Choi
Department of Industrial Design
KAIST
Daejeon, Republic of Korea
jaeyoungchoi@kaist.ac.kr

Hyangkyeong Oh
Yonsei University College of Medicine
Institute of Behavioral Sciences in Medicine
Seoul, Republic of Korea
ohk92090@yuhs.ac

Ujin Lee
Yonsei University College of Medicine
Institute of Behavioral Sciences in Medicine
seoul, Republic of Korea
becoming496712@naver.com

Eunjoo Kim
Yonsei University College of Medicine
Seoul, Republic of Korea
ejkim96@yuhs.ac

Tak Yeon Lee[*]
Industrial Design department
KAIST
Daejeon, Republic of Korea
takyeonlee@kaist.ac.kr

Uichin Lee[†]
School of Computing
KAIST
Daejeon, Republic of Korea
uclee@kaist.ac.kr

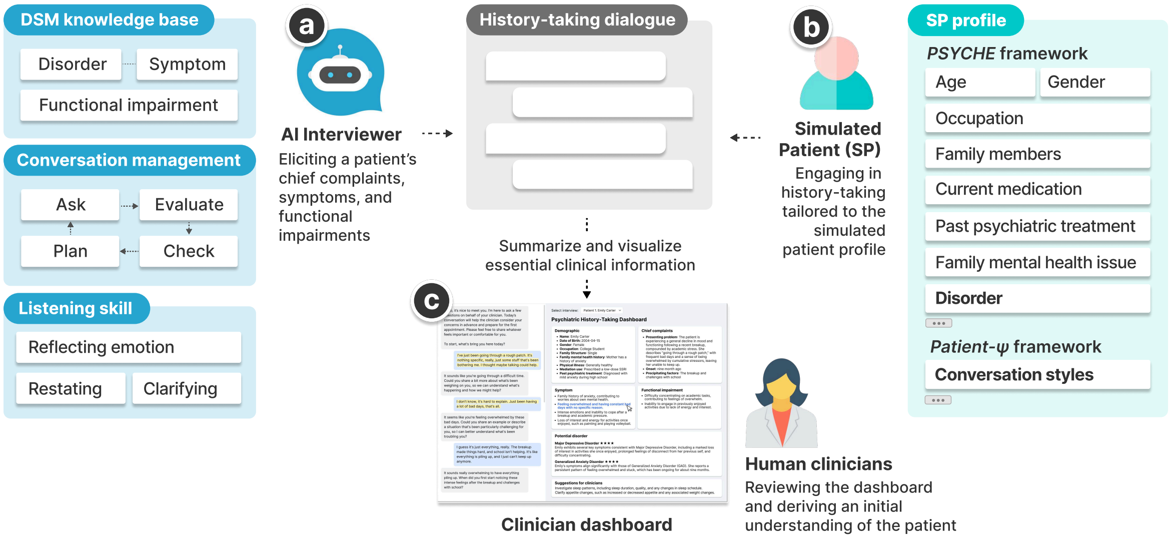


**Figure 1: Overall architecture of our exploratory system design. (a) AI interviewer integrates a DSM knowledge base, conversation management, and listening skills for initial history-taking (Section 4.1). (b) Simulated patients engage in history-taking dialogue tailored to their predefined profiles (Section 4.2). (c) The clinician dashboard summarizes and visualizes essential clinical information (Section 4.3), enabling human clinicians to review the dashboard and derive an initial understanding.**

## Abstract

The initial psychiatric interview centers on patients' chief complaints, symptoms, and functional impairments, forming the basis of diagnostic impressions. In real clinical practice, however, interviews are constrained by limited time and the unpredictability of patient responses, making it difficult to secure essential information efficiently. While prior conversational agents have focused on conversationalizing validated instruments or advancing interview systems in general medical domains, little research has addressed the distinctive challenges of initial psychiatric history-taking from clinicians' perspective. We present a flexible psychiatric interviewer that dynamically adapts question flow and prioritizes clinically essential information within time constraints, with a clinical dashboard for efficient review. We evaluated the system through 1,440 simulated patient dialogues and follow-up interviews with 19 clinicians. Results show that it captures essential information within a limited time while preserving conversational flexibility and empathy, highlighting design implications for coachable and responsible AI interviewers that align with clinical practice.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Health informatics.*

## Keywords

psychiatric interview, conversational agents, simulated patients, human-computer interaction

**ACM Reference Format:**
Yugyeong Jung, Thu Hoang Anh Vo, Hyun Seung Moon, Jae Young Choi, Hyangkyeong Oh, Ujin Lee, Eunjoo Kim, Tak Yeon Lee, and Uichin Lee. 2026. Toward Flexible Psychiatric History-Taking and Visualization: Exploring Clinician Perspectives with Large Language Models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3790970

## 1 Introduction

The initial psychiatric interview is the first encounter between a patient and a clinician, shaping subsequent diagnosis and treatment planning [9, 42, 102]. Unlike other medical fields, where decisions are guided by objective tests, psychiatry relies primarily on patients' verbal accounts as diagnostic material [8, 34, 76]. How patients describe their symptoms and how these difficulties manifest across academic, daily, and social contexts are key determinants of clinical decisions [19, 45, 96]. Clinicians begin with the patient's chief complaint and core symptoms and extend to functional impairments across multiple life domains. This process goes beyond standardized questioning and unfolds as a dynamic interaction in which the clinician flexibly adapts and elaborates questions in response to the patient's narrative [95].

In practice, these interviews are carried out under a number of constraints. Foremost among them is limited time, which makes the efficient use of this resource essential [33, 124]. Clinicians must capture a broad range of symptoms and contextual background within these limits while avoiding overly long sessions that reduce patient concentration or disrupt workflow. Variability and unpredictability in patient speech add further challenges, often obscuring core issues or complicating interpretation [76, 96]. To address this, clinicians are required to guide the conversation naturally and flexibly to secure essential diagnostic information [75], while also fostering a therapeutic relationship [18, 43].

To support clinical information gathering, HCI research has explored LLM-based interviewers in general medical contexts, demonstrating their potential for patient engagement and rapport building [30, 62, 66]. In psychiatry, conversational agents grounded in structured diagnostic tools such as the PHQ-9 [58] and the MINI [99] have been developed to enhance the reliability of information collection by enforcing standardized protocols [6, 14, 68]. Yet these systems have largely focused on conversationalizing validated instruments or advancing in general medical domains, while paying less attention to the distinctive challenges of the *initial psychiatric interview*. In particular, little research has examined how clinicians' perspectives and the realities of time-constrained history-taking should inform system design.

Building on this background, our study addresses this gap by foregrounding clinicians' perspectives in the design and evaluation of an AI interviewer for initial psychiatric interviews. Through in-depth interviews with six clinicians, we identified key challenges and translated them into system requirements. The system employs an adaptive interview design that dynamically adjusts questions to patient input, prioritizing essential information within limited time while maintaining natural dialogue. Combined with listening techniques, this design moves beyond simple information gathering to foster patient-centered interaction, supporting both the completeness of information and rapport building. A clinical dashboard summarizes and visualizes key interview data, helping clinicians quickly grasp patient status.

We evaluated the system through 1,440 simulated patient dialogues and interviews with 19 clinicians. Due to ethical and safety concerns inherent in experimenting with real patients in mental health contexts, we employed clinically validated simulated patients [60, 118] for our evaluation. Results show that the AI interviewer effectively captured essential information within limited time and was positively received by clinicians, particularly for ensuring high completeness while preserving conversational flexibility and empathy. However, clinicians also emphasized the need to balance broad coverage with in-depth probing and to carefully handle sensitive topics. Building on these insights, we discuss the potential role of AI interviewers and design implications in initial psychiatric interviews.

In summary, our contributions are as follows:

- Through interviews with six clinicians, we identify practical needs in initial psychiatric interviews and present an AI

[*] Corresponding author
[†] Corresponding author

*CHI '26, Barcelona, Spain*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790970

interviewer that prioritizes the collection of essential clinical information under time constraints.

- Through a mixed-method evaluation of 1,440 simulated dialogues with 19 clinicians, we demonstrate the feasibility of adaptive psychiatric interviewing and provide empirical insights into managing AI-patient conversation.
- We propose several design implications for the development of coachable and responsible psychiatric interviewer systems for clinical practice.

## 2 Background and Related Work

In this section, we review existing literature relevant to initial psychiatric interviews and the systems developed to support clinical interviews in the medical domain.

### 2.1 Initial psychiatric interview

The initial psychiatric interview constitutes the starting point of psychiatric history-taking and assessment. Because psychiatry relies heavily on patients' speech and narratives [39, 50], careful elicitation and translation of patients' accounts into clinically meaningful information are essential [95, 106]. To systematize this process, the interview is usually divided into two parts: history-taking and the mental status examination (MSE) [116]. History-taking explores psychiatric and personal background, the onset and course of symptoms, and their social context [51]. The MSE offers a cross-sectional view of the patient's current condition, including mood, thought, memory, and judgment [53, 110]. Clinicians integrate both to form provisional diagnoses and treatment plans [84, 103].

Along with this structure, standardized self-report measures such as the PHQ-9 and GAD-7 are also widely used in practice [100, 104]. However, key information about symptom duration, severity, and functional impairment must still be confirmed through the interview [32]. Moreover, patients often struggle to articulate their conditions due to limited insight, avoidance, or stigma, making clinicians' questioning strategies critical [13, 96]. Reflecting these demands, prior studies show that structured interviews improve diagnostic reliability and efficiency in information gathering [27, 88], while empathic techniques and affective reflection facilitate patients' emotional expression and spontaneous narratives [128]. Thus, the initial psychiatric interview is a complex practice that demands both procedural rigor and interactional flexibility. Despite extensive descriptive work, few studies have translated these insights into system design, underscoring the need to incorporate clinicians' practical constraints in the *initial* psychiatric interview.

### 2.2 Supporting clinical interview in medical domain

*2.2.1 General medical domain.* Recent work shows that conversational agents are feasible in diverse clinical settings [48], leading to systems that support history-taking and diagnostic workflows. One line of research targets efficient history-taking [40, 41, 65]. Li et al. [65] propose a doctor agent that asks symptom-related questions based on patient history. *AnCha* [40], a rule-based chatbot built on IBM Watson Assistant [49], collects medical history, chief complaints, social background, and prevention-related information. *Quro* [41] leverages a medical ontology to extract symptoms and conditions from user input and returns an initial assessment.

Another line of research focuses on improving diagnostic performance through conversational data [77, 112, 117]. *SDBench* [77] models physicians' stepwise reasoning, coordinated by multiple agents, to make accurate diagnoses. Google's *AMIE* [112] assigns patient, clinician, critic, and coordinator roles to agents and trains them via self-play, achieving diagnostic performance comparable to that of human clinicians. Wang et al. [117] propose a multi-agent framework in which a general practitioner leads the assessment and specialists contribute to the final diagnosis.

Recent work in HCI and the medical domain has emphasized not only diagnostic accuracy but also patient experience, trust, and interaction quality. Prior studies show that elements such as question clarity [64, 112], appropriate follow-up questions [62], backchanneling behaviors [30], and empathic expression [66] are key factors that enhance chatbots' perceived thoughtfulness, patients' trust, and willingness to engage, thereby improving interaction quality. Moreover, when safety and clinical expertise are communicated transparently, both trust in and acceptance of the system can be strengthened [57, 93]. Together, this growing body of work reflects a shift beyond efficiency-focused information gathering toward patient-centered conversational experiences [112]. These studies demonstrate the promise of conversational agents in medicine, yet they are designed for contexts in which objective tests can validate conversational findings. Psychiatry, however, relies on patient narratives under time pressure, creating distinct design needs.

*2.2.2 Psychiatric domain.* Building on this contrast, psychiatric disorders often involve overlapping symptoms and lack objective biomarkers, which makes flexible and context-sensitive interviewing essential. To support such interviewing, recent studies have introduced two main approaches: *survey-driven interfaces* that conversationalize validated instruments and *tree-based controllers* that manage turn-by-turn progression.

Survey-driven systems convert validated screening instruments such as the PHQ-9 (Patient Health Questionnaire-9) [58], GAD-7 (Generalized Anxiety Disorder-7) [104], and PCL-5 (PTSD Checklist) [15] into conversational formats, retaining standardization while improving engagement [6, 17, 68, 79]. *Perla* transforms the PHQ-9 into a dialogue and automatically scores free-text responses [6]. *aiCARE* extends chatbot delivery to multiple instruments with empathic feedback [17]. Furthermore, *DSM5AgentFlow* brings DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [7]) items into conversation, ensuring that all questions are asked and triggering re-queries when responses are insufficient [79]. Lucas et al. reconstruct PTSD screening surveys as a virtual agent to foster self-disclosure [68].

Tree-based approaches provide finer control of conversational flow by branching along decision trees and using targeted follow-up questions when information is incomplete [14, 111, 123]. *MAGI* overlays a MINI [99]-based tree with four collaborating agents that handle flow management, diagnostic exploration, response validation, and inference [14]. *TRUST* implements the CAPS-5 protocol [119] as a tree and selects whether to ask, empathize, or summarize, advancing once sufficient information is gathered [111]. *WiseMind* models the DSM-5 tree as a knowledge graph and coordinates a

*Reasonable Mind Agent* for diagnostic reasoning with an *Emotional Mind Agent* for empathy [123].

These approaches have made meaningful progress in improving structured interviewing and clinical completeness. Furthermore, the quality of psychiatric interviews is shaped not only by *what* questions are asked but also by *how* they are delivered [59, 83]. Thus, beyond efficiently capturing relevant symptoms, it is critical to support emotional care and maintain a natural, patient-centered conversational flow [101]. Because patients' internal experiences serve as key clinical evidence, interaction skills such as empathic responses are essential for ensuring conversational quality and for building rapport and trust [59, 111].

While prior systems have advanced protocols for interviewing, they have not fully addressed the challenges of initial psychiatric history-taking, a time-consuming and demanding stage of the interview. Most existing work has focused on conversationalizing validated questionnaires or improving diagnostic performance, without generating deeper insights into how clinicians actually conduct and evaluate psychiatric interviews. Our study complements this body of work by focusing on the practical challenges of psychiatric history-taking and by engaging clinicians to surface their perspectives. In particular, our study differs from prior approaches in the following ways:

- **Adaptive psychiatric interviewer prioritizing key clinical information under time constraints.** Prior studies have focused on broad coverage without considering time constraints. However, limited time makes efficiency a critical factor for practical utility. Our work addresses this by proposing strategies that secure the essential items, such as chief complaints, symptoms, and functional impairments, within the available time.
- **Clinician-centered design and evaluation.** Earlier studies emphasized diagnostic accuracy or patient experience, with limited attention to clinicians' needs during the initial psychiatric interview. Our study differs by deriving design requirements and conducting an evaluation with clinicians. This process foregrounds clinicians' perspectives, ensuring that the system addresses the realities of psychiatric practice.

## 3 Formative Study

In this section, we present a formative study focusing on insights gained from interviews with clinicians experienced in conducting initial psychiatric interviews. The study protocol was approved by the Institutional Review Board (IRB).

### 3.1 Method

To identify design requirements for an AI interviewer supporting psychiatric history-taking, we conducted one-hour semi-structured online interviews with six clinicians in South Korea (2 clinical psychologists, 4 psychiatrists; 3 female; *M* age = 40.5, *SD* = 8.04). We recruited clinicians through snowball sampling, beginning with a small group of psychiatrists and clinical psychologists and expanding our sample through their professional recommendations. Snowball sampling, a non-probabilistic sampling technique, is commonly used to access hard-to-reach professional populations in qualitative HCI research [4, 56, 125]. Interviews addressed (1) the process of initial psychiatric interviews, (2) challenges, and (3) expectations for AI support. All sessions were audio-recorded with participants' consent, and each participant received approximately USD 30 as compensation. Using inductive qualitative analysis [12], three authors collaboratively applied affinity diagramming to cluster insights, identify patterns, and derive higher-level themes until consensus was reached.

### 3.2 Result

Although clinicians varied in their levels of experience and individual interview styles, several common themes emerged. They described a shared interview structure, highlighted recurring challenges in time-constrained settings, and outlined expectations for AI support. These insights informed four design requirements.

**Design Requirement 1. Supporting flexible history-taking according to potential disorders.** Clinicians emphasized that the primary goal of initial psychiatric interviews is to collect information relevant to potential disorders. The process typically starts with the *chief complaints*, which describe patients' most pressing concerns along with duration and causes. Clinicians then probe for *symptoms* (e.g., psychological or somatic difficulties) and *functional impairments* (e.g., reduced ability to manage daily, social, occupational, or academic activities). Based on these accounts, clinicians form preliminary hypotheses about possible disorders and develop subsequent questions accordingly.

Importantly, these hypotheses (e.g., patients' potential disorders inferred by clinicians) evolve as the conversation unfolds. In other words, the interview does not rely on a fixed set of questions but rather shifts its direction based on the patient's symptoms, functional impairments, and corresponding potential disorders. As P2 explained, "*We never know exactly what the diagnosis will be from the start. As the patient talks, new symptoms emerge, and I continuously adjust my questions and working hypotheses.*" Similarly, P1 noted, "*I often start with one hypothesis, but if the patient suddenly mentions something unexpected, I change direction. This interview is a process of reshaping the questions so that we don't miss important information.*" Therefore, it is crucial to enable flexible adaptation to patient responses while maintaining a structured process that guarantees the collection of essential clinical information.

**Design Requirement 2. Collecting essential clinical information within time constraints.** Time is a critically scarce resource, particularly during initial interviews where it often becomes the greatest bottleneck. Although clinicians are trained for comprehensive history-taking, schedules rarely allow such depth. P6 highlighted, "*If we had an hour or two, we could gather all the information we need, but the real issue is that the time we're given is always limited.*" In about 30 minutes, clinicians often struggle to cover patients' key issues thoroughly. These constraints make it difficult to address the full range of necessary information or to explore patient concerns in depth. P3 explained, "*Time is a real challenge. Especially in initial interviews where there is so much to cover, we often cannot hear everything in full and end up missing important points.*" P1 added, "*Some patients have so much they want to say. When the session ends due to time, they feel frustrated that they could not share everything. We also regret not being able to fully understand them.*"

Therefore, clinicians envisioned AI playing a workflow-aligned role by assisting with rapid history-taking. In current practice, patients often wait a long time before seeing a psychiatrist, during which they may complete a mental health questionnaire, depending on the hospital system. Typically, psychiatric residents first conduct a history-taking interview to collect key information such as symptoms and functional impairments, which then informs the attending psychiatrist's diagnostic interview.[1] As a result, the initial psychiatric interview is divided between two clinicians: the resident, who focuses on history-taking, and the psychiatrist, who conducts the main diagnostic process. Clinicians suggested that the preliminary history-taking role of residents could be complemented by an AI system. P2 mentioned, "*If patients could briefly go over their history with the AI before meeting me, it would help save valuable time.*"

**Design Requirement 3. Minimizing psychological burden using listening skills.** All clinicians emphasized that in psychiatric interviews, how information is elicited is as important as what is collected. Because the disclosure of personal experiences is highly sensitive, interactions must help patients feel heard and respected. Listening skills, in particular, were seen as crucial for reducing patients' psychological burden, fostering trust, and eliciting richer narratives. As P4 noted, "*One of the most basic skills is listening, which means paying close attention to what the patient is saying and continuing the conversation with respect.*" P1 added, "*When a patient shares their emotions, it is important to acknowledge and reflect them. For example, saying something like 'That must have been very difficult for you' or 'You must have felt sad at that time' can help convey empathy and make the patient feel validated.*" Accordingly, clinicians expected AI systems to incorporate such techniques to encourage patients to open up more quickly: "*People tend to open up much more quickly when their emotions are empathized with and acknowledged. If various counseling skills could be built into AI, I believe it would make it much easier to build trust with patients.*" (P4)

**Design Requirement 4. Summarizing major clinician information to support diagnostic decision-making.** Clinicians emphasized that while AI can help ask questions and elicit narratives, the ultimate diagnosis must remain with human clinicians. Therefore, they expressed an expectation that AI systems extract and organize information from conversations in a way that supports diagnostic decision-making. As P6 explained, "*I think there are clear limits to expecting AI to make a definitive diagnosis. The ideal scenario is for AI to ask the questions we want, record the patient's responses, and then generate a summary. That way, the clinician still makes the final diagnosis, but the AI can greatly support the initial assessment process.*" Clinicians identified three categories of information as critical for diagnosis: the "*chief complaints that bring the patient to the clinic*" (all clinicians),"*the patient's symptoms and functional impairments* (P1, P4, P6), and "*the potential candidate disorders inferred from this information, which serve as hypotheses guiding further questioning and differential diagnosis*" (P3, P4, P5). Clinicians stressed that summaries should go beyond surface complaints to capture clinically relevant details and synthesize them into plausible diagnostic directions, without overstepping into making a definitive diagnosis.

## 4 Exploratory System Design

Building on the design requirements identified in the formative study, we developed an exploratory prototype that addresses the practical challenges of initial psychiatric interviews. The system is designed to manage a flexible conversational flow that captures essential clinical information within limited time (*Design Requirements 1 and 2*), incorporate empathic listening skills (*Design Requirement 3*), and provide transparent visualizations to support clinician review (*Design Requirement 4*). As shown in Figure 1, the system comprises three core components: (1) *an AI interviewer* that collects chief complaints, symptoms, and functional impairments related to potential disorders while applying listening techniques to facilitate disclosure (Figure 1a), (2) *simulated patients* that emulate diverse psychiatric conditions and conversational styles to enable safe, controlled evaluation before real-world deployment (Figure 1b), and (3) *a clinical dashboard* that summarizes the dialogue to assist human clinicians in making diagnostic decisions (Figure 1c).

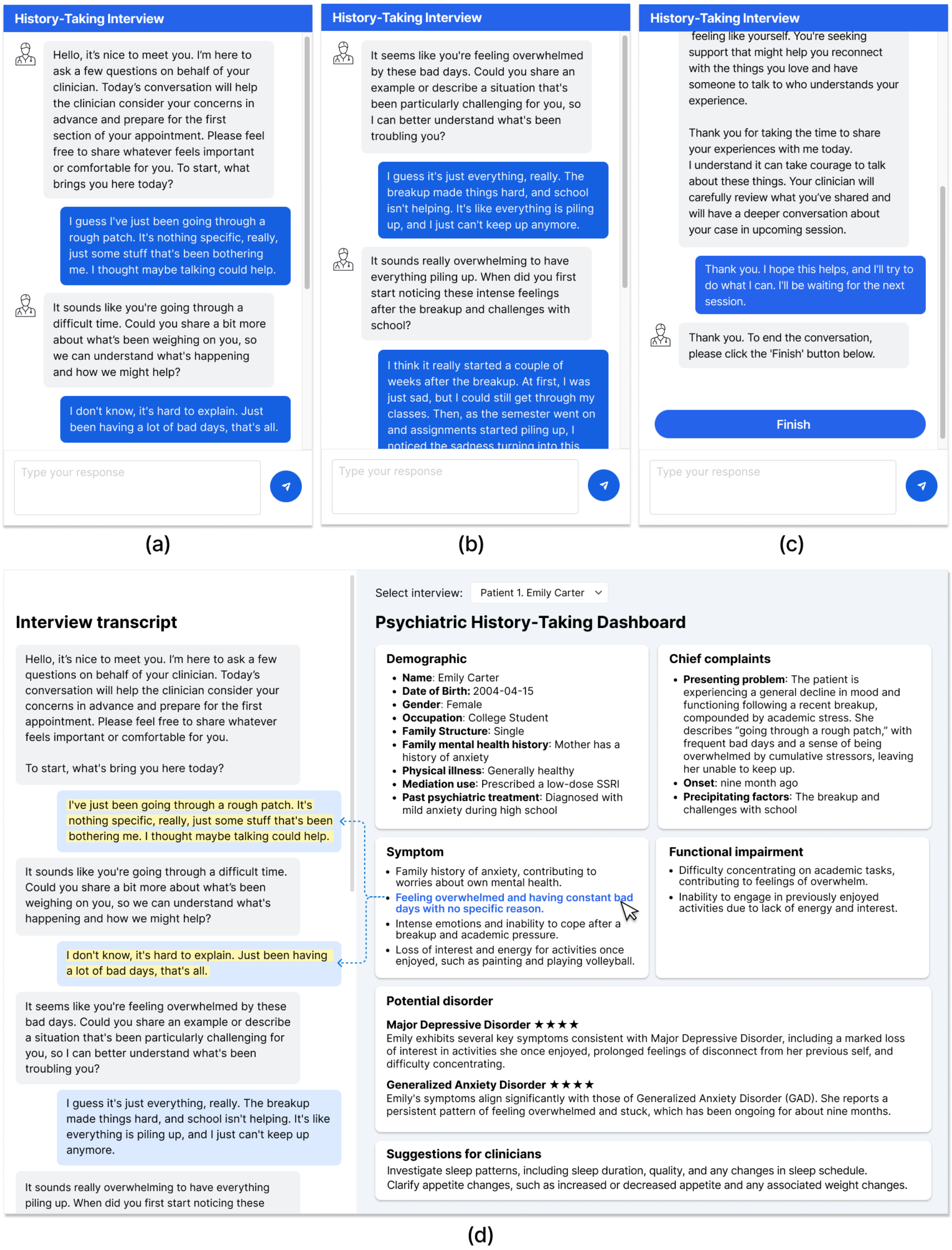
Figure 2 shows the overall interface design of the system, which proceeds in two stages: (1) simulated patients engage in dialogue with the AI interviewer (Figure 2a–c), and (2) clinicians review a structured summary through the clinical dashboard (Figure 2d). In the first stage, patients interact with the AI interviewer via a text-based chat interface, where the interviewer asks questions and patients respond in natural language (Figure 2a, Figure 2b). Once the conversation ends (Figure 2c), clinicians use the dashboard to review the interview results (Figure 2d). The dashboard presents a structured summary highlighting key symptoms and functional impairments relevant to potential disorders. Details of the interviewer design and the implementation of the clinical dashboard are described in the following sections.

### 4.1 AI interviewer design

The AI interviewer is intended to support initial psychiatric interviews, where clinically relevant information is primarily elicited from patients' narratives. This requires flexible approaches that adapt to patients' unfolding stories while remaining mindful of time constraints. Our goal is to develop a system that generates clinically realistic conversations by reflecting how essential information is gathered. The interviewer design is guided by three goals:

- **Knowledge base aligned with DSM diagnostic criteria.** The system incorporates a knowledge base that organizes symptoms and functional impairments in relation to potential disorders.
- **Flexible and time-aware flow.** The dialogue flow dynamically prioritizes essential clinical questions to ensure that chief complaints, symptoms, and functional impairments are addressed within a limited time.
- **Supportive conversational style.** The interviewer applies conversational strategies that reduce patients' psychological burden, acknowledging that trust and comfort are prerequisites for eliciting meaningful information.

[1] A "resident" is a clinician in postgraduate training after completing a professional degree (e.g., medical school or clinical psychology), usually practicing under supervision. A "psychiatrist" is a licensed medical doctor who has completed residency training and is qualified to diagnose mental disorders, prescribe medications, and provide psychotherapy in clinical or hospital settings.



History-Taking Interview

Hello, it's nice to meet you. I'm here to ask a few questions on behalf of your clinician. Today's conversation will help the clinician consider your concerns in advance and prepare for the first section of your appointment. Please feel free to share whatever feels important or comfortable for you. To start, what brings you here today?

I guess I've just been going through a rough patch. It's nothing specific, really, just some stuff that's been bothering me. I thought maybe talking could help.

It sounds like you're going through a difficult time. Could you share a bit more about what's been weighing on you, so we can understand what's happening and how we might help?

I don't know, it's hard to explain. Just been having a lot of bad days, that's all.

Type your response

(a)

History-Taking Interview

It seems like you're feeling overwhelmed by these bad days. Could you share an example or describe a situation that's been particularly challenging for you, so I can better understand what's been troubling you?

I guess it's just everything, really. The breakup made things hard, and school isn't helping. It's like everything is piling up, and I just can't keep up anymore.

It sounds really overwhelming to have everything piling up. When did you first start noticing these intense feelings after the breakup and challenges with school?

I think it really started a couple of weeks after the breakup. At first, I was just sad, but I could still get through my classes. Then, as the semester went on and assignments started piling up, I noticed the sadness turning into this

Type your response

(b)

History-Taking Interview

feeling like yourself. You're seeking support that might help you reconnect with the things you love and have someone to talk to who understands your experience.

Thank you for taking the time to share your experiences with me today. I understand it can take courage to talk about these things. Your clinician will carefully review what you've shared and will have a deeper conversation about your case in upcoming session.

Thank you. I hope this helps, and I'll try to do what I can. I'll be waiting for the next session.

Thank you. To end the conversation, please click the 'Finish' button below.

Finish

Type your response

(c)

Interview transcript

Hello, it's nice to meet you. I'm here to ask a few questions on behalf of your clinician. Today's conversation will help the clinician consider your concerns in advance and prepare for the first appointment. Please feel free to share whatever feels important or comfortable for you.

To start, what's bring you here today?

I've just been going through a rough patch. It's nothing specific, really, just some stuff that's been bothering me. I thought maybe talking could help.

It sounds like you're going through a difficult time. Could you share a bit more about what's been weighing on you, so we can understand what's happening and how we might help?

I don't know, it's hard to explain. Just been having a lot of bad days, that's all.

It seems like you're feeling overwhelmed by these bad days. Could you share an example or describe a situation that's been particularly challenging for you, so I can better understand what's been troubling you?

I guess it's just everything, really. The breakup made things hard, and school isn't helping. It's like everything is piling up, and I just can't keep up anymore.

It sounds really overwhelming to have everything piling up. When did you first start noticing these

Select interview: Patient 1. Emily Carter

Psychiatric History-Taking Dashboard

Demographic
- **Name**: Emily Carter
- **Date of Birth:** 2004-04-15
- **Gender**: Female
- **Occupation**: College Student
- **Family Structure**: Single
- **Family mental health history**: Mother has a history of anxiety
- **Physical illness**: Generally healthy
- **Mediation use**: Prescribed a low-dose SSRI
- **Past psychiatric treatment**: Diagnosed with mild anxiety during high school

Chief complaints
- **Presenting problem**: The patient is experiencing a general decline in mood and functioning following a recent breakup, compounded by academic stress. She describes "going through a rough patch," with frequent bad days and a sense of being overwhelmed by cumulative stressors, leaving her unable to keep up.
- **Onset**: nine month ago
- **Precipitating factors**: The breakup and challenges with school

Symptom
- Family history of anxiety, contributing to worries about own mental health.
- Feeling overwhelmed and having constant bad days with no specific reason.
- Intense emotions and inability to cope after a breakup and academic pressure.
- Loss of interest and energy for activities once enjoyed, such as painting and playing volleyball.

Functional impairment
- Difficulty concentrating on academic tasks, contributing to feelings of overwhelm.
- Inability to engage in previously enjoyed activities due to lack of energy and interest.

Potential disorder

**Major Depressive Disorder ★★★★**
Emily exhibits several key symptoms consistent with Major Depressive Disorder, including a marked loss of interest in activities she once enjoyed, prolonged feelings of disconnect from her previous self, and difficulty concentrating.

**Generalized Anxiety Disorder ★★★★**
Emily's symptoms align significantly with those of Generalized Anxiety Disorder (GAD). She reports a persistent pattern of feeling overwhelmed and stuck, which has been ongoing for about nine months.

Suggestions for clinicians

Investigate sleep patterns, including sleep duration, quality, and any changes in sleep schedule.
Clarify appetite changes, such as increased or decreased appetite and any associated weight changes.

(d)

**Figure 2: Overall interface design of the AI interviewer system. (a, b) Patients interact with the AI interviewer through a chat-based interface to conduct a psychiatric history-taking interview. (c) Upon completing the dialogue, patients can press the 'Finish' button to end the session. (d) The system then generates a summary dashboard that supports clinicians' diagnostic decision-making. The dashboard consists of a dialogue transcript view (left) and a structured summary view (right) highlighting symptoms, functional impairments, and potential disorders.**

*4.1.1 Constructing a DSM knowledge base.* To support history-taking in line with established clinical standards, we constructed a knowledge base grounded in DSM-5 diagnostic criteria [7]. This knowledge base defines the core symptoms associated with each disorder, enabling the system to elicit essential information relevant to patients' potential disorders. In conventional interviews, when patients provide long or ambiguous narratives, the selection of follow-up questions often depends on clinicians' subjective judgment and experience. By contrast, our approach leverages the diagnostic knowledge base to structure this process, ensuring the capture of clinically meaningful information.

Based on DSM criteria, we extracted information related to *disorders, symptoms, and functional impairments*. We first selected five common *disorders* frequently seen in clinical practice [87, 98]: major depressive disorder (MDD) [70], generalized anxiety disorder (GAD) [108], post-traumatic stress disorder (PTSD) [97], attention-deficit/hyperactivity disorder (ADHD) [1], and insomnia disorder [86]. For each disorder, we specified categories of *symptom* information. For example, persistent depressed mood or loss of interest can be mapped to major depressive disorder (MDD). To capture how these symptoms affect patients' daily lives, we defined dimensions of *functional impairment*, drawing on the World Health Organization Disability Assessment Schedule (WHODAS) [114], which is included in the DSM:

- **Cognition**: impairments in thinking, memory, concentration, problem-solving, and communication
- **Mobility**: restrictions in physical movement within and outside the home
- **Self-care**: challenges in basic self-maintenance such as hygiene, dressing, eating, and health management
- **Get along**: challenges in interacting with other people and difficulties that might be encountered with this life domain due to a health condition
- **Life activities**: difficulties in managing household tasks, chores, or family caregiving and impairments in academic or occupational functioning
- **Participation in society**: reduced engagement in social relationships, leisure, and community involvement

The DSM knowledge base is subsequently integrated into the conversation flow, guiding the AI interviewer to systematically explore patients' symptoms and functional impairments.

*4.1.2 Conversation flow design.* Figures 3 and Figure 4 illustrate the overall conversation flow and an example dialogue between the AI interviewer and a patient. In line with *Design Requirement 1*, we designed the flow to draw on both the DSM knowledge base and patients' responses, allowing flexible exploration of symptoms and functional impairments linked to potential disorders. The flow follows an iterative loop with four recurring stages: *Ask, Evaluate, Check, and Plan* (see "main conversation flow" in Figure 3). This loop repeats throughout the dialogue to progressively gather essential clinical information (Figure 4). Guided by *Design Requirement 2*, the flow incorporates a time-management strategy that prioritizes the elicitation of core symptoms and functional impairments, ensuring essential information is captured under time constraints. We next describe how this iterative loop is applied.

**Stage 1. Ask.** The Ask stage begins the dialogue by posing questions to the patient (Figure 31 and Figure 41). It starts with an open-ended question to elicit the chief complaint, such as "*What brings you here today?*" (Figure 4, Q1), encouraging patients to articulate their concerns. After this opening, the system proceeds with questions drawn from the interview plan, which is derived from the DSM knowledge base and patients' responses (details in Stage 4. Plan). In this way, the Ask stage supports to collect cues that guide the subsequent flow of the interview.

**Stage 2. Evaluate.** The Evaluate stage assesses whether the patient's response sufficiently addresses the question posed in the Ask stage (Figure 32 and Figure 42). Building on prior research conducted in the context of psychiatric interviews [63], we adopted four metrics to evaluate response quality:

- **Relevancy:** Whether the response is directly related to the question.
- **Clarity:** Whether the response is unambiguous and comprehensible.
- **Informativeness:** Whether the response provides new information which was not explored before.
- **Depth:** Whether the response goes beyond superficial remarks and includes sufficient detail.

If the system determines that the response meets the criteria, the dialogue advances to the next stage. If the response is insufficient or ambiguous, the system generates up to two follow-up questions to clarify the patient's intent.

**Stage 3. Check.** The Check stage evaluates whether the interview should end after assessing the patient's response, ensuring completion within the limited timeframe (Figure 33 and Figure 43). This reflects *Design Requirement 2* and serves as a time-management strategy to guarantee interview completion within the allotted period. Two termination criteria were implemented: first, a *time-based criterion* ensures that the interview ends once the pre-specified maximum duration (such as 30 minutes) is reached. Second, an *information-based sufficiency criterion* allows the interview to conclude earlier if the collected symptom and functional impairment are sufficient to account for the potential disorder before the time limit. If neither condition is met, the interview continues, and the system moves to the *Plan* stage.

**Stage 4. Plan.** In the Plan stage, the system updates and refines diagnostic hypotheses by integrating patient responses collected thus far. Based on extracted symptoms and functional impairments, the AI interviewer revises potential disorders, identifies missing information guided by DSM criteria and WHODAS domains, and prioritizes the next questions according to clinical importance and remaining time (Figure 3.4 a-e and Figure 44 a-e). The detailed matching process, filtering logic, and prioritization rules are described in the Appendix (Section 8).

*4.1.3 Applying listening skills.* Reflecting *Design Requirement 3*, we incorporated core listening techniques to encourage patients to feel comfortable sharing their stories. Drawing on counseling literature [52], we focused on three skills especially relevant to psychiatric history-taking: *reflecting emotion, restating, and clarifying. Reflecting emotion* acknowledges and verbalizes the patient's feelings. For example, in response to a patient's difficulties, the system might say,"*You felt sad at that time*", mirroring the patient's emotions. *Restating* condenses long explanations into concise core points, such as, "*In sum, you found it difficult because your father was authoritative*," capturing the essence of the patient's account. *Clarifying* structures the dialogue by organizing the patient's narrative. For instance, if a patient says, "*I was stressed at work, and then my sleep*

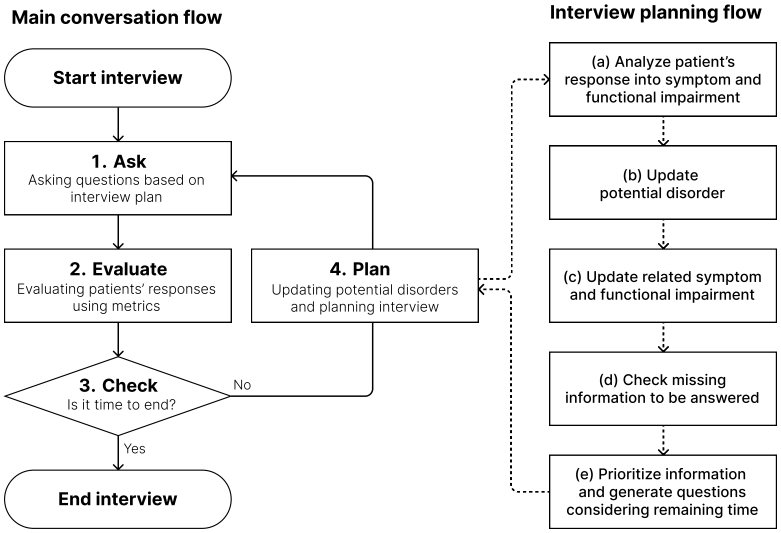

Main conversation flow
Start interview
1. Ask
Asking questions based on interview plan
2. Evaluate
Evaluating patients' responses using metrics
3. Check
Is it time to end?
No
Yes
End interview
4. Plan
Updating potential disorders and planning interview
Interview planning flow
(a) Analyze patient's response into symptom and functional impairment
(b) Update potential disorder
(c) Update related symptom and functional impairment
(d) Check missing information to be answered
(e) Prioritize information and generate questions considering remaining time

**Figure 3: Flow diagram of the AI-supported psychiatric interview process. The main conversation loop follows an Ask–Evaluate–Check–Plan cycle that guides the dialogue with the patient.**

*got worse, and I felt anxious most of the time*," the system might respond, "*So you're saying that work stress affected your sleep, which then increased your anxiety*." Together, these strategies support a more empathetic and structured interview, reducing psychological burden and encouraging patients' disclosure.

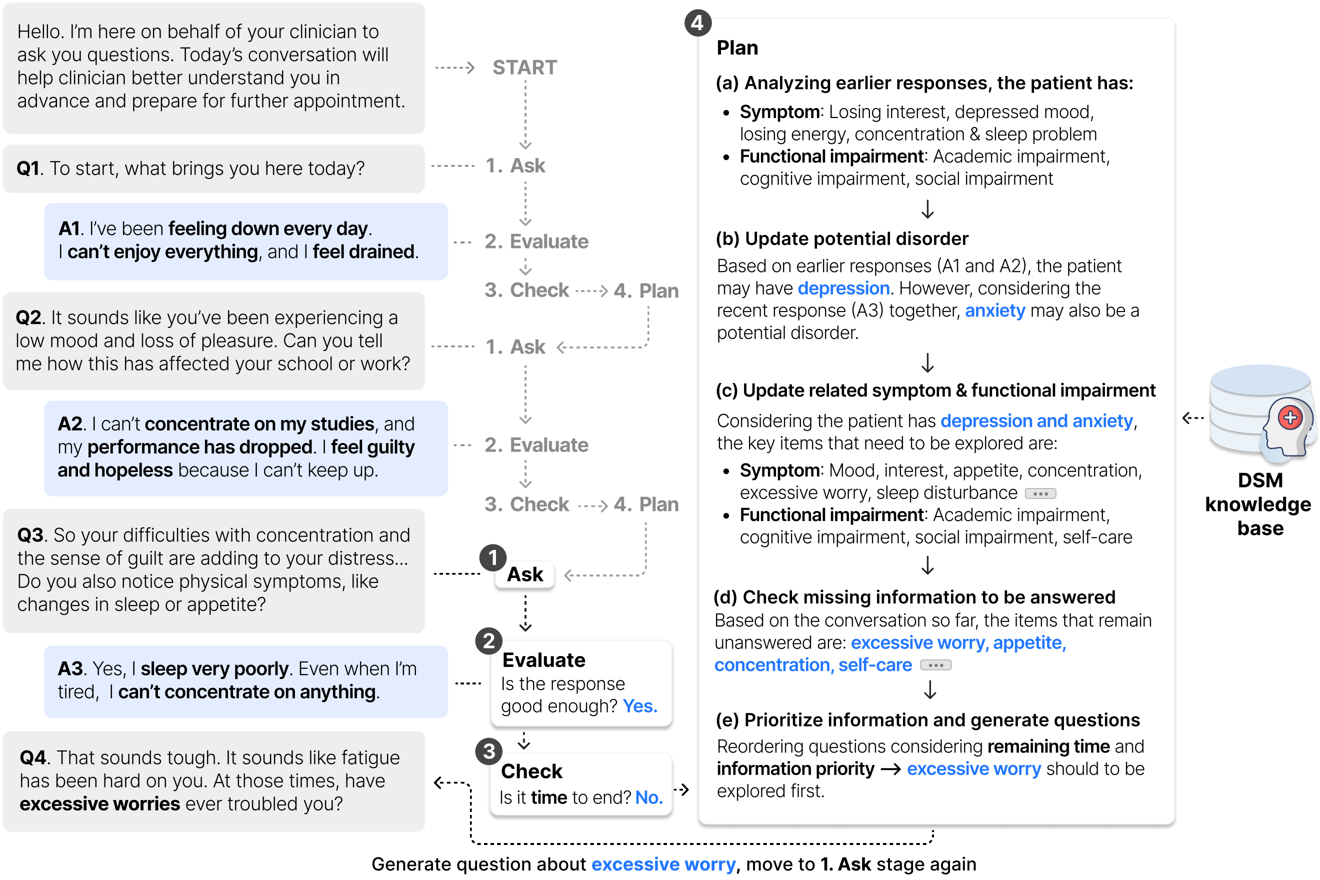
## 4.2 Simulated patients

We use Simulated Patients (SP) to interact with our system, a method widely adopted in prior studies [16, 47, 60, 115, 118]. The rationale for using SP is twofold. First, applying the system to real patients could expose them to inappropriate questions or impose unnecessary psychological burden, whereas SP mitigates these risks while still allowing us to evaluate whether the system effectively captures essential information [60]. Second, involving a large number of real patients would demand substantial time and effort and would limit the diversity of patient cases. By contrast, SP enables the systematic generation of diverse diagnostic profiles and styles, thereby offering a scalable evaluation.

We combined two existing frameworks to generate our SPs. First, the PSYCHE framework [60] is highly relevant as it provides a methodology for constructing SPs with different mental disorders using Large Language Models. We applied five common single disorders (major depressive disorder (MDD) , generalized anxiety disorder (GAD), post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD) , and insomnia disorder) and three comorbid disorders (MDD and GAD, MDD and PTSD, MDD and ADHD). The selection of disorders was informed by obtaining comments from clinicians and literature reviews [11, 36, 69, 72, 90, 127]. Second, to study how the AI interviewer adapts to different patient communication styles, we utilize the prompts from PATIENT-Ψ framework [118] to resemble the complex dynamics of real patient communication (plain, verbose, upset, reserved, tangent, and pleasing). These six conversational styles were identified through their formative study with mental health experts. The specific details on how these SP profiles were generated are included in the Appendix 8.

## 4.3 Dashboard design

Applying *Design Requirement 4*, we designed a dashboard that enables clinicians to review both the interview dialogue and the corresponding clinical information (Figure 2.d). Psychiatric interviews typically generate extensive dialogue records, which are time-consuming to review in full. Our dashboard summarizes patient utterances into essential clinical information, thereby supporting clinicians' diagnostic decision-making. The dashboard consists of two primary views. The *dialogue view* displays the dialogue record between the AI interviewer and the patient (left panel in Figure 2.d), while the *summary view* presents a structured summary distilled from the conversation (right panel in Figure 2.d). The summary content was informed by our formative study, reflecting the information that clinicians considered most critical. Specifically, it includes the patient's demographic information, chief complaints, core symptoms, functional impairments, and potential disorders inferred from the combination of symptoms and impairments. The dashboard highlights information that the AI interviewer may not have been able to collect within the limited time, thereby ensuring that clinically important symptoms and functional impairments are not overlooked (see "Suggestions for clinician" in Figure 2.d). To further enhance interpretability, we incorporated an interaction feature that links each item in the summary view to its source in the dialogue. When users click on a specific item in the summary view (see the mouse cursor in the Figure 2.d), the corresponding patient utterance is highlighted.



Hello. I'm here on behalf of your clinician to ask you questions. Today's conversation will help clinician better understand you in advance and prepare for further appointment.

**Q1**. To start, what brings you here today?

**A1**. I've been **feeling down every day**. I **can't enjoy everything**, and I **feel drained**.

**Q2**. It sounds like you've been experiencing a low mood and loss of pleasure. Can you tell me how this has affected your school or work?

**A2**. I can't **concentrate on my studies**, and my **performance has dropped**. I **feel guilty and hopeless** because I can't keep up.

**Q3**. So your difficulties with concentration and the sense of guilt are adding to your distress... Do you also notice physical symptoms, like changes in sleep or appetite?

**A3**. Yes, I **sleep very poorly**. Even when I'm tired, I **can't concentrate on anything**.

**Q4**. That sounds tough. It sounds like fatigue has been hard on you. At those times, have **excessive worries** ever troubled you?

**Plan**

**(a) Analyzing earlier responses, the patient has:**

- **Symptom**: Losing interest, depressed mood, losing energy, concentration & sleep problem
- **Functional impairment**: Academic impairment, cognitive impairment, social impairment

**(b) Update potential disorder**

Based on earlier responses (A1 and A2), the patient may have depression. However, considering the recent response (A3) together, anxiety may also be a potential disorder.

**(c) Update related symptom & functional impairment**

Considering the patient has depression and anxiety, the key items that need to be explored are:

- **Symptom**: Mood, interest, appetite, concentration, excessive worry, sleep disturbance ⋯
- **Functional impairment**: Academic impairment, cognitive impairment, social impairment, self-care

**(d) Check missing information to be answered**

Based on the conversation so far, the items that remain unanswered are: excessive worry, appetite, concentration, self-care ⋯

**(e) Prioritize information and generate questions**

Reordering questions considering **remaining time** and **information priority** → excessive worry should to be explored first.

Generate question about excessive worry, move to **1. Ask** stage again

**Figure 4: Example of a dialogue between a patient and the AI interviewer. The excerpt demonstrates how the system progresses through the Ask–Evaluate–Check–Plan cycle**

## 5 Evaluation

In this section, we describe how we evaluated our system through simulated dialogues and expert interviews.

### 5.1 Quantitative evaluation

*5.1.1 Method.* The goal of the quantitative analysis was to assess how comprehensively the AI interviewer gathered chief complaints, symptoms, and functional impairments within time constraints. To this end, we simulated large-scale dialogues with simulated patients and measured whether the system collected sufficient clinical information.

**Generating simulated patients.** To evaluate the AI interviewer across diverse scenarios, we generated simulated patients that varied by disorder and conversational style. Specifically, we combined eight disorders (MDD, GAD, PTSD, ADHD, Insomnia, MDD+GAD, MDD+PTSD, MDD+ADHD) with six conversation styles (plain, verbose, upset, reserved, tangent, pleasing), yielding 48 combinations. For each combination, we created 30 patient profiles balanced by gender, within an age range of 19–33 years to capture young adulthood, a stage marked by educational, occupational, and social transitions and elevated risk for mental health conditions [122]. This process produced 1,440 simulated patients, each of whom engaged in an interview, resulting in 1,440 dialogues for evaluation.

**Dialogue simulation.** We set the interview duration to up to 30 minutes, reflecting the timeframe frequently mentioned in the formative study. To avoid the inefficiencies of real-time interaction while still modeling realistic conversational pacing, we employed a virtual clock that simulated reading, thinking, and typing times. Further details on how each component of the simulation was quantitatively estimated are provided in the Appendix (Section 8).

**Metric.** During each interview, the AI system collected patient information, including chief complaints, symptoms, and functional impairments. Based on formative study feedback, we defined *essential items* for each patient: chief complaints (onset and causes), disorder-related symptoms (symptoms relevant to potential disorders), and six WHODAS functional impairment dimensions [114] (the list of items is provided in Table 2). *Completeness* was measured as the proportion of these essential items collected during the dialogue, evaluated at 5-minute intervals up to 30 minutes:

$$\text{Completeness}(t) = \frac{\text{Number of essential items collected up to time t}}{\text{Total number of essential items for the patient}}$$

**Analysis.** We examined the effects of disorder type and conversational style on completeness. A two-way ANOVA was considered, but Levene's test indicated heterogeneity of variances. Therefore, we adopted a generalized linear model (GLM) as the primary analysis and used the Games-Howell test for post hoc comparisons.

*5.1.2 Result.* Figure 5 presents changes in completeness over time for the Plain style. Overall, completeness steadily increased across most disorders as time progressed. By the 30-minute mark, all disorders except MDD+PTSD, ADHD, and MDD+ADHD had reached ≥ 0.90 completeness. For single-disorder cases such as insomnia, MDD, and GAD, the system rapidly achieved high coverage, surpassing 0.90 completeness by 10 minutes and approaching near-complete levels (> 0.95) by 15 minutes. PTSD followed a slightly slower trajectory, reaching 0.95 at 20 minutes and stabilizing at 0.98 by 30 minutes. In contrast, comorbid cases posed greater challenges due to the larger volume of information required. While MDD+GAD and MDD+PTSD improved steadily, reaching 0.96 and 0.86 completeness at 30 minutes, respectively, ADHD and MDD+ADHD showed the lowest rates, plateauing at 0.83 and 0.78. Similar patterns were observed for other conversation styles, with details provided in the Appendix (Figure 6).

GLM revealed significant main effects of disorder type ($\chi^2(7)$ = 2385.40, p < .001), conversational style ($\chi^2(5)$ = 125.09, p < .001), and their interaction ($\chi^2(35)$ = 322.24, p < .001) on completeness. Games–Howell post hoc tests showed that completeness was significantly lower for ADHD and MDD+ADHD than for other disorders. GAD, insomnia, MDD, and MDD+GAD reached near-ceiling levels, whereas PTSD and MDD+PTSD fell in between. Differences across styles were smaller than those across disorders, but the upset style consistently produced the lowest completeness, whereas the reserved style yielded the highest.

Finally, there was a significant Style × Disorder interaction ($\chi^2(35)$ = 322.24, p < .001), indicating that the impact of conversational style was particularly pronounced in the MDD+ADHD and ADHD groups but negligible in other diagnostic categories. This occurred because ADHD-simulated patients often report daily-life difficulties, such as concentration problems, restlessness, or impulsive mistakes, which overlap with depressive or anxiety-related symptoms, particularly in upset styles. These overlaps make it challenging for the AI interviewer to accurately distinguish the underlying disorder. As a result, the system may misinterpret these symptoms and generate less appropriate or less targeted questions. The system demonstrates a tendency to prioritize questions concerning depression or anxiety over those targeting ADHD. For example, some ADHD simulated patient utterances include: *"**I feel very stressed and exhausted**. I make a lot of mistakes at work… I just want things to get better... Yes, organizing daily tasks is really hard. I often miss important deadlines, which makes me feel guilty toward my coworkers, and that pressure feels heavier. I put off household chores, and **I talk less with my family**… **I feel lonely and anxious.**"* The phrases (highlighted in bold) shifted the interview toward depressive or anxiety-related questioning.

## 5.2 Expert evaluation

*5.2.1 Method.* To examine the feasibility of applying the AI interviewer in clinical settings, we conducted an expert evaluation focused on its practical applicability. The study protocol was reviewed and approved by the Institutional Review Board (IRB), and written informed consent was obtained from all participants.

**Participants.** We contacted eleven hospitals in South Korea and recruited 19 clinicians (12 psychiatrists and 7 clinical psychologists; 10 female; age: *M* = 38.8, *SD* = 7.3) with extensive experience in psychiatric history-taking across university hospitals and private medical clinics. Their average clinical experience was 11.6 years. We employed snowball sampling to recruit participants. We first identified a small group of psychiatrists and clinical psychologists with whom initial contact was feasible and designated them as seed participants. These participants were then asked to recommend other clinicians with relevant professional experience, allowing us to expand the sample.

**Study procedure.** Each 90-minute evaluation session followed a structured sequence. After being introduced to the study purpose and system features, participants reviewed five dialogues between simulated patients and the AI interviewer. Simulated patients were generated using the PSYCHE and PATIENT-Ψ frameworks, spanning eight disorders (MDD, GAD, PTSD, ADHD, Insomnia, MDD+GAD, MDD+ADHD, MDD+PTSD) and six conversation styles (plain, verbose, reserved, upset, tangent, pleasing). Five dialogues were selected to cover three single-disorder and two comorbid-disorder cases, each with a distinct conversational style.

For each dialogue, participants performed three tasks. First, they examined the dialogue using the clinical dashboard (left panel in Figure 2d), assessing whether the AI interviewer appropriately elicited symptoms and functional impairments relevant to the patient's potential disorders. Next, they reviewed the dashboard (right panel in Figure 2d), which presented key information extracted from the dialogue. Finally, they completed a survey using the PIQSCA (Psychiatric Interview Quality Scale for Conversational Agents) questionnaire [60], which assesses conversational agents across three factors: (1) process, evaluating session structure; (2) techniques, assessing facilitative interventions and avoidance of obstructive behaviors; and (3) information diagnosis, evaluating the extent to which the agent collects comprehensive information for diagnosis (the full set of items is provided in Table 3). Each factor was rated on a five-point Likert scale.

After the case reviews, participants engaged in a 30-minute semi-structured interview. At the beginning of the interview, we introduced the design goals and design requirements identified through the formative study and asked participants whether they considered these requirements appropriate and important for the design of an AI interviewer for psychiatric history-taking, as well as whether they had any disagreements. Afterwards, the remainder of the interview was organized to support an evaluation of the system across the following areas: perceptions of the AI interviewer's conversational flow, adequacy of data collection, and quality of summarized information; suggestions or expectations for improvement; and potential risks or safety concerns in clinical practice. All interviews were audio-recorded and transcribed. We conducted reflexive thematic analysis [21] in six phases: familiarizing with the data, generating codes, searching for themes, reviewing themes, defining and naming themes, and producing the final report. Two researchers independently coded the transcripts, inductively assigned thematic

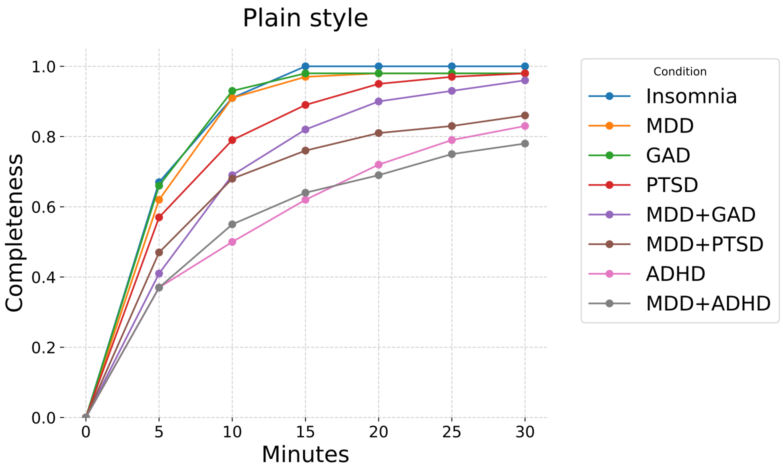


**Figure 5: This graph shows how the completeness of information from interviews changes over time. Each line on the graph represents a different patient condition (e.g., Insomnia, MDD, GAD, PTSD, etc.), and the data points show the average completeness for 30 persona with that condition and a plain speech style.**

**Table 1: GLM results for the effects of patients’ disorders and conversational styles on completeness.**

| Source | Wald Chi-Square | df | Sig. |
|---|---|---|---|
| (Intercept) | 264251.733 | 1 | <.001 |
| Style | 125.094 | 5 | <.001 |
| Disorder | 2385.404 | 7 | <.001 |
| Style × Disorder | 322.241 | 35 | <.001 |

labels, and refined them by merging overlapping codes and prioritizing frequent themes, continuing until consensus was reached on the final set.

*5.2.2 Result.* For the PIQSCA scores (rated on a five-point Likert scale), the *process* of the dialogue was rated 3.96 (*SD* = 0.64), the *techniques* of the conversation 3.84 (*SD* = 0.63), and the *information for diagnosis* 3.69 (*SD* = 0.82). Overall, clinicians acknowledged the importance of the key design requirements identified in the formative study, such as maintaining flexible questioning under time constraints, demonstrating listening skills, and providing structured summaries of clinical information. We provide a detailed account of clinicians’ experiences with the system’s key features.

**Overall Perception: Well structured and flexible, though comparable to a trainee level.** Clinicians perceived the AI interviewer as being on par with “*a first-year resident who forms an initial diagnostic impression and then proceeds to confirm the necessary information*” (P2), or as “*a first-year clinical psychologist who seeks to capture the symptoms comprehensively according to diagnostic criteria*” (P4). Eleven clinicians highlighted as a clear strength the AI interviewer’s ability to maintain a well-defined questioning structure while flexibly covering essential information within the limited time. P7 reflected, “*It seems to cover at least eight of the nine MDD criteria. Honestly, when I was a first-year, I missed a lot of those. But the AI makes sure nothing is overlooked, which I think is quite helpful when making a diagnosis.*” The conversational flow was generally perceived as natural. P19 said, “*The interview flowed quite naturally. I especially appreciated that it was able to cover most of the essential information needed.*”

Alongside these positive evaluations, five clinicians noted that the AI interviewer fell short in “*capturing patients’ emotional nuances, eliciting related narratives*” (P7), or “*probing into concrete episodes and personal histories* (P10). P4, who emphasized the importance of attending to subtle emotional cues, remarked: “*When a patient shows signs of irritation, it’s important to add something like, ‘I’m asking this to better help you.’ But I don’t think the AI was able to pick up on that kind of nuance or respond accordingly.*” Taken together, the AI interviewer was positively evaluated for its ability to structure information with precision, but it lacks in sensitively reading patients’ emotions and eliciting personal narratives, which can lead to the relatively low level of perception in the interviewer’s capability.

**Perceived adequacy of empathy.** Thirteen clinicians noted that the AI interviewer’s empathic responses and listening skills aligned with real clinical practices and played a key role in eliciting patient narratives. As P13 remarked, “*The AI is not just asking questions. It shows empathy toward what the patient says and expands the conversation accordingly, which feels similar to the strategies we use to draw out patient narratives. These responses seem to give patients more room to continue sharing their stories.*” These expressions of empathy were seen as tied to time constraints. P5 mentioned, “*I* *was surprised at how well the AI expressed empathy. It's important to provide enough empathy and support, but that's usually difficult within a limited time. The AI was actually doing it quite well."*

At the same time, eight clinicians noted that lengthy empathic responses are unnecessary in a time-constrained context. P17 said, "*If empathy becomes too long or detailed, there is a risk of failing to secure essential clinical information within the limited time.*" Furthermore, they cautioned that premature or overly elaborate empathy in an initial encounter could be perceived as *hollow empathy*, remarking that: "*I think the level of emotional reflection the AI provides now is sufficient; it doesn't need to go into deep empathy. When a patient says, 'Work has been tough lately,' then this response like, 'That must have been difficult for you,' is enough. But saying something like, 'That must have deeply shaken your sense of self-esteem,' could come across as hollow or even excessive.*" (P1) In sum, clinicians positively evaluated the AI interviewer's ability to provide appropriate empathic responses even within limited time. They thought deep or overly frequent empathy was unnecessary, viewing the current level of empathic engagement as enough.

**Breadth of coverage vs. depth of narrative.** Interestingly, clinicians' expectations of the AI interviewer's information-gathering approach diverged into two directions. ten clinicians valued *breadth of coverage*, its strength in covering as many symptom and functional impairment domains as possible within the limited time. P6 stressed, "*If the purpose of the interview is screening at the initial stage, it's more important to cover a wide range of domains without missing anything rather than probing too deeply.*" This perspective reflects that initial interviews should at minimum secure a baseline overview, which can later be supplemented through the follow-up: "*At the first encounter, getting an overall sense of the key problems is more important than drawing out deep personal history.*" (P14)

In contrast, nine clinicians prioritized *depth of narrative*, delving more deeply into the patient's narrative is important for accurate diagnosis. P8 underscored: "*A good interview is judged by the use of 'exploratory questions'. If a patient says they are stressed, then what matters is the content and quality of that stress, and how they interpret it through their own cognitive framework. I felt the AI lacked questions that probed into those details.*" P5 explained, "*When judging the severity of depression, we consider not only the number of reported symptoms but also qualitative aspects such as the patient's cognitive style, manner of speech, tone of voice, and sense of hope for the future. This kind of contextual depth needs to be further explored.*" However, they also concerned that such detailed exploration is difficult under time constraints. P18 said, "*It would be ideal to explore all narratives deeply, but in a time-constrained setting, essential areas need to come first and other details can be further handled by clinicians.*" In that sense, four clinicians envisioned hybrid possibilities, suggesting that the system could incorporate mechanisms for selective depth. P9 commented, "*Even within a fixed time, if the system could include an algorithm that breaks down questions to allow deeper exploration of the symptoms emphasized in diagnostic criteria, it would be very helpful for clinicians when making a diagnosis.*" Ultimately, whether to prioritize breadth or depth was not a matter of *right or wrong* but related to clinician's interview style and preferences.

**Should everything be asked, or should boundaries be kept?** One of the central concerns clinicians raised was how the system would handle sensitive topics, such as suicidal ideation or traumatic experiences (e.g., sexual assault, car accidents). This concern stemmed from the fact that while such sensitive topics may be essential to make further diagnosis, they also carry the risk of causing psychological distress or unnecessary harm. Clinicians expressed differing views: some emphasized the need for *open exploration*, while others argued for maintaining a *certain degree of caution.*

Regarding open exploration, clinicians noted that if the interview takes place in the safe setting of a hospital, allowing patients to discuss sensitive topics with the AI interviewer is not problematic. They emphasized that such exploration is not only appropriate but essential for accurate assessment. Sensitive domains, including suicidal ideation or self-harm tendencies, are often central to understanding a patient's condition. As P11 noted, "*If the goal is to diagnose depression, then suicidal ideation must naturally be included in the questions. Otherwise, making an accurate diagnosis would be very difficult.*" P7 added, "*Some people may find it easier to respond to an AI. In front of a doctor, they might try to hide such experiences, but with an AI, I think they could be more willing to answer.*"

In contrast, some clinicians emphasized the risks of probing too deeply into sensitive domains, warning that such exploration could trigger strong emotional reactions that an AI might struggle to manage. P8 voiced this concern: "*When it comes to sensitive issues like sexual assault or abuse, patients' reactions can be explosive, and I think it would be very difficult for an AI to handle that. It could even escalate into extreme situations like self-harm. These topics are highly risky for AI to address and must be handled directly by clinicians.*" From this perspective, the AI interviewer's role should be limited to confirming the *presence* of sensitive experiences without delving into details. P10 noted, "*Sensitive topics cannot be entirely omitted, but they should be addressed in only a minimal way, with patients then guided to discuss the details with their clinicians. If patients want to discuss, the AI can talk with, but it is crucial to make clear that there are limits to what the AI can respond to.*" Taken together, handling sensitive topics emerged as a matter of balancing the diagnostic necessity with the imperative of protecting patients, with clinicians' opinions diverging across this spectrum.

# 6 Discussion

In this section, we discuss our findings and the design implications for AI interviewers in initial psychiatric interviews, drawing on the quantitative and qualitative insights.

## 6.1 AI interviewer as a coachable apprentice

In this study, we designed an AI interviewer to support initial psychiatric interviews. Clinicians positively evaluated its ability to systematically gather key clinical information within limited time. However, their expectations for how the AI should conduct interviews diverged. Some preferred broad symptom coverage, while others emphasized the importance of deeply exploring personal narratives. These findings reflect that psychiatric interviewing is a flexible practice shaped by clinicians' interview styles and preferences [39, 50], and that no single gold standard commonly applies.

Therefore, rather than adhering to a single strategy, the AI interviewer can adapt its interviewing approach by reflecting clinicians' feedback. To operationalize this adaptability, clinicians' guidance can be structured along several dimensions: strategy, modality, and scope. In the *strategy* dimension, clinicians can adjust questioning strategies (e.g., breadth and depth, level of exploratory questioning), tones, and specify questions that must be included or avoided. In the *modality* dimension, feedback can be delivered through natural language instructions, short demonstrations [24], or UI manipulations such as sliders for control. In the *scope* dimension, feedback can be distinguished between case-specific feedback that applies only to particular disorders and general feedback that applies consistently across patients. This approach can be connected with the concept of *interactive machine learning*, where users gradually train and refine the model [5]. Similar to prior cases [31, 37], clinicians could correct misinterpretations, demonstrate questioning examples, or adjust interviewing rules. By embedding clinical expertise, the AI interviewer can progressively acquire interviewing strategies.

From this perspective, we envision the AI interviewer not as a *partner* based on complete delegation of work or equal collaboration [55], but as a *coachable apprentice* that continuously develops under expert guidance [120]. In the context of human-AI collaboration, our system would operate within a *human-in-the-loop* framework, progressively adjusting its interviewing strategies based on clinicians' oversight and feedback [22, 61, 91]. Such a structure would retain clinical judgment and accountability with human professionals while allowing the AI to expand its adaptive interviewing capabilities under trustworthy and supervised boundaries. Positioning the AI interviewer as an apprentice reinforces important clinical AI by keeping sensitive psychiatric information under expert control and maintaining privacy-preserving and responsible boundaries around patient data use.

## 6.2 Balancing efficiency and flexibility under time constraints

The AI interviewer can achieve a high level of completeness within a limited time, which indicates improved clinical utility compared to prior work [6, 14] that did not account for time constraints. At the same time, clinicians also noted that the presence of time constraints can shape how the interview unfolds and influence the resulting outcomes. If empathic responses or exploratory questions become prolonged, they risk delaying the acquisition of essential information. Conversely, minimizing such expressions enables rapid collection of information, but may lead to the loss of deeper narratives. Similar challenges have been reported in clinical practice, where time pressure often restricts empathic communication and hinders the elicitation of critical diagnostic signals [3, 29].

Therefore, balancing efficiency and flexibility under time constraints becomes a core challenge for designing interviews. To achieve this, exploratory questions, pacing, and empathic expressions can be modeled as *adjustable parameters* that are flexibly regulated within the available time. For instance, the system can monitor the proportion of uncollected information relative to the remaining time and, when necessary, reduce empathic language while prioritizing essential symptom coverage. On the other hand, when time allows, it can incorporate more emotional responses or invite deeper narratives. Contextual signals extracted from patient utterances, such as tone or emotional nuance, may serve as qualitative cues for flexibly adapting the level of empathy or degree of exploratory questioning. However, these cues are often implicit and were difficult for the current system to reliably detect. This limitation could be addressed by integrating affective cue detection [85, 113] or conversation analysis [25]. Finally, this parameterized design can extend beyond psychiatric interviewing to broader HCI domains, where both efficient information collection and emotionally responsive interaction are required under time limits, such as counseling [92], coaching [105], and customer support [2].

## 6.3 Handling complex or comorbid cases

The AI interviewer broadly achieved high completeness, even in scenarios involving additional disorders, such as bipolar disorder [44], panic disorder [89], obsessive-compulsive disorder [107], and social anxiety disorder [109] (see Figure 6 in the Appendix). However, ADHD and MDD+ADHD cases remained challenging due to overlapping and ambiguously expressed symptoms, which sometimes led the AI interviewer to pursue depression-focused questioning. Such patterns are also reported in clinical practice, where comorbid ADHD complicates diagnostic reasoning [38, 54]. Moreover, comorbid cases require confirmation of a broader set of diagnostic criteria compared with single disorders, which can lower completeness. These cases imply an increased likelihood of misdiagnosis, which poses risks to patient safety by potentially leading to inappropriate treatment decisions or delayed interventions.

Thus, when such conditions are suspected early in the interview, disorder-specific questioning can help clinicians clarify diagnostic meaning. For example, if a patient reports concentration difficulties, targeted follow-up questions that clarify whether the issue stems from mood disturbances or executive function deficits can improve diagnostic reasoning. Similarly, incorporating developmental history can be crucial when inferring ADHD, since symptoms such as inattention or hyperactivity typically originate in childhood [74, 121]. In comorbid situations where it is difficult to cover all symptom elements within a limited time, the system should assign higher priority weights to risk-related and diagnostically discriminative items. In the Plan stage of our Ask–Evaluate–Check–Plan framework, these priorities allow the system to surface unaddressed high-value items, prompting clinicians to decide whether to revisit them or defer remaining elements to follow-up interviews. These design implications enhance interview efficiency and safety, thereby improving the practical utility of the system for complex cases.

## 6.4 Safety guardrails for handling sensitive topics in hospital settings

One of the primary concerns expressed by clinicians was the possibility that patients might disclose highly personal and sensitive topics, such as suicidal ideation or traumatic experiences (e.g., sexual assault, traffic accidents). Our findings highlighted a trade-off between securing diagnostically relevant information and ensuring patient safety. Some clinicians worried that if patients avoided discussing sensitive topics, the system could miss critical diagnostic signals, leading to misinterpretation or irrelevant symptom exploration. In contrast, others emphasized that probing into such topics could trigger emotional outbursts, raising concerns about the system's ability to respond appropriately.

These findings suggest that when vulnerable patients discuss sensitive topics, *guardrails* are needed to both ensure safety and preserve access to clinically important information. Prior studies have emphasized such guardrails for conversational agents [81, 94]. For example, systems have been designed to connect users to hotlines when suicidal ideation is detected [46, 73], or to restrict direct responses to high-risk questions [71]. However, these approaches have usually been developed for everyday life contexts. Our study is situated in hospital, where patients can be directly connected with clinicians immediately, and clinical staff can intervene in crisis situations. This contextual distinction highlights the need for guardrails that are tailored to hospital settings. The goal of initial psychiatric interviews is to gather key information while ensuring patient safety. We propose guardrails applicable to clinical settings.

- **Depth-limiting guardrail**: The system should confirm only the minimal diagnostic signal for inferring a potential disorder (e.g., the presence of suicidal ideation) while in-depth probing is deferred to clinicians.
- **Patient-agency guardrail**: Patients should retain autonomy over sensitive disclosures. The system may provide guidance such as,"*You may share briefly if you prefer. Detailed discussions can be addressed with your clinician*", to ensure that sensitive disclosures occur in a safe context.
- **Emotional breakdown guardrail**: If the system detects affective cues indicating heightened emotional distress or risk of breakdown, the AI interviewer should suspend exploration and either alert or hand off to a clinician.

Finally, it is not only the design of guardrails that requires careful consideration, but also the question of responsibility when those guardrails fail. System designers and medical institutions must assume clear roles. Designers bear responsibility for validating whether guardrails function as intended and for ensuring that an immediate human handoff is triggered in failure. Medical institutions should define the AI interviewer as a supportive tool, with ultimate responsibility for diagnosis resting with clinicians. They must establish and oversee protocols that specify how the tool will be used and how accountability will be distributed when problems arise. In sum, the AI interviewer can contribute meaningful value only when safety and responsibility are embedded at its core.

## 6.5 Simulated patients in exploratory system design

In the sensitive context of initial psychiatric interviews, safety becomes a critical consideration when introducing AI. Therefore, thorough testing must precede clinical deployment, and our study demonstrates that simulated patients can serve as valuable exploratory resources in this process. Using validated simulated patient models [60, 118], we showed that it is possible to conduct early-stage experiments in a safe and ethical environment without involving real patients. This prevents inappropriate questions from reaching patients while enabling diverse scenarios to be explored.

Moreover, simulated patients can serve not only as evaluation tools but also as resources for ensuring early-stage safety and guiding design decisions. By utilizing SPs, researchers and clinicians can observe key dimensions such as question flow, completeness, and empathy in a safe sandbox environment, enabling iterative refinement based on immediate feedback. This facilitates systematic improvement of questioning strategies and system requirements to better align with clinical needs. SP-based experiments may further extend to independent validation by clinicians, who can rapidly test alternative dialogue strategies and compare outcomes by exploring diverse scenarios that would be difficult to observe in actual practice. Similar to the multiverse simulation approach [82], this approach allows the construction of different cases to anticipate how system modifications might influence the interview.

Despite these advantages, SP-based evaluation may not fully capture the complexity of real clinical interactions. Although clinicians found SP conversations realistic, actual patients often exhibit unpredictable behaviors such as diverse symptom expressions, fluctuating engagement, and inconsistent follow-up behaviors [28, 35]. Thus, completeness in SP simulations likely represents an upper bound of real-world performance, and our study should be viewed as an exploratory step rather than a finalized solution. While the aforementioned coachable dimensions demonstrate adaptability in questioning, more advanced clinical adaptation will be required, including managing avoidant communication or emotional complexity. Moreover, safety guardrails must expand into multidimensional approach, including distress detection or clinician intervention triggers. Possible pathways for future real-world validation can include patient data-driven simulations [111] and semi-clinical studies with screened participants under strict safety protocols.

## 6.6 Limitation and future work

Our study has a few limitations. First, the design requirements were derived from a formative study with only six clinicians. Although the expert evaluation confirmed the importance and feasibility of these requirements, there remains a need to examine a broader diversity of perspectives. In addition, participants were recruited from a single geographic region, which may have limited the representation of different healthcare systems, cultural norms of clinical communication, and local psychiatric practices. Future work should engage clinicians from more diverse global contexts to enhance generalizability. Moreover, the current design did not involve patient perspectives, which limits the extent to which the AI interviewer accounts for patients' emotional comfort, privacy concerns, and willingness to engage. Incorporating patient voices will help improve the system to meet user needs and mitigate any psychological burden associated with interacting with an AI interviewer.

In addition, clinicians evaluated the system using a summarized dashboard of simulated conversations rather than through live interaction with the AI interviewer. While this approach allowed a controlled assessment of interview completeness and adaptivity, it constrained the ability to evaluate ecological validity, including real-world usability and utility within busy clinical settings. In practice, additional workflow considerations emerge, such as the timing of the interview, influence on clinician-patient rapport, and integration with hospital information systems. Future studies should deploy the system in clinical workflows and examine its roles.

Finally, our AI interviewer was designed and tested on a limited set of disorders. Although these categories were chosen based on prior literature and clinical consultation as representative profiles commonly observed in adults, real clinical practice encompasses a broader spectrum and more complex comorbidities. Disorders such as schizophrenia spectrum disorders, substance use disorders, and personality disorders were not included in this study. Future work should therefore extend evaluation to a wider range of psychiatric conditions and comorbid presentations.

## 7 Conclusion

We present an AI interviewer for initial psychiatric interviews that integrates clinicians' practical needs, supports time-efficient information gathering, and provides structured visualizations for diagnostic decision-making. Evaluations with simulated patients showed strong coverage of key information within a limited time, though performance varied with patient context. Qualitative feedback from 19 clinicians confirmed the system's feasibility and highlighted considerations for history-taking, empathy, and handling sensitive topics. We envision that this approach can be extended to diverse AI-assisted interviewing solutions by positioning the AI as a coachable apprentice that adapts under clinician supervision, thereby supporting clinician-oriented and responsible practices.

## Acknowledgments

This research was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2025-02305801, RS-2025-02305705).

## References

[1] Elie Abdelnour, Madeline O Jansen, and Jessica A Gold. 2022. ADHD diagnostic trends: increased recognition or overdiagnosis? *Missouri medicine* 119, 5 (2022), 467.

[2] Garima Agrawal, Riccardo De Maria, Kiran Davuluri, Daniele Spera, Charlie Read, Cosimo Spera, Jack Garrett, and Don Miller. 2025. Redefining CX with Agentic AI: Minerva CQ Case Study. *arXiv preprint arXiv:2509.12589* (2025).

[3] Florian Ahrweiler, Melanie Neumann, Hadass Goldblatt, Eckhart G Hahn, and Christian Scheffer. 2014. Determinants of physician empathy during medical education: hypothetical conclusions from an exploratory qualitative survey of practicing physicians. *BMC medical education* 14, 1 (2014), 122.

[4] Sarah Aldaweesh, Deemah Alateeq, Max Van Kleek, and Nigel Shadbolt. 2024. "If Someone Walks In On Us Talking, Pretend to be My Friend, Not My Therapist": Challenges and Opportunities for Digital Mental Health Support in Saudi Arabia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[5] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.

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

[7] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders*. American psychiatric association.

[8] Julie M Aultman. 2016. Psychiatric diagnostic uncertainty: challenges to patient-centered care. *AMA Journal of Ethics* 18, 6 (2016), 579–586.

[9] John W Barnhill. 2017. Chapter 1. The Initial Interview. In *Co-occurring Mental Illness and Substance Use Disorders: A Guide to Diagnosis and Treatment*. 3–12.

[10] John N Bassili and B Stacey Scott. 1996. Response latency as a signal to question problems in survey research. *Public opinion quarterly* 60, 3 (1996), 390–399.

[11] Katja Beesdo, Daniel S Pine, Roselind Lieb, and Hans-Ulrich Wittchen. 2010. Incidence and risk patterns of anxiety and depressive disorders and categorization of generalized anxiety disorder. *Archives of general psychiatry* 67, 1 (2010), 47–57.

[12] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.

[13] Dinesh Bhugra. 2008. Decision-making in psychiatry: what can we learn? 3 pages.

[14] Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, et al. 2025. MAGI: Multi-Agent Guided Interview for Psychiatric Assessment. *arXiv preprint arXiv:2504.18260* (2025).

[15] Christy A Blevins, Frank W Weathers, Margaret T Davis, Tracy K Witte, and Jessica L Domino. 2015. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of traumatic stress* 28, 6 (2015), 489–498.

[16] Anna Bodonhelyi, Christian Stegemann-Philipps, Alessandra Sonanini, Lea Herschbach, Marton Szep, Anne Herrmann-Werner, Teresa Festl-Wietek, Enkelejda Kasneci, and Friederike Holderried. 2025. Modeling Challenging Patient Interactions: LLMs for Medical Communication Training. *arXiv preprint arXiv:2503.22250* (2025).

[17] Rares Boian, Ana-Maria Bucur, Diana Todea, Andreea Luca, Traian Rebedea, and Ioana R Podina. 2025. A conversational agent framework for mental health screening: Design, implementation, and usability. *Behaviour & Information Technology* 44, 10 (2025), 2364–2378.

[18] Hannah Bowker, David Saxon, and Jaime Delgadillo. 2025. First impressions matter: The influence of initial assessments on psychological treatment initiation and subsequent dropout. *Psychotherapy Research* 35, 3 (2025), 368–378.

[19] Paola Bozzatello, Benedetta Giordano, Cristiana Montemagni, Paola Rocca, and Silvio Bellino. 2022. Real-world functioning in psychiatric outpatients: Predictive factors. *Journal of Clinical Medicine* 11, 15 (2022), 4400.

[20] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language* 109 (2019), 104047.

[21] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

[22] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[23] Stuart K Card, Thomas P Moran, and Allen Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM* 23, 7 (1980), 396–410.

[24] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–8.

[25] Benjamin L Cook, Ana M Progovac, Pei Chen, Brian Mullin, Sherry Hou, and Enrique Baca-Garcia. 2016. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Computational and mathematical methods in medicine* 2016, 1 (2016), 8708434.

[26] Ruth Elizabeth Corps and Martin J Pickering. 2024. The role of answer content and length when preparing answers to questions. *Scientific Reports* 14, 1 (2024), 17110.

[27] A Cox, K Hopkinson, and M Rutter. 1981. Psychiatric interviewing techniques II. Naturalistic study: eliciting factual information. *The British Journal of Psychiatry* 138, 4 (1981), 283–291.

[28] Veena Das, Benjamin Daniels, Ada Kwan, Vaibhav Saria, Ranendra Das, Madhukar Pai, and Jishnu Das. 2022. Simulated patients and their reality: an inquiry into theory and method. *Social Science & Medicine* 300 (2022), 114571.

[29] Frans Derksen, Jozien Bensing, and Antoine Lagro-Janssen. 2012. Effectiveness of empathy in general practice: a systematic review. *The British Journal of General Practice* 63, 606 (2012), e76.

[30] Zijian Ding, Jiawen Kang, Tinky Oi Ting Ho, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. 2022. TalkTive: A conversational agent using backchannels to engage older adults in neurocognitive disorders screening. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[31] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[32] Giovanni A Fava, Nicoletta Sonino, David C Aron, Richard Balon, Carmen Berrocal Montiel, Jianxin Cao, John Concato, Ajandek Eory, Ralph I Horwitz, Chiara Rafanelli, et al. 2024. Clinical interviewing: an essential but neglected method of medicine. *Psychotherapy and psychosomatics* 93, 2 (2024), 94–99.

[33] Irosh Fernando, Rahul Gupta, Kate Simpson, Stuart Szwec, Mariko Carey, Agatha Conrad, Todd Heard, and Lisa Lampe. 2025. Improving the time-efficiency of initial mental health assessment (triaging) using an online assessment tool followed by a clinical interview via phone: a randomised controlled trial. *BMC psychiatry* 25, 1 (2025), 635.

[34] Elizabeth H Flanagan, Larry Davidson, and John S Strauss. 2007. Issues for DSM-V: Incorporating patients' subjective experiences. 391–392 pages.

[35] Octavia L Flanagan, Kristina M Cummings, and Kristina Cummings. 2023. Standardized patients in medical education: a review of the literature. *Cureus* 15, 7 (2023).

[36] Janine D Flory and Rachel Yehuda. 2015. Comorbidity between post-traumatic stress disorder and major depressive disorder: alternative explanations and

treatment considerations. *Dialogues in clinical neuroscience* 17, 2 (2015), 141–150.

[37] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*. 29–38.

[38] Xinyu Fu, Weige Wu, Yuru Wu, Xiaofu Liu, Wanting Liang, Rongchuan Wu, and Yun Li. 2025. Adult ADHD and comorbid anxiety and depressive disorders: a review of etiology and treatment. *Frontiers in Psychiatry* 16 (2025), 1597559.

[39] Thomas Fuchs. 2010. Subjectivity and intersubjectivity in psychiatric diagnosis. *Psychopathology* 43, 4 (2010), 268–274.

[40] Floriana Gashi, Selina F Regli, Richard May, Philipp Tschopp, and Kerstin Denecke. 2021. Developing intelligent interviewers to collect the medical history: lessons learned and guidelines. In *Navigating healthcare through challenging times*. IOS Press, 18–25.

[41] Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. 2018. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. In *Connecting the System to Enhance the Practitioner and Consumer Experience in Healthcare*. IOS Press, 51–56.

[42] Merton Gill, Richard Newman, Fredrick C Redlich, and Margaret Col Sommers. 1954. Interviewing. (1954).

[43] Deana Shevit Goldin. 2022. *Fast Facts for Psychopharmacology for Nurse Practitioners*. Springer Publishing Company.

[44] Iria Grande, Michael Berk, Boris Birmaher, and Eduard Vieta. 2016. Bipolar disorder. *The Lancet* 387, 10027 (2016), 1561–1572.

[45] Kassahun Habtamu, Atalay Alem, and Charlotte Hanlon. 2015. Conceptualizing and contextualizing functioning in people with severe mental disorders in rural Ethiopia: a qualitative study. *BMC psychiatry* 15, 1 (2015), 34.

[46] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. 2025. Randomized trial of a generative AI chatbot for mental health treatment. *Nejm Ai* 2, 4 (2025), AIoa2400802.

[47] Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling, et al. 2024. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education* 10, 1 (2024), e59213.

[48] Grace Hong, Margaret Smith, and Steven Lin. 2022. The AI will see you now: feasibility and acceptability of a conversational AI medical interviewing system. *JMIR Formative Research* 6, 6 (2022), e37028.

[49] IBM. [n. d.]. IBM Watson Assistant. https://www.ibm.com/watson/assistant/. Accessed: 2025-09-10.

[50] KS Jacob. 2012. Patient experience and psychiatric discourse. *The Psychiatrist* 36, 11 (2012), 414–417.

[51] Lennart Jansson, Julie Nordgaard, et al. 2016. *The psychiatric interview for differential diagnosis*. Vol. 270. Springer.

[52] Jin-Ryung Kang. 2025. *Counseling Practice: Therapeutic Conversation Skills (2nd ed.)*. Hakjisa. 360 pages. Format: 4×6 Bae-pan; Binding: Softcover.

[53] Indunil Karunarathna, Palith Aluthge, Neha Perera, Sena Gunathilake, Kap De Alvis, PNG Rodrigo, Asok Jayawardana, Cam Fernando, USG Vidanagama, UPN Ekanayake, et al. [n. d.]. The Mental Status Examination: A Multidimensional Tool for Psychiatric and Medical Assessment Across Clinical Contexts. ([n. d.]).

[54] Martin A Katzman, Timothy S Bilkey, Pratap R Chokka, Angelo Fallu, and Larry J Klassen. 2017. Adult ADHD and comorbid disorders: clinical implications of a dimensional approach. *BMC psychiatry* 17, 1 (2017), 302.

[55] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[56] Alexandra Kitson, Petr Slovak, and Alissa N Antle. 2024. Supporting cognitive reappraisal with digital technology: A content analysis and scoping review of challenges, interventions, and future directions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[57] Naja Kathrine Kollerup, Joel Wester, Mikael B Skov, and Niels Van Berkel. 2024. How Can I Signal You To Trust Me: Investigating AI Trust Signalling in Clinical Self-Assessments. In *Proceedings of the 2024 ACM designing interactive systems conference*. 525–540.

[58] 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* 16, 9 (2001), 606–613.

[59] Hyein S Lee, Colton Wright, Julia Ferranto, Jessica Buttimer, Clare E Palmer, Andrew Welchman, Kathleen M Mazor, Kimberly A Fisher, David Smelson, Laurel O'Connor, et al. 2025. Artificial intelligence conversational agents in mental health: Patients see potential, but prefer hum ans in the loop. *Frontiers in Psychiatry* 15 (2025), 1505024.

[60] Jingoo Lee, Kyungho Lim, Young-Chul Jung, and Byung-Hoon Kim. 2025. Psyche: A multi-faceted patient simulation framework for evaluation of psychiatric assessment conversational agents. *arXiv preprint arXiv:2501.01594* (2025).

[61] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[62] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[63] Brenna Li, Saba Tauseef, Khai N Truong, and Alex Mariakakis. 2025. A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[64] Dehe Li, Heman Zhang, Chuntao Lu, and Chunxia Miao. 2025. Use of Artificial Intelligence-Assisted Conversational Agents to Improve Patient Experience Related to Physicians: Cross-Sectional Study in China. *Journal of Medical Internet Research* 27 (2025), e76540.

[65] Xueshen Li, Xinlong Hou, Nirumapa Ravi, Ziyi Huang, and Yu Gan. 2025. A two-stage proactive dialogue generator for efficient clinical information collection using large language model. *Expert Systems with Applications* (2025), 127833.

[66] Dingdong Liu, Yujing Zhang, Bolin Zhao, Shuai Ma, Chuhan Shi, and Xiaojuan Ma. 2025. Scaffolded Turns and Logical Conversations: Designing Humanized LLM-Powered Conversational Agents for Hospital Admission Interviews. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[67] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172* (2023).

[68] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 4 (2017), 51.

[69] Rachel G Klein Salvatore Mannuzza. 2009. Comorbidity in adult attention-deficit hyperactivity disorder. *Key issues in mental health* (2009), 126.

[70] Wolfgang Marx, Brenda WJH Penninx, Marco Solmi, Toshi A Furukawa, Joseph Firth, Andre F Carvalho, and Michael Berk. 2023. Major depressive disorder. *Nature Reviews Disease Primers* 9, 1 (2023), 44.

[71] Ryan K McBain, Jonathan H Cantor, Li Ang Zhang, Olesya Baker, Fang Zhang, Alyssa Burnett, Aaron Kofner, Joshua Breslau, Bradley D Stein, Ateev Mehrotra, et al. 2025. Evaluation of Alignment Between Large Language Models and Expert Clinicians in Suicide Risk Assessment. *Psychiatric Services* (2025), appi–ps.

[72] Diane McIntosh, Stan Kutcher, Carin Binder, Anthony Levitt, Angelo Fallu, and Michael Rosenbluth. 2009. Adult ADHD and comorbid depression: a consensus-derived diagnostic algorithm for ADHD. *Neuropsychiatric disease and treatment* (2009), 137–150.

[73] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176, 5 (2016), 619–625.

[74] Terrie E Moffitt, Renate Houts, Philip Asherson, Daniel W Belsky, David L Corcoran, Maggie Hammerle, HonaLee Harrington, Sean Hogan, Madeline H Meier, Guilherme V Polanczyk, et al. 2015. Is adult ADHD a childhood-onset neurodevelopmental disorder? Evidence from a four-decade longitudinal cohort study. *American Journal of Psychiatry* 172, 10 (2015), 967–977.

[75] David E Ness and Scott F Kiesling. 2007. Language and connectedness in the medical and psychiatric interview. *Patient education and counseling* 68, 2 (2007), 139–144.

[76] Julie Nordgaard, Louis A Sass, and Josef Parnas. 2013. The psychiatric interview: validity, structure, and subjectivity. *European archives of psychiatry and clinical neuroscience* 263, 4 (2013), 353–364.

[77] Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. 2025. Sequential Diagnosis with Language Models. *arXiv preprint arXiv:2506.22405* (2025).

[78] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ Accessed: 2023-09-05.

[79] Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, Xin Sun, and Junxiao Wang. 2025. Trustworthy AI Psychotherapy: Multi-Agent LLM Workflow for Counseling and Explainable Mental Disorder Diagnosis. *arXiv preprint arXiv:2508.11398* (2025).

[80] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do people type on mobile devices? Observations from a study with 37,000 volunteers. In *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*. 1–12.

[81] Jinkyung Park, Vivek Singh, and Pamela Wisniewski. 2024. Toward safe evolution of artificial intelligence (AI) based conversational agents to support adolescent mental and sexual health knowledge discovery. *arXiv preprint arXiv:2404.03023* (2024).

[82] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[83] Pierre Philip, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, and Jean-Arthur Micoulaud-Franchi. 2020. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *NPJ digital medicine* 3, 1 (2020), 2.

[84] Patricia J Polanski and J Scott Hinkle. 2000. The mental status examination: Its use by professional counselors. *Journal of Counseling & Development* 78, 3 (2000), 357–364.

[85] Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with Twitter data. *Scientific reports* 7, 1 (2017), 13006.

[86] Dieter Riemann, Fee Benz, Raphael J Dressle, Colin A Espie, Anna F Johann, Tessa F Blanken, Jeanne Leerssen, Rick Wassing, Alasdair L Henry, Simon D Kyle, et al. 2022. Insomnia disorder: State of the science and challenges for the future. *Journal of sleep research* 31, 4 (2022), e13604.

[87] Soo Jung Rim, Bong-Jin Hahm, Su Jeong Seong, Jee Eun Park, Sung Man Chang, Byung-Soo Kim, Hyonggin An, Hong Jin Jeon, Jin Pyo Hong, and Subin Park. 2023. Prevalence of mental disorders and associated factors in Korean adults: national mental health survey of Korea 2021. *Psychiatry investigation* 20, 3 (2023), 262.

[88] Richard Rogers. 2003. Standardizing DSM-IV diagnoses: The clinical applications of structured interviews. *Journal of Personality Assessment* 81, 3 (2003), 220–225.

[89] Peter P Roy-Byrne, Michelle G Craske, and Murray B Stein. 2006. Panic disorder. *The Lancet* 368, 9540 (2006), 1023–1032.

[90] Nina K Rytwinski, Michael D Scur, Norah C Feeny, and Eric A Youngstrom. 2013. The co-occurrence of major depressive disorder among individuals with posttraumatic stress disorder: A meta-analysis. *Journal of traumatic stress* 26, 3 (2013), 299–309.

[91] Gonesh Chandra Saha, Sanjay Kumar, Avinash Kumar, Hasi Saha, TK Lakshmi, and Niyati Bhat. 2023. Human-AI collaboration: Exploring interfaces for interactive machine learning. *Tuijin Jishu/Journal of Propulsion Technology* 44, 2 (2023), 2023.

[92] Harsh Sakhrani, Saloni Parekh, and Shubham Mahajan. 2021. Coral: An approach for conversational agents in mental health applications. *arXiv preprint arXiv:2111.08545* (2021).

[93] Mana Samiee, Joel Wester, Rune Møberg Jacobsen, Michael Skovdal Rathleff, and Niels van Berkel. 2025. General Practitioners' Perspectives on a Pre-Consultation Chatbot for Shared Decision-Making. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 3117–3131.

[94] Surjodeep Sarkar, Manas Gaur, L Chen, Muskan Garg, Biplav Srivastava, and Bhaktee Dongaonkar. 2023. Towards explainable and safe conversational agents for mental health: A survey. *arXiv preprint arXiv:2304.13191* (2023).

[95] Enikö Èva Savander, Jukka Hintikka, Mariel Wuolio, and Anssi Peräkylä. 2021. The patients' Practises disclosing subjective experiences in the psychiatric intake interview. *Frontiers in Psychiatry* 12 (2021), 605760.

[96] Enikö É Savander, Liisa Voutilainen, Jukka Hintikka, and Anssi Peräkylä. 2024. What to take up from the patient's talk? The clinician's responses to the patient's self-disclosure of their subjective experience in the psychiatric intake interview. *Frontiers in psychiatry* 15 (2024), 1352601.

[97] Alexa Schincariol, Graziella Orrù, Henry Otgaar, Giuseppe Sartori, and Cristina Scarpazza. 2024. Posttraumatic stress disorder (PTSD) prevalence: an umbrella review. *Psychological Medicine* (2024), 1–14.

[98] Jeong-Cheol Seo, Duk-In Jon, Se-Hoon Shim, Hyung-Mo Sung, Young Sup Woo, Jeongwan Hong, Sung-Yong Park, Jeong Seok Seo, and Won-Myong Bahk. 2022. Prevalence and comorbidities of attention deficit hyperactivity disorder among adults and children/adolescents in Korea. *Clinical Psychopharmacology and Neuroscience* 20, 1 (2022), 126.

[99] David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, Geoffrey C Dunbar, et al. 1998. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J clin psychiatry* 59, Suppl 20 (1998), 22–33.

[100] Mark Shevlin, Sarah Butter, Orla McBride, Jamie Murphy, Jilly Gibson-Miller, Todd K Hartman, Liat Levita, Liam Mason, Anton P Martinez, Ryan McKay, et al. 2022. Measurement invariance of the Patient Health Questionnaire (PHQ-9) and Generalized Anxiety Disorder scale (GAD-7) across four European countries during the COVID-19 pandemic. *BMC psychiatry* 22, 1 (2022), 154.

[101] Sverker Sikström, Rebecca Astrid Boehme, Mariam Mirström, Thibaud Agbotsoka, Gergő Győri, Marta Lasota, Mona Tabesh, Lotta Stille, and Danilo Garcia. 2025. Generative AI-assisted clinical interviewing of mental health. *Scientific Reports* 15, 1 (2025), 37737.

[102] Iver F Small, Joyce G Small, RAMON GONZALEZ, and MALCOLM D GYNTHER. 1964. Content reliability of a structured psychiatric interview. *Archives of General Psychiatry* 11, 2 (1964), 192–196.

[103] HN Sno and JJ van Croonenborg. 2006. The guideline'Psychiatric evaluation in adults'. *Nederlands Tijdschrift Voor Geneeskunde* 150, 1 (2006), 24–27.

[104] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* 166, 10 (2006), 1092–1097.

[105] Vidya Srinivas, Xuhai Xu, Xin Liu, Ayush Kumar, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, and Tim Althoff. 2025. Substance over style: Evaluating proactive conversational coaching agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 20848–20880.

[106] Giovanni Stanghellini. 2004. The puzzle of the psychiatric interview. *Journal of Phenomenological Psychology* 35, 2 (2004), 173–195.

[107] Dan J Stein, Daniel LC Costa, Christine Lochner, Euripedes C Miguel, YC Janardhan Reddy, Roseli G Shavitt, Odile A van den Heuvel, and H Blair Simpson. 2019. Obsessive–compulsive disorder. *Nature reviews Disease primers* 5, 1 (2019), 52.

[108] Murray B Stein and Jitender Sareen. 2015. Generalized anxiety disorder. *New England Journal of Medicine* 373, 21 (2015), 2059–2068.

[109] Murray B Stein and Dan J Stein. 2008. Social anxiety disorder. *The lancet* 371, 9618 (2008), 1115–1125.

[110] Paula T Trzepacz and Robert W Baker. 1993. *The psychiatric mental status examination.* Oxford University Press.

[111] Sichang Tu, Abigail Powers, Stephen Doogan, and Jinho D Choi. 2025. TRUST: An LLM-Based Dialogue System for Trauma Understanding and Structured Assessments. *arXiv preprint arXiv:2504.21851* (2025).

[112] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature* (2025), 1–9.

[113] Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems* 124 (2021), 480–494.

[114] T Bedirhan Üstün. 2010. *Measuring health and disability: Manual for WHO disability assessment schedule WHODAS 2.0.* World Health Organization.

[115] Henrik Voigt, Yurina Sugamiya, Kai Lawonn, Sina Zarrieß, and Atsuo Takanishi. 2025. LLM-Powered Virtual Patient Agents for Interactive Clinical Skills Training with Automated Feedback. *arXiv preprint arXiv:2508.13943* (2025).

[116] Rachel Voss and Joe Das. 2024. Mental status examination. *StatPearls* (2024).

[117] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Beyond direct diagnosis: LLM-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107* (2024).

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

[119] Frank W Weathers, Michelle J Bovin, Daniel J Lee, Denise M Sloan, Paula P Schnurr, Danny G Kaloupek, Terence M Keane, and Brian P Marx. 2018. The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): Development and initial psychometric evaluation in military veterans. *Psychological assessment* 30, 3 (2018), 383.

[120] Daniel Weitekamp, Erik Harpstead, and Ken R Koedinger. 2020. An interaction design for machine teaching to develop AI tutors. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.

[121] Ozge C Williams, Sakshi Prasad, Amanda McCrary, Erica Jordan, Vishi Sachdeva, Sheryl Deva, Harendra Kumar, Jayati Mehta, Purushottam Neupane, and Aditi Gupta. 2023. Adult attention deficit hyperactivity disorder: a comprehensive review. *Annals of medicine and surgery* 85, 5 (2023), 1802–1810.

[122] David Wood, Tara Crapnell, Lynette Lau, Ashley Bennett, Debra Lotstein, Maria Ferris, and Alice Kuo. 2017. Emerging adulthood as a critical stage in the life course. *Handbook of life course health development* (2017), 123–143.

[123] Yuqi Wu, Guangya Wan, Jingjing Li, Shengming Zhao, Lingfeng Ma, Tianyi Ye, Ion Pop, Yanbo Zhang, and Jie Chen. 2025. WiseMind: Recontextualizing AI with a Knowledge-Guided, Theory-Informed Multi-Agent Framework for Instrumental and Humanistic Benefits. *arXiv preprint arXiv:2502.20689* (2025).

[124] Joel Yager, Edward R MacPhee, Alexis D Ritvo, and Rakel M Salamander. 2022. Thirty-Minute Psychiatric Management Visits in Academic Medical Centers: Framing and Exploring Distinct Clinical-Educational Social Processes. *The Journal of Nervous and Mental Disease* 210, 2 (2022), 77–82.

[125] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[126] Wei Zhang, Zhenhong Zhou, Kun Wang, Junfeng Fang, Yuanhe Zhang, Rui Wang, Ge Zhang, Xavier Li, Li Sun, Lingjuan Lyu, et al. 2025. LIFEBench: Evaluating Length Instruction Following in Large Language Models. *arXiv*

*preprint arXiv:2505.16234* (2025).

[127] Yongjie Zhou, Zhongqiang Cao, Mei Yang, Xiaoyan Xi, Yiyang Guo, Maosheng Fang, Lijuan Cheng, and Yukai Du. 2017. Comorbid generalized anxiety disorder and its association with quality of life in patients with major depressive disorder. *Scientific reports* 7, 1 (2017), 40511.

[128] William Zinn. 1993. The empathic physician. *Archives of Internal Medicine* 153, 3 (1993), 306–312.

# 8 Appendix

## Detailed Workflow of Stage 4: Plan

The Plan stage updates the set of potential disorders and associated symptoms with functional impairments based on patient responses collected thus far. It iteratively refines the interview plan by analyzing utterances, updating hypotheses, and prioritizing subsequent questions (Figure 3.4 a-e and Figure 4.4 a-e). First, patient responses are classified into symptoms and functional impairments (Figure 3.a). For example, synthesizing responses A1, A2, and A3, the system identifies loss of interest, depressed mood, loss of energy, impaired concentration, and sleep problems as symptoms, while academic, cognitive, and social issues as functional impairments (Figure 4.a).

These items are then matched against DSM criteria to update potential disorders (Figure 3.b). With only responses A1 and A2, the system initially inferred depression, but after adding A3 it also inferred possible comorbidity (e.g., depression and anxiety) (Figure 4.b). The system then queries the DSM knowledge base to identify further information needs (Figure 3.c and Figure 4.c). For symptoms, it lists features associated with the identified disorders. For functional impairments, it uses the WHODAS domains defined in the knowledge base and determines the most relevant areas by taking into account the patient's demographic, reported symptoms, and their relation to the potential disorders. Since some information may already have been addressed, the system filters for items with missing responses (Figure 3.d). At this point, it determines that additional information is still needed to evaluate both depression and anxiety (Figure 4.d).

Finally, the system prioritizes unanswered items and generates corresponding questions to collect missing information (Figure 3.e and Figure 4.e). Prioritization is dynamically adjusted on remaining time and diagnostic importance. This process reflects the time-management strategy of *Design Requirement 2*, allowing the AI interviewer to focus on the most clinically important information within the available time. Priorities are assigned using the rules: (1) questions about the chief complaint, including onset and causes, are given the highest priority; (2) if symptoms of the potential disorder have not yet been confirmed, those items are prioritized; (3) once the chief complaint and symptoms are covered, questions about functional impairments are addressed; and (4) information closely tied to the patient's most recent response is selected as a follow-up. As shown in Figure 4.e, given the time left and priorities, the system identified *excessive worry* as the most urgent and generated a related follow-up question (e.g., returning to Stage 1. Ask).

## Procedure for Generating Simulated Patients

The design of simulated patients in this study builds upon two existing frameworks: PSYCHE and PATIENT-Ψ. The PSYCHE framework focuses on generating patient personas, including diagnostic information and the history of present illness, making it particularly suited for history-taking exercises. In contrast, the PATIENT-Ψ framework emphasizes cognitive modeling and identifies six conversational styles commonly observed in patients, based on formative work with mental health experts. Since PSYCHE does not explicitly incorporate patient interaction styles, and PATIENT-Ψ provide prompt for that, we combine the two approaches to construct a more comprehensive simulation framework.

## Step 1. Generate SP persona (SP-profile, SP-history, SP-behavior) (PSYCHE)

- Input variables: diagnosis, age, and gender.
  - *Diagnosis*: Major Depressive Disorder (MDD), Generalized Anxiety Disorder (GAD), Insomnia, Post-Traumatic Stress Disorder (PTSD), Attention-Deficit/Hyperactivity Disorder (ADHD), and comorbid conditions (MDD+GAD, MDD+ADHD, MDD+PTSD).
  - *Age Range*: 19–33 years.
  - *Gender*: male or female.

From these inputs, LLM generates an SP-profile, which includes:

(1) Identifying data
  - Age:
  - Gender:
  - Marital status:
  - Occupation:
  - Date of birth:

(2) Chief complaint:
  - Presenting problem:
  - Onset:
  - Precipitating factors:

(3) Symptoms
  - Symptom lists based on diagnostic categories.

(4) Functional impairment
  - Cognition (memory, decision making):
  - Self-care (eating, sleeping, get dress):
  - Mobility (physical movement, motor activities, standing):
  - Work (impact of illness to work):
  - Home (impact of illness to home life activities):
  - Get along (impact of illness to relationship, getting along with family, co-worker):
  - Participation in society (on their participation in society such as markedly diminished interest or participation in significant activities, social withdrawal or neglect of pleasurable avocations)

(5) Past psychiatric history
  - Present (yes/no):
  - Description:

(6) Past medical history
  - Present (yes/no):
  - Description:

(7) Current medication
  - Present (yes/no):
  - Description:

(8) Family mental health history
  - Present (yes/no):
  - Description (who, diagnosis):

(9) Developmental social history
  - Childhood history:
  - School history:
  - Work history:

Based on this structured profile, LLM generates an SP-history, which represents a dynamic life story, generated based on the SP-Profile. This contains a lifetime biography, including their present illness and developmental history.

Subsequently, LLM generate SP-Behavior with the given input, SP-Profile, and SP-History, characterized by:

- General appearance, attitude, and demeanor
- Mood and affect
- Spontaneity and verbal productivity
- Tone of voice
- Social judgment and insight
- Reliability of reporting
- Perceptual abnormalities
- Thought process and thought content

## Step 2. Defining Patient Interaction Style (PATIENT-Ψ)

To simulate variability in patient–clinician interactions, the model applies one of six conversational styles identified in the PATIENT-Ψ framework:

- **Plain**: The client is designed as a standard patient who has no specific types.
- **Verbose**: A verbose client may 1) provide detailed responses to questions, even if directly relevant, 2) elaborate on personal experiences, thoughts, and feelings extensively, and 3) demonstrate difficulty in allowing the therapist to guide the conversation.
- **Upset**: An upset client may 1) exhibit anger or resistance towards the therapist or the therapeutic process, 2) you may be be challenging or dismissive of the therapist's suggestions and interventions, 3) have difficulty trusting the therapist and forming a therapeutic alliance, and 4) be prone to arguing, criticizing, or expressing frustration during therapy sessions.
- **Reserved**: A reserved client may 1) provide brief, vague, or evasive answers to questions, 2) demonstrate reluctance to share personal information or feelings, 3) require more prompting and encouragement to open up, and 4) express distrust or skepticism towards the therapist.
- **Tangent**: A client who goes off on tangent may 1) start answering a question but quickly veer off into unrelated topics, 2) share personal anecdotes or experiences that are not relevant to the question asked, 3) demonstrate difficulty staying focused on the topic at hand, and 4) require redirection to bring the conversation back to the relevant points.
- **Pleasing**: A pleasing client may 1) minimize or downplay your own concerns or symptoms to maintain a positive image, 2) demonstrate eager-to-please behavior and avoid expressing disagreement or dissatisfaction, 3) seek approval or validation from the therapist frequently, and 4) agree with the therapist's statements or suggestions readily, even if they may not fully understand or agree.

## Step 3. Conversational Response Generation

For each clinician–patient interaction turn, SP responses are generated in two stages:

(1) **Persona-based generation**: Responses are first generated based on the SP's profile, history, and behavioral.

```
Prompt: Imagine you are a patient experiencing mental
health challenges.
Patient's profile: {sp_profile}
Patient's history: {sp_history}
Patient's behavior: {sp_behavior}
You will engage in a conversation with the clinician.
Adhere to the following guidelines:
1. Use natural language, including hesitations, pauses,
and emotional expressions, to enhance the realism
of your responses.
2. Gradually reveal deeper concerns and core issues,
as real patients often require extensive dialogue
before delving into more sensitive topics. This
gradual revelation creates challenges for clinicians
in identifying the patient's true thoughts and emotions.
3. Ensure that your responses align with the provided
background information.
4. If asked about a symptom not included in the
patient profile, respond neutrally and indicate
that you do not experience problems with that symptom.
5. Do not ask the clinician any questions.
6. Do not state your diagnosis directly, only report
the symptoms.
7. Limit your responses to a maximum of 50 words.
Respond in Korean.
Recent conversation history:
{conversation_history}
```

(2) **Style adaptation**: The response is then modified according to the assigned conversational style, ensuring alignment with real-world patient variability.

```
Prompt: You will engage in a conversation with
the clinician regarding the following situation and
behavior.
This is the information that you can reply to
clinician: {initial_response}.
However as a mental health patient, you follow this
style: {sp_style}. Update the response to match your
style.
Do not include prefix "Patient:".
Do not ask the clinician any questions.
Limit your responses to a maximum of 50 words.
Respond in Korean.
Recent conversation history:
{conversation_history}
```

We separated the process into two steps because the persona contains a large amount of information, which can sometimes cause the LLM reduce adherence to the patient's conversational style. For example, it may generate extensive information even when the patient's style is reserved. Prior work shows that models instruction-following accuracy declines as prompt length excessively increases. [67, 126]

## Time estimation for dialogue simulation

The AI interviewer's response time was fixed at 2.6 seconds, corresponding to the average API call latency. For the simulated patient, *response time* was modeled as the sum of *reading, thinking, and typing times* following the Keystroke-Level Model (KLM) framework [23].

Reading time was estimated using the average adult silent reading rate of 238 words per minute [20]. Specifically, the number of words in the question was divided by this rate to calculate the time required:

$$\text{Reading time (seconds)} = \frac{\text{Question length (in words)}}{\text{238 words}} \times 60\text{seconds}$$

Thinking time was modeled as a function of response length, based on evidence that longer answers require more cognitive planning [26]. Specifically, we defined thinking time, bounded between 3 seconds to 5 seconds [10]:

$$\text{Thinking time (seconds) = 0.1 x Response Length (in words)}$$

Typing time was estimated using the average adult typing rate of 36.2 words per minute [80]. Specifically, the length of the response was divided by this rate to calculate the required time:

$$\text{Typing time (seconds)} = \frac{\text{Response length (in words)}}{\text{36.2 words}} \times 60\text{seconds}$$

Both AI interviewer and SP are based on GPT-4o-mini [78], version 2024-07-18.

**Table 2: Information slots to be elicited during the interview in relation to potential psychiatric disorders, including the chief complaint, representative symptom examples, and six domains of functional impairment based on the WHODAS framework.**

| Slot name | Slot Description | Potential Disorders |
|---|---|---|
| | **Chief complain** | |
| Presenting problem | The presenting problem for which the patient seeks professional help. | – |
| Onset | When did this problem first appear | – |
| Precipitating factors | Precipitating factors need to be collected. Precipitating factors is stress or environmental change around the time this issue began. | – |
| | **Symptom (72 items)** | |
| Mood | (core symptom) Whether the patient has felt depressed, sad, empty, hopeless, or "down" in the past 2 weeks. Frequency of these feelings. | MDD |
| Interest | (core symptom) Activities the patient usually enjoys (e.g., hobbies, social events, work tasks). Whether they still enjoy these activities. | MDD |
| Appetite | Whether appetite decreased or increased or weight decrease or increase without trying intentionally. | MDD |
| Sleep difficulty | (core symptom) Whether the patient has sleep difficulty. How often do patient experience difficulty sleeping. | MDD, GAD, PTSD, insomnia, BP |
| Wakeup midnight | (core symptom) Whether the patient wake up ad midnight and hard to get back to sleep | MDD, GAD, PTSD, insomnia, BP |
| Psychomotor | whether the patient has experienced noticeable psychomotor agitation (restlessness, inability to sit still) or psychomotor retardation (slowed movements or speech). | MDD |
| ... | ... | ... |
| | **Funtional impairment (6 items)** | |
| D1. Cognitive | whether the patient has signs of cognitive impairment. Typical features include slowed thinking, increased pauses before answering, difficulty concentrating, and impaired decision-making. Patients may appear easily distracted, complain of memory difficulties, or report a general inability to think clearly. | – |
| D2. Mobility | whether the patient has signs of impaired physical movement. Typical features include slowed body movements, reduced motor activity, or observable physical slowness during tasks such as walking, standing up, or general movement. It may reflect psychomotor retardation in some mental health conditions. | – |
| D3. Self care | Whether the patient have difficulty in maintaining basic self-care routines including sleep, appetite, get dress. Typical impairments include neglect of personal hygiene, irregular eating or sleeping habits, failure to dress appropriately, or inability to manage medications or health appointments. | – |
| D4. Get along | Whether patient has problems related to getting along with family members, friends, co-worker. Typical impairments may include conflicts with family members, withdrawal from family interactions, or interpersonal difficulties with supervisors and coworkers. In cases where ADHD is a potential diagnosis, peer relationships are often affected by rejection, neglect, or teasing. | – |
| D5. Life activities School/Work | Impact of patient's problem on their ability to perform school/work-related tasks. Typical indicators include reduced productivity, difficulty meeting deadlines, increased errors, absenteeism, poor concentration during tasks, or inability to sustain attention or effort throughout the workday. | – |
| D5. Life activities Household | Impact of patient's problem on their ability to manage daily activities at home. Typical impairments may include difficulty maintaining household responsibilities (e.g., cleaning, cooking, childcare). | – |
| D6. Participant in society | Impact of patient's problem on their participation in society such as marked diminished interest or participation in significant activities, social withdrawal or neglect of pleasurable avocations. Only collect information that is related to social participant. | – |

**Table 3: Three dimensions and descriptions for PIQSCA measure from PSYCHE framework [60]**

**Dimension 1: Process of the interview**

1 = No components of the interview process were followed.

- Example: The interview lacked any recognizable structure, and key steps such as greeting, inquiry, or closing

2 = Vital components from two stages of the interview process (initial, near the end, or at the end) are missing.

- Example: Critical elements, such as a proper introduction or closure, were absent in two parts of the interview

3 = A vital component from one stage of the interview process (initial, near the end, or at the end) is missing.

- Example: The interview was generally structured, but one key component, such as greeting the patient warmly or properly closing the session, was absent.

4 = Some aspects of the interview process were incomplete, but the overall structure of the interview was maintained.

- Example: While a few minor steps were skipped, the overall flow - from introduction to conclusion - was mostly preserved.

5 = The interview process was fully followed: The therapist warmly greeted the patient and introduced themselves, began the session with open-ended inquiry, informed the patient near the end that the interview was concluding soon, and, at the end, either summarized the diagnosis and treatment options or informed the patient about the next session.

- Example: The therapist demonstrated full adherence to the ideal interview structure, ensuring a smooth start, transition, and conclusion to the session.

**Dimension 2: Techniques**

1 = No facilitating interventions were used.

- Example: The therapist did not use any techniques to encourage or support the patient's communication, making the session purely one-sided or unproductive.

2 = More obstructive interventions were used than facilitating ones.

- Example: The therapist's interventions hindered the patient's communication more than they helped, such as close-ended questions, compound questions, and premature advice.

3 = An equal number of facilitating and obstructive interventions were used.

- Example: The therapist's facilitating interventions (e.g., reinforcement, reflection, and acknowledgement of emotion) were balanced by obstructive behaviors (e.g., close-ended questions, compound questions, premature advice).

4 = Some obstructive interventions were used, but there were clearly more facilitating interventions.

- Example: Although the therapist occasionally used obstructive techniques (e.g., close-ended questions, compound questions, premature advice), the majority of their interventions helped the patient express themselves and engage meaningfully.

5 = All facilitating interventions were used appropriately, with no obstructive interventions.

- Example: The therapist consistently used supportive interventions, such as reinforcement, reflection, and acknowledgement of emotion, with no obstructive behaviors, creating a smooth and productive session.

**Dimension 3: Information for Diagnosis**

1 = No information or vital information is missing.

- Example: The interview lacked essential details needed to form any diagnostic impressions.

2 = Some information gathered, but additional details are needed to make a reliable diagnosis.

- Example: A partial history or symptom discussion was obtained, but key elements required for a first impression diagnosis were not covered.

3 = Sufficient information for a first impression diagnosis, but no effort to gather further details to rule out other possible disorders.

- Example: The primary symptoms were discussed, allowing for a preliminary diagnosis, but differential diagnoses were not explored.

4 = Enough information for a first impression diagnosis, but not enough to rule out other possible disorders.

- Example: Information gathered allowed for a likely diagnosis, but additional investigation is necessary to rule out alternatives.

5 = Comprehensive information was gathered, allowing for both a first impression diagnosis and ruling out other disorders.

- Example: The interview covered all areas, including differential diagnosis considerations, providing a well-rounded assessment.

Section 5

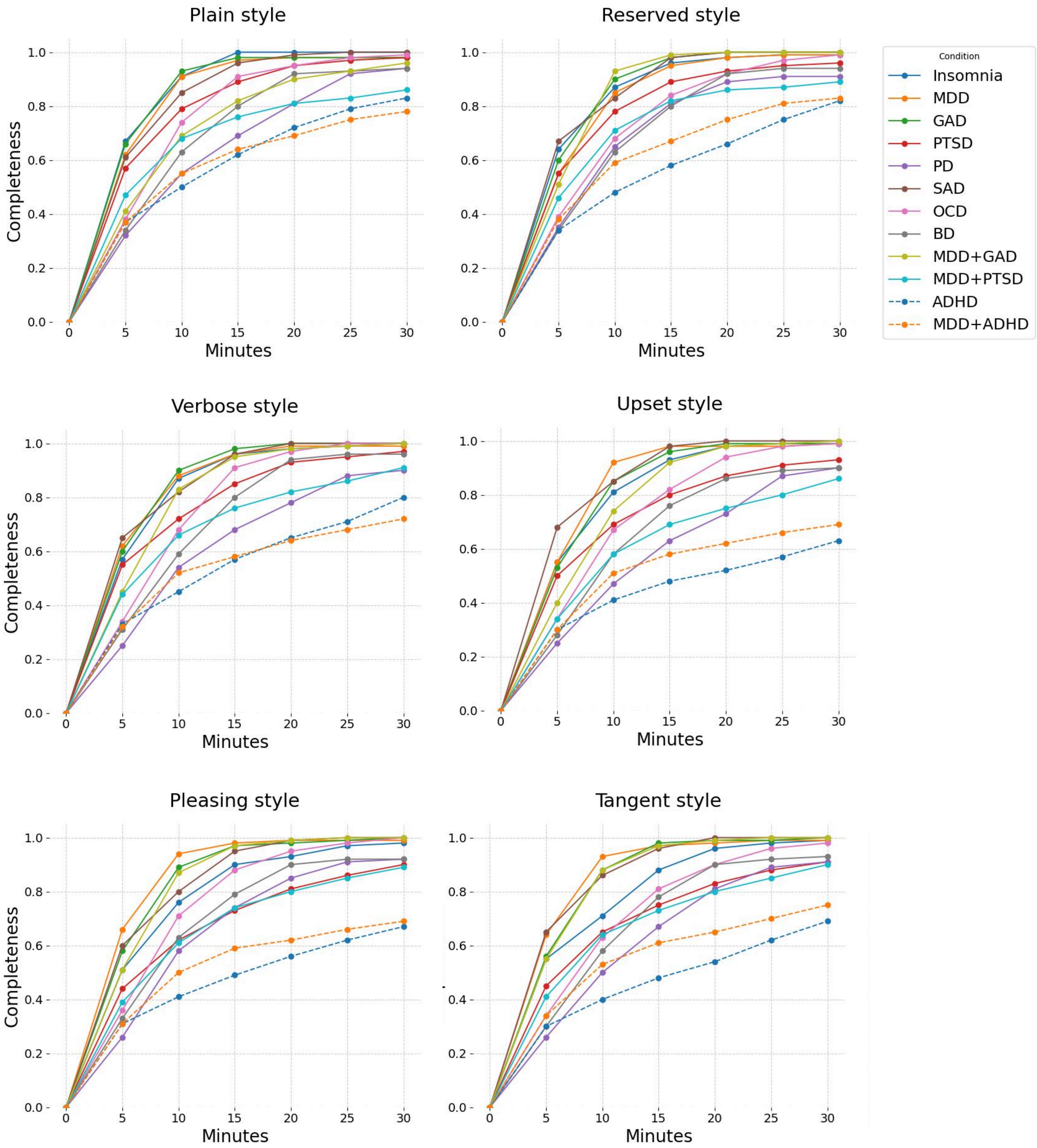


**Figure 6: This graph shows how the completeness of information from interviews changes over time. Each line on the graph represents a different patient condition, and the data points show the average completeness for 30 persona with a mental condition and a conversational style. Each disorder represents: Major Depressive Disorder (MDD), Generalized Anxiety Disorder (GAD), Post-Traumatic Stress Disorder (PTSD), Panic Disorder (PD), Social Anxiety Disorder (SAD), Obsessive-Compulsive Disorder (OCD), Bipolar Disorder (BD), and Attention-Deficit/Hyperactivity Disorder (ADHD).**